A spatial spectral domain integral equation solver for electromagnetic scattering in dielectric layered media

A spatial spectral domain integral equation solver for electromagnetic scattering in dielectric layered media

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op woensdag 22 november 2017 om 16:00 uur

door

Roeland Johannes Dilz

geboren te Utrecht

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. A. B. Smolders
1e promotor:	dr.ir. M. C. van Beurden
2e promotor:	prof.dr. A. G. Tijhuis
leden:	prof. E. Heyman (Tel Aviv University)
	prof. R. Orta (Politecnico di Torino)
	prof.dr. W. M. J. M. Coene (TU Delft)
	prof.dr.ir. A. P. M. Zwamborn
	dr. J. J. G. M. van der Tol

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

This work is part of the research programme HTSM with project number 12786, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

This work has been facilitated by a significant financial and intellectual contribution from ASML.

A spatial spectral domain integral equation solver for electromagnetic scattering in dielectric layered media
by R. J. Dilz
Technische Universiteit Eindhoven, 2017

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-94-6233-774-9
NUR: 958 - Elektrotechniek

This thesis was prepared with the LaTeX 2$_\varepsilon$ documentation system
Reproduction: Gildeprint Drukkerijen, Enschede, The Netherlands

Voor mijn vader

Contents

Summary

Electromagnetic (EM) scattering from dielectric objects embedded in stratified dielectric media is relevant for the design of optical integrated circuits, inverse imaging in integrated-circuit metrology, and the design of metamaterials. Design and optimization for these applications requires knowledge about the responses of the aforementioned structures to incident EM waves. Numerical methods are often employed to find this response. To optimize one single design, numerous applications of such a numerical method are required. Numerical methods available to calculate such a response are available, however, the time required for such a large number of computations is often impractical for design and optimization.

The goal of this research is to find a more efficient numerical method to calculate this scattering response. The numerical methods in existence can be classified into so-called global methods and local methods, where local methods such as FEM, FDTD and FIT, approximate Maxwell's equations on elements on a mesh or grid and couple these elements. Global methods, such as domain and boundary integral equations, approximate Maxwell's equations on a larger domain at once. The translation symmetry in the transverse direction of the stratified medium, which is a global symmetry, requires a global method to exploit it. Our method of choice is a domain-integral equation formulation, which consists of a field-material interaction that calculates a contrast current density, and Green function convolutions that calculate the electric field generated by this contrast current density.

The Green function in our integral equation can be represented in the spatial and the spectral domain. An analytical expression for the Green function exists in the spectral domain, and it is possible to calculate a spatial-domain Green function through tedious Sommerfeld integrals for a spatial volume or boundary integral equation. However, we prefer a spectral representation for the Green function. For periodically repeating scattering objects, the periodicity can be exploited to represent functions on a discrete spectral lattice, e.g. the periodic volume integral method and rigorous coupled-wave analysis. Contrary to that, finitely sized scatterers require a continuous spectral domain, hence a discretization in the spectral domain is needed. Additionally, the field-material interaction is most efficiently represented in the spatial domain. Therefore, a discretization in the spatial domain is needed as well and a fast way of transforming between spatial and spectral domain should be available.

The development of a mixed spatial-spectral numerical method takes place in stages with an increasing difficulty, starting from the two-dimensional case of a homogeneous background in TE polarization, which yields a scalar problem. To discretize the contrast

current and the electric field we use a Gabor frame, i.e. a representation in terms of a sum of weighted Gaussians that are shifted with a uniform spacing and modulated with a uniform frequency step. Since the Gabor frame consists of frame functions that can be Fourier transformed analytically, a fast and exact transformation of the electric field and contrast current density between the spatial and spectral domain exists. The homogeneous-medium Green function contains branch-cuts along which rapid oscillations occur. These oscillations are hard to represent accurately with a small number of Gabor coefficients. To retain an accurate algorithm, we use a scaling of the spectral coordinate to smoothen these oscillations. The discretized combination of the field material interaction and Green function convolution are used as a matrix-vector product in an iterative solver. The Gabor frame allows to compute the matrix-vector product in an amount of time that scales as $O(N \log N)$ with the number of unknowns.

Subsequently a two-dimensional scattering problem with transverse electric polarization in a layered medium is solved. In a layered medium complex poles can appear in the Green function. These poles are not integrable, so a coordinate stretch cannot be used to smoothen the peaks originating from these poles. By representing the contrast current density and the electric field on a specially chosen path through the complex plane, the poles can be evaded. The special choice of the path allows for fast transformations to and from the path. Along the complex-plane path, the poles and oscillations along the branchcuts are smoothened so that the Gabor frame can be used to accurately and efficiently represent the Green function. It is shown again that the computational complexity of the matrix-vector product scales as $O(N \log N)$ with respect to the number of unknowns, that scales with the size of the embedded scatterer.

The same complex-plane spectral path is then used to solve a scattering problem with TM polarization, which is the first vectorial problem that we tackle. It is well known that periodic spectral solvers can have difficulties in solving scattering problems with TM polarization because two functions with spatial discontinuities are convolved in the spectral domain, a convolution that is not well-defined. We show that also for a continuous spectral solver such a convolution yields a poor convergence with respect to the range in the spectral domain that is included in the simulation. To overcome this effect, we apply an auxiliary-field formulation. Again, this leads to an $O(N \log N)$ complexity for the matrix vector product.

By applying the knowledge gained from the two-dimensional solvers, a three-dimensional solver has been constructed. The complex-plane path integration in the spectral domain for the 2D case is replaced by a two-dimensional manifold in the complex plane, which consists of nine regions. The auxiliary-field formulation is combined with a normal-vector field formulation for an efficient handling of the field-material interaction for discontinuous field components at material interfaces. To further increase the computational efficiency, operations on Gabor coefficients are approximated by faster methods that require a smaller number of fast Fourier transformations and less overhead. This yields a full-wave three dimensional solver for stratified dielectric media that scales as $O(N \log N)$ on the number of unknowns and that scales well to large-sized scatterers.

All algorithms were tested against a finite element method solver to at least three digits accuracy. For the three dimensional algorithm a computation of the scattering far field is shown of a grating with a contrast $\chi = 1.25$ consisting of 10 dielectric lines on a substrate. The size of the problem is $\lambda/4.1 \times 16\lambda \times 16\lambda$, and it was discretized using $5.1 \cdot 10^6$ unknowns and the system was solved in under two hours on a single core of an i7-4600U cpu while using 10GB of RAM.

The thesis is completed by arguing that the use of Gabor frames is not vital for this type of spatial-spectral methods. Any discretization in the transverse plane that is a good approximation in the spatial as well as the spectral domain can yield good accuracy. Additionally, when a fast means of Fourier transformation exists between the spatial domain and a complex path in the spectral domain, then it is suitable to apply in a spatial-spectral solver. As an example it is shown that a Hermite-interpolation-based discretization can also yield an efficient algorithm.

Chapter 1

Introduction

1.1 Wafer metrology

Since the invention of the triode vacuum tube in 1906 [1], the developments in electronic circuitry progressed at a rapid rate. With an increasing reliability and decreasing costs, more and more advanced circuitry became achievable. The increasing possibilities allowed more advanced communication techniques and the advances in communication techniques also increased the demand for more advanced circuitry. During the second World War, the state of electronic technology allowed it to be applied for deciphering coded messages in the first computer, the Colossus at Bletchly park [2]. This opened up a new, major field where electronic circuits were applied: computing.

During the following years computers consisted of vast numbers of discrete components and connecting wires, which was hard to manage. One method to cope with this problem was to integrate multiple components inside a single package. Technically, the first integrated circuit, the Loewe 3NF that was designed by Manfred von Ardenne, already dates back to 1925 [3, Chapter 11]. However, it was the invention of semiconductor-based integrated circuits by Geoffry Dummer in 1956 that made it possible to easily incorporate vast numbers of components on a single flat piece of semiconductor material [4]. With this integrated-circuit technology, the required number of production steps does not depend very strongly on the number of components in a circuit, i.e., just a few production steps can produce a vast number of components.

Currently, integrated circuits are still constructed on semiconductor material via a lithographic process. On a flat piece of semiconductor, a so-called wafer, materials can be deposited or etched away. By applying photosensitve layers and illuminating them with a pattern that characterizes electronic components, an integrated circuit can be constructed. The number of integrated components on a single integrated circuit has risen from several hundreds around 1965 [5] to more than five billion in 2017 [6]. The exponential growth of this number was already predicted in 1965 by Gordon Moore [7]. To allow such large quantities of components in a single package, the size of these components has been greatly reduced as well. Modern integrated circuits for use in computers are produced with a detail of 10 nm [8].

To enable cost-effective lithographic production of so many components at this level of detail, several key parameters in the production process have to be closely monitored. The critical dimensions (CD) is the width of the smallest features that are produced. A good control of the CD ensures that the components are not erroneously interrupted when they are too thin, or touch other components when they are too thick. The alignment of different layers is critical, since structures and connections in a layer need to connect precisely to the structures in other layers. This alignment is measured in an overlay measurement. For this type of measurement it is required that the measurement technique penetrates somewhat into the layers on the wafer, since that will allow to assess the alignment of a pattern in one layer with respect to a pattern in another layer. Another parameter is the focus of the illumination stage. An unfocused beam will not illuminate as precisely as is required. The illumination dose is also important, since an illumination that is too short or too long can produce artefacts in the pattern on the wafer.

Keeping the machinery optimally calibrated requires a continual monitoring of the lithographic process. Therefore, fast measurements are preferred for continual calibrations. Ideally, a lithography machine is calibrated on each wafer that is produced, such that the calibration error of the previous wafer is compensated immediately for the next wafer. Integrated circuits are often produced in many illumination steps on several machines, so a high accuracy is of vital importance. When a measurement technique allows the calibration of the lithographic process within the production line at each newly produced wafer, the accuracy is monitored very well. However, such a mode of calibration requires measurements with a nanometer-precision instrument at a high speed.

There are several popular measurement techniques for this type of problem.

- Atomic force microscopy (AFM) measures structures down to atom level, by measuring the force between a tip and the sample [9]. The tip is moved over the sample, and the force between the tip and the sample is recorded. From this data an image can be constructed of the sample. The accuracy of this method is very high. However, it takes time to move the tip over the sample and only the surface can be characterized. Since only the surface can be measured, a destructive technique is required if measurements below the surface of the integrated chip are required.

- Scanning electron microscopy (SEM) measures the reflection of a very narrowly focused electron beam from the surface [10, Chapter 9], [11]. The accuracy of the method is high and measurements can be carried out moderately fast. The technique is especially popular for CD measurements, although it can be used for overlay measurements as well at high voltages [12].

- Scatterometry is an optical technique where details much smaller than the diffraction limit can be measured [13]. However, the diffracted light that is registered is not directly related to the shape of the scatterer and it requires a significant computational effort to recover the calibration parameters. Light of several wavelengths can be used, so wavelengths can be chosen that penetrate into the integrated circuit, which allows for very fast non-destructive measurements.

- Ellipsometry is an optical technique where light with a known polarization is reflected from a surface. From the polarization of the reflected light information can be retrieved about the reflecting surface [14]. In semiconductor manufacturing this non-destructive process is popular for characterizing the thickness of the deposited layers [15].
- Soft X-ray works similar to scatterometry [16], but with a much shorter wavelength, on the order of 10 nm. At the moment, this technique is still rather experimental, but it has the advantage of employing light at much shorter wavelengths, which allows for even more precision and non-destructive measurements. A downside is the low contrast value of materials at these wavelengths. A difficulty with this technique are the focussing optics, since they are relatively absorbing.
- Hard X-ray employs X-rays with a wavelength much shorter than the detail size [17]. Here, a transmission approach is followed, instead of measuring the scattering. Optics are even more difficult at such short wavelengths, and, again, this technique has not matured to the level of e.g. SEM and scatterometry. However, the technique promises fast, non-destructive measurements at a high precision.

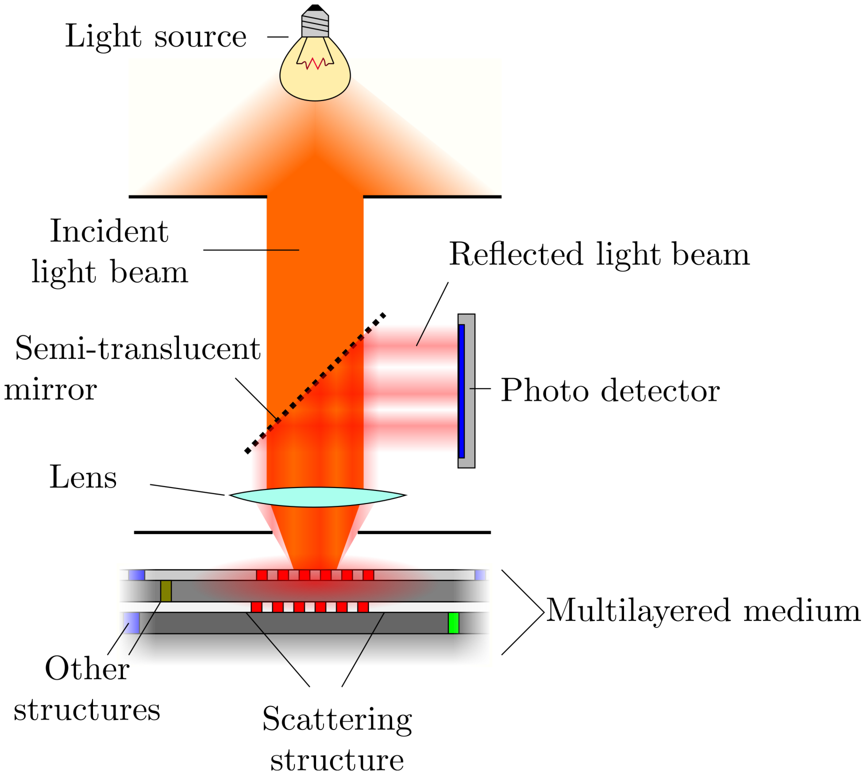


Figure 1.1: A simplified setup for a scatterometry measurement.

Non-destructiveness of a measurement technique is very important. Therefore, scatterometry is a popular method. The method is schematically depicted in Figure 1.1. In practice, special metrology targets are constructed on the wafer for these measurements. One of the challenges for scatterometry is that the scattered light, as it is registered, has no direct relation to shape of the target, in the sense of e.g. a microscopy image. Given a certain metrology target, it is possible to compute how the light is scattered in a so-called forward computation. However, the other way around, the complete shape of the metrology target as it is produced by the lithography apparatus cannot be reconstructed directly from the scattered light, because a lot of the information about the target structure does not radiate into the far field and is therefore not registered. To reconstruct the shape of the measured scattering target, the technique of inverse inverse scattering is applied.

In this case, the technique of inverse scattering involves fitting a slightly deformed metrology target as it is produced by the mis-calibrated lithography machine. A large number of forward computations is required, each of them corresponding to a different deformation in the realized metrology target. By comparing the results of the computed scattered light with the measured scattered light, a good fit can be found to the realized metrology target. So via the use of prior information about the target, sub-wavelength features about the metrology target can be recovered. When the dimensions of the realized target are known, recalibration parameters can be computed. Clearly, a large number of forward computations are required to recalibrate after measuring a target structure. To mitigate high-demands on the solver, a library-based approach is often preferred. Such a library samples the parameter-space of mis-calibrations, which can be interpolated to find the most suitable recalibration. However, even though the library can be computed outside of the production process, it is challenging to fill this library in acceptable computation times.

When the shape of the target is well-chosen, this can yield reliable measurement results. It is essential that the diffraction pattern is sensitive to the parameters of the metrology target on which the calibration is carried out. Optimizing the design of such a metrology target also requires a large number of forward computations. Typically, a periodically repeating target is chosen, which has the benefit of stronger reflections and a clear distinction between reflection orders. Another advantage is that a forward computation can be carried out much faster with the assumption of an infinitely repeating scatterer. This is important, since these computations are complex and require a significant amount of time and computational resources.

1.2 Research question

The demands on the accuracy of the scatterometry measurements increases since the feature size of integrated circuits is still decreasing. The required accuracy of the measurement technique is already orders of magnitude smaller than the wavelength of the light source. Additionally, there is a demand for smaller metrology targets, since it can be beneficial to locate the targets within the integrated circuit, where the space occupied by the target competes with the space occupied by the integrated circuit itself. At a certain number of repeating elements and a certain accuracy level, the assumption of infinitely repeating metrology target breaks down. Additionally, surface waves that are present in infinitely repeating structures [18] are not present in finite structures. To obtain some insight into the limits of this periodicity assumption, forward computations need to be carried out for the entire metrology target. Although computational methods for this problem exist, the metrology targets are often too large to conveniently apply them on current computers, both because of the memory requirements and the computation time. An important difficulty for the computational methods is that an integrated device is built from a large number of different materials, stacked as layers on top of each other. For many computational methods such a multilayered medium requires a significantly larger computational effort.

Metrology is not the only field where solvers for finite scatterers in multilayered dielectric media are required. Another use for such a solver would be metasurfaces, i.e., the science of creating surfaces with microscopic objects that interact with electromagnetic fields on a macroscopic scale by combining all microscopic interactions. Examples include flat lenses [19, 20, 21, 22]. Also in integrated optics lightwaves interact with each other via complex structures of e.g. waveguides, that are created with the aid of techniques similar to those for ordinary integrated circuits [23, 24, 25]. Therefore, these waveguides are also deposited on a multilayered medium and the design of these also requires a solver for multilayered media.

The goal of this study is to create an efficient full-wave electromagnetic solver for finitely sized dielectric objects embedded in a multilayered medium. Important is the flexibility of this solver. It should be possible to define the shape of the scatterer independent of the expansion of the electric fields. This requirement leads to a continuous dependence of the result of a forward computation with respect to the scatterer geometry, which is advantageous for the fitting procedure in inverse scattering. The solver should also be very flexible with respect to changes in the multilayered medium, such as layer thickness and dielectric constant.

The primary goal of this project is to obtain a solver with which the scattering from large, finite objects embedded in a layered medium can be studied. To do so, both a high accuracy of 10^{-3} and a high computational efficiency in both memory and computation time are required. A secondary goal would be to be able to apply such a solver directly in the production process, which would require forward computation times in the order of seconds for complete metrology targets.

1.3 Existing numerical methods

A very large variety of computational methods for electromagnetic scattering problems exists. We will divide the existing numerical methods in two classes, i.e. in local and global methods. A local method is based directly on Maxwell's equations in differential or integral form. It uses only short-distance interactions and connects local interactions into a single large system, where interactions at a distance are coupled via a chain of locally connected interactions. Global methods employ derived formulations. The response of the entire system to a single point source in space is computed and therefore all source points are interacting with all observation points directly. Global methods include Green function methods and modal methods, which will be discussed shortly in Section 1.3.1-1.3.3. Some important examples of both methods, will be discussed in the context that we search for a method that

- is applicable to multilayered media,
- is tailored towards finitely sized objects,
- is efficient for single-wavelength-sized scattering problems.

1.3.1 Local methods

A straightforward version of a local method is the finite difference time domain (FDTD) method [26, 27, 28]. This method approximates the differential operators in Maxwell's equations by finite differences. In a large number of discrete time steps the response of the system to an incident electromagnetic pulse is found. The strong point is its very wide range of applicability and ease of implementation. The time-domain nature of FDTD will yield results for a continuous spectrum of wavelengths. This can be advantageous for certain problems, but it comes at the cost of incorporating the time dimension explicitly as a variable in the computation, requiring four dimensions for a problem with three spatial dimensions.

Finitely sized objects embedded in an infinite space can be incorporated through applying radiation boundary conditions. These conditions are applied at the boundary of the simulation domain and guarantee that no waves reflect back at the edge of the open boundary. This mimics the wave-propagation of an infinite homogeneous medium outside the simulation domain. Both the Mur absorbing boundary [29] and perfectly matched layers (PMLs) [27, Chapter 7 and 8], [28, Sections 3.2 and 4.2] are popular methods to implement radiation boundary conditions. Multilayered media can be incorporated as well and since they extend to infinity, they can be incorporated together with absorbing boundary conditions, e.g. PMLs. A downside for multilayered media is that the simulation domain has to extend to all layers of the multilayered medium [29, 30].

Where the FDTD method discretizes the differential form of the Maxwell equations, the finite integration technique (FIT) discretizes the integral form of the Maxwell equations [31]. Similar to the FDTD, Cartesian grids are used in the spatial domain and a marching-on-in-time scheme is used for FIT in the time dimension. Advantages of FIT are that Maxwell's laws are not approximated, which leads to e.g. conservations of energy and charge. However, the approximations are made in the constitutive relations. Again, radiative boundary conditions are used for finite scatterers in multilayered media.

Another popular local method is the finite-element method (FEM) [32]. In this method the time dimension is eliminated by assuming a harmonic excitation of the object and a linear response of the scattering object. The object can be meshed nonuniformly, such that a finer meshing is possible where the field is singular, such as around corners and edges of objects. On each mesh cell a polynomial expansion represents the fields inside the cell and on its boundary. The field at the boundaries of neighboring cells are coupled to satisfy the continuity conditions of the Helmholtz equation and the resulting linear equation is usually solved using a fast, direct solver. Recently, a fast solver was proposed to solve the pertaining linear system in $O(N)$ operations [33], with N unknowns. However, in most cases a solver is employed that does not scale that well to large problem sizes. FEM tackles finitely sized problems with a radiative boundary condition as well.

Both FEM and FDTD are popular choices for scattering problems in metrology [34, 35, 36]. Higher-order basis functions and the absence of the time dimension in FEM are advantageous for reaching high levels of accuracy [36], although all the results in [34, 35, 36] rely on the assumption of periodic scatterers to keep the simulation domain small.

Global methods: Free-space methods

An important branch of the global methods uses a Green function, i.e. the response of the background medium (e.g. vacuum) to a point-like source. However, modal methods do not use a Green function and an example of such a method, the Fourier modal method, is briefly described in Section 1.3.3. In most cases, the Green function for a harmonic time dependence is used. The field at the location of the scatterer is discretized by a set of basis functions, each of which represents a small part of the field on the complete scatterer. The Green function is then employed to compute the response of one basis function due to fields due to another basis function. This leads to a large matrix equation, which can be solved directly or via iterative techniques. When iterative techniques are used, the full matrix does not have to be calculated. A method to compute the electric field from a given source distribution is the only requirement. Both volume integral equations (VIEs) and surface integral equations (SIEs) follow this approach.

An important example of a VIE is the CGFFT method [37]. Since the Green function is translation invariant, the integral over the Green function takes the form of a spatial convolution. The CGFFT method exploits the fact that a convolution over uniformly sampled function values or expansion functions with equal spacing can be computed efficiently using fast Fourier transformations (FFTs) [38, Section 13.1]. The conjugate-gradient (CG) iterative technique is used to solve the resulting matrix system [39]. Since the number of iterations does not depend strongly on the number of basis functions, N, the computation time for such a scattering problem scales as $O(N \log N)$ with the number of unknowns.

A different approach employing a SIE formulation is the Boundary Element Method (BEM), which is useful for scattering from homogeneous objects [40]. In this surface-integral formulation, the fields are only treated at the boundaries of the scattering objects. These boundaries are meshed, e.g. into triangles (Rao Wilton Glisson functions [41]), to approximate the fields accurately. By approximating the object boundaries by small triangles, they can be approximated accurately. On each of these triangles a surface-current is used on both sides, to model the transmission of fields to the other sides. These surface current densities are modeled through a set of basis functions on each triangle. The Green functions are used to compute the interactions inside and outside the homogeneous objects from triangle to triangle. Boundary conditions are used to match the fields on both sides of the object boundaries. This results in a matrix equation that can be solved both directly in $O(N^3)$ steps and iteratively in $O(N^2)$ steps to find the surface current density on the mesh and in a post-processing step the fields can be computed at any other location, when no further optimizations are applied for the matrix-vector product.

For both VIE and SIE formulations optimization schemes exist. Several improvements have been made on CGFFT, such as the adaptive integral method (AIM) [42] for SIEs and pre-corrected FFT [43, 44] for VIEs. These methods employ two discretizations, a fine one for local interactions and a coarser one with equidistant sampling for long-range interactions. The long range interactions can be approximated on the equidistant grid and are handled by FFTs comparable to CGFFT, which results in $O(N \log N)$ computation time. Another important method to speed up computation is by making low-rank approximations of the integrals over the Green function by multipoles for interactions at a larger distance. This so-called fast multipole method [45, 46] only employs the full expression for the Green function integrals for short distance interaction and employs the approximation for larger distances, which is significantly faster. Using this method a matrix-vector product can be computed in $O(N^{3/2})$ time. For even larger systems, the multi-level fast multipole algorithm (MLFMA) extends this into a hierarchical structure where the interactions are computed through a tree-structure for both SIE formulations [47] and VIE formulations [48]. It is possible to compute matrix-vector products with the MLFMA method in $O(N \log N)$ time. Extensions of the method are available for multilayered media [49]. An alternative to the fast multipole expansion is the fast inhomogeneous plane wave expansion (FIPWA) [50, 51]. This method computes long-distance interactions using an expansion in inhomogeneous plane waves, instead of multipoles. Multilevel variants of this method exist as well. It is interesting to see how the free-space methods mentioned in Section 1.3.1 can be used for multilayered media [52]. The implementation of the MLFMA method for multilayered media consists of changing the Green function with one of the multilayered versions discussed in Section 1.3.2.

There are important differences in the computational complexity between VIEs, such as CGFFT, and SIEs such as BEM. For scattering from large objects, SIEs have a smaller number of unknowns than VIEs, since the number of unknowns scales with the surface area of the scatterer instead of with the volume. On the other hand, a SIE is significantly more complex, since the surface elements require more intricate testing and basis functions that project onto the surface, whereas for a VIE all elements are defined already in three-dimensional coordinate system to which the Green function applies. This leads to a relatively longer computation time per unknown for SIEs. The structures that are encountered in metrology are often of sizes comparable to or smaller than the wavelengths, for which the number of unknowns for both a SIE and VIE are comparable. In such cases, and when the iterative solver converges fast, VIEs are advantageous. For this reason we focus on a VIE, although a SIE does perform better in some cases, e.g. for scatterers of large size and large dielectric contrast.

1.3.2 Global methods: Green function for multilayered media

Most global methods somehow incorporate the Green function. This Green function represents the electromagnetic field on the whole computational domain radiated by a point source. By integrating the Green function over a source function, the electromagnetic field as radiated by the source can be found. A major advantage of the use of the Green function for the problem at hand is that the multilayered medium can be incorporated directly in the Green function [53]. When a source is located in a multilayered medium, an integration over the multilayered medium Green function computes the electromagnetic fields in the full multi-layered medium, including the reflections at the layer interfaces. In this way, the computational domain can be limited to the support of the objects in the layered medium. This could potentially allow for a computational domain of a much smaller size than a local method would need.

For a homogeneous medium of infinite extent, Green's function is available in closed form. However, an analytical expression for the multilayered Green's function in the spatial domain does not exist. On the other hand, an analytical expression for the multilayered Green's function exists as a function of the spectral variables conjugate to the coordinates in the transverse plane, i.e. parallel to the layer interfaces. Via a Fourier transformation in the transverse plane, the result of which in this case is also known as a Sommerfeld integral, it is possible to compute the Green function in the spatial domain. The computation of these Sommerfeld integrals [54] is difficult because of branch cuts and poles that are present in the spectral Green function [55, Chapter 8], [56, Chapter 5], [57, Chapter 4], [58, Chapter 2].

Several approaches have been proposed to compute Sommerfeld integrals efficiently. The Green tensor is rotationally invariant in the plane of stratification, a fact which is often exploited by working in cylindrical coordinates (k_ρ, φ). The integral over φ is trivial because of the rotational symmetry, whereas the poles are located in the integral over k_ρ. The poles can then be circumvented by a deformation of the integration path into the complex domain. An important approach that uses such a deformation of the integration path makes use of the steepest descent path (SDP) [55, 59]. This SDP is chosen such that there are as few oscillations in the integral as possible, which enables a fast convergence of the numerical integration. Care has to be taken that the SDP passes some of the poles at the wrong side of the integration contour, which means that their residual contributions must be accounted for in a separate step. Since the locations and residues of the poles are not known analytically, they have to be obtained numerically, which can be cumbersome for general multilayered media.

A different approach to the deformation of the integration path is to choose the deformation much closer to the real coordinate axis [60, 61]. This has the advantage that the path can be chosen such that all poles are circumvented at the same side as the original integration path, and therefore the residues do not have to be added individually. Sometimes the location of the poles is used to create a path that passes around each individual pole [61, 62]. These methods have the disadvantage that the integrand is not as well behaved as on the SDP, so a larger number of quadrature points is needed. However, on the positive side, less or no information about the locations and residues of the poles are needed, which significantly simplifies the whole process and makes it more robust.

A radically different approach to a multilayered medium Green function is the discrete complex image method (DCIM) [63, 64, 65]. The main idea behind this method is that the reflection from a perfectly reflecting layer interface can be modeled by adding a fictitious source on the other side of the interface. It is easy to implement this, for example by adding a displaced Green function to the original Green function. Reflections from dielectric media can be implemented by attenuating the reflection depending on the angle at which the interface is crossed. Multiple layers can be implemented by increasing the number of reflections, which involves bookkeeping and choosing a suitable truncation to the number of reflections. A major advantage is that a free-space code can be employed directly when these reflections are added.

For completeness, it should be noted that Green-function methods are not limited to time-harmonic problems. Time-dependent Green functions exist as well. They also depend on the time difference between when a source emits an electric field and the moment when the electric field is observed. Variants of a time-dependent Green function for multilayered media are reported in [66, 67, 68].

It might seem that the CGFFT method would not generalize well to multi-layered media, since the translation symmetry in the direction normal to the layer interfaces is lost. However, where the integral over the homogeneous-medium Green function can be written as a single convolution, the reflection from the layer stack above and below a scatterer can be added as two correlations added to the homogeneous-medium Green function [69]. In the plane parallel to the layer interfaces, there is still a translation symmetry, so in these directions we can still use a fast FFT-based implementation.

1.3.3 Global methods: Spectral methods for multilayered media

Since the Green function is known analytically in the spectral domain of the transverse plane, it can be advantageous to solve the problem in the transverse-plane spectral domain completely. In principle, the CGFFT method already uses the spectral domain, since the Green function convolution is sped up using FFTs. However, a true spectral method uses testing and basis functions completely in the spectral domain. Consequently, even the shape of the scatterer is transformed to the spectral domain.

An example of such a spectral method is the periodic volume integral method (pVIM), which employs the spectral Green function directly [70, 71]. This method can be employed only for scatterers that are periodic in the transverse direction. The periodicity of the scatterer implies that the electric and magnetic fields are infinitely periodic as well. This periodicity means that the spectral domain decomposes into a set of discrete Floquet modes. Since the spectral domain can be represented accurately by a discrete set of modes, the poles in the Green function are not problematic here, since the chance of a pole coinciding exactly with one of the mode numbers is negligible. In the direction normal to the layer interfaces, a spatial discretization is used, that is handled efficiently using Gohberg and Koltracht's first-order recursion [72]. The field-material interaction in the transverse direction takes the form of a convolution in the spectral domain, which can be computed efficiently using FFTs. This leads to an $O(N \log N)$ matrix vector product.

Another spectral method for periodic scatterers is the rigorous coupled-wave analysis (RCWA), also known as the Fourier modal method [73, 74]. Here, the decomposition in separate modes is used as well. The transmission of the modes in the longitudinal direction, i.e. the direction normal to the layer interfaces, is treated analytically. At every change of the transverse geometry, when moving in the longitudinal direction, this longitudinal transmission changes in shape, and therefore a coupling matrix is computed. This approach leads to a linear system of equations, that is often solved directly. RCWA is very efficient when the number of transverse changes in the longitudinal direction is small, since then the number of coupling matrices for the transmission lines is small as well. However, it can only handle scattering objects with boundaries directed completely in the longitudinal or the transverse direction. When borders are at a slope in both transverse and longitudinal direction they need to be approximated by a staircase approximation [75]. A further disadvantage is that the size of the coupling matrix that needs to be solved increases as $O(N^2)$ for N the number of modes.

A final important spectral method is the C-method [76, 77]. With this method the scattering from a perfectly conducting surface with a periodic modulation is computed. With the aid of tensor calculus, a coordinate transformation is applied to Maxwell's equations in which the surface is aligned with one of the new coordinates. The problem is then solved in this new coordinate system by using a Fourier expansion similar to RCWA. However it allows for more general scattering surfaces, which can be beneficial compared to RCWA [78, 75]. The method has also been generalized to non-conducting materials [79].

An important complication for spectral methods is that spatially discontinuous fields decay slowly in the spectral domain. In principle, this slow decay is not problematic in itself, since most methods make some sort of approximation for a discontinuous field and we are only interested in the accuracy of the radiating part, which means small wavenumbers. However, problems can occur when the gratings we are interested in exhibit a discontinuous dielectric constant function. This discontinuous dielectric contrast means that also the fields are discontinuous for two-dimensional transverse magnetic (TM) and three-dimensional scattering. In the spatial domain, the contrast current density is computed by multiplying the contrast function with the electric field and both are discontinuous at material boundaries. In the spectral domain, such a multiplication is represented by a convolution, and this convolution does not converge uniformly [80], due to the slow decay of discontinuous functions in the spectral domain. In [81] the Li rules are presented, which state that the spectral convolution of two functions that have discontinuities at the same position in the spatial domain, leads to intolerable errors. By a proper reformulation this can be avoided easily for two-dimensional scattering problems [82, 83]. For three-dimensional problems, the electric-field component normal to an interface of discontinuous dielectric constant is discontinuous, but the electric flux density is continuous. However, at such an interface the tangential components of the electric flux density are discontinuous, whereas those of the electric field are continuous. This is exploited in [84], by representing the fields by a combination of electric flux density and electric fields that is continuous everywhere via a so-called normal-vector field formulation.

1.4 Thesis outline

We propose to use a combination of a spatial and a spectral discretization. The method therefore combines strong points of both CGFFT, i.e. the efficient Green function convolution, and strong points of pVIM, i.e., Gohberg and Koltracht's recursion and the exact multilayered Green function. In Chapters 2 and 3 the formulation is presented in which we work. Since the spectral domain does not decompose in discrete modes for a finitely sized (non-periodic) scatterer, a discretization with continuous functions is required. We propose to use the Gabor frame to discretize functions and a short summary of Gabor frames is given in Chapter 4.

In Chapters 2-4 we have set up a guideline along which we devise a full 3D algorithm. However, we start with simpler two-dimensional problems to test whether the approach is feasible for full 3D scattering. We first solve a TE-polarized scattering problem in two dimensions for a homogeneous scatterer in Chapter 5. Since branch cuts are present in the homogeneous Green function, a coordinate stretch in the spectral domain is applied to accurately represent this. Subsequently, in Chapter 6, the TE-scattering problem in a multilayered medium is addressed. Since the multilayered Green function also contains poles, the coordinate stretch is replaced by a deformation of the path of representation from the real line into the complex plane. In Chapter 7, the TM-scattering problem in a multilayered medium is solved, which is the first vectorial scattering problem. A reformulation of the field-material interaction is applied to satisfy the Li rules. This reformulation is extended to normal-vector fields in Chapter 8, where the method is extended to three dimensions in a multilayered medium. Chapter 9 shows the scaling of this method to large simulation domains. Additionally, a simulation result is given for a large scatterer, combined with an optimization for scatterers that are of large longitudinal extent. An improvement of the discretization that partly differs from the Gabor frame is proposed in Chapter 10. This approximation significantly reduces the computation time, while it does not significantly impact the accuracy. The main idea is to change from the Gabor-frame representation, which requires a large number of small-sized FFTs and much overhead, to a list-based representation with a small number of large FFTs and reduced overhead. The results presented in this chapter can be considered as state of the art. Subsequently, in Chapter 11, ideas that came up during the construction of the main algorithm are explored in more detail. The key point here is to put the discretization and the deformation of the spectral representation onto the complex plane into a broader perspective. A Hermite interpolation-based discretization is proposed. Various methods are proposed to transform to and from the complex-plane spectral path, which are applicable to a wider range of paths for representation in the complex spectral domain. Numerical results for 2D TE-polarized scattering are shown for an algorithm based on these ideas. The final chapter, Chapter 12, contains conclusions and an outlook for further improvements on the methods developed in this thesis.

Chapters 5, 6, 7, 8, and 10 are word by word copies of published papers. Hence, they contain abstracts, introductions and conclusions. Additionally, there are slicht deviations in notation in these chapters.

Chapter 2

Formulation

2.1 Description of the geometry

A layered medium consists of a stack of $N-1$ layers of different materials located between two half spaces and stacked in the z-direction. An example of such a medium is drawn in Figure 2.1, together with a Cartesian coordinate system and we indicate the position vector $\mathbf{x} = (x, y, z) = x\hat{x} + y\hat{y} + z\hat{z}$, where symbols with hats, e.g. $\hat{x}$ indicate unit vectors. In this work vectors are distinguishable from scalars by their bold font. We start counting from $n = 0$ in the upper half space, the half space where $z < 0$, towards $n = N$ in the lower half space, where $z > z_N$. Layer n ranges from z_n to z_{n+1}, with thickness d_n and the material has uniform isotropic relative complex permittivity $\varepsilon_{rb,n}$. We assume a uniform permeability μ_0 equal to that of free space and no conductivity.

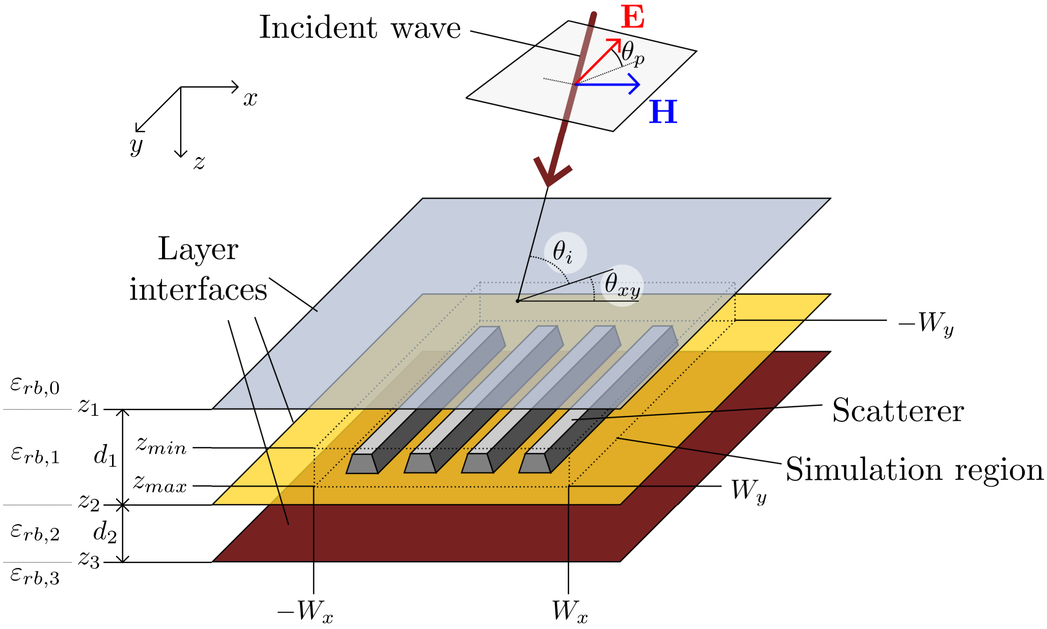


Figure 2.1: A scattering setup for $N = 3$ and a small finite grating as scatterer.

In layer i ($i \in \{1, \cdots, N-1\}$), an object of finite size is embedded. This object is described by the permittivity function $\varepsilon_r(\mathbf{x})$, where $\mathbf{x} = (x, y, z)$ denotes the position vector. The object is contained within the rectangular box $[-W_x, W_x] \times [-W_y, W_y] \times [z_{min}, z_{max}]$, with $z_i \leq z_{min} < z_{max} \leq z_{i+1}$, which we will call the simulation domain D. In this work, the object and simulation domain are assumed to be embedded in a single layer.

2.2 Waves in a multilayer medium

We start from Maxwell's equations to derive equations for waves in a homogeneous medium in a representation, where we use the spectral domain in the xy plane and the spatial domain in the z direction. From there we continue by identifying two polarizations for waves moving through a homogeneous medium in absence of the scatterer. Transmission and reflection coefficients for both polarizations are combined into a single transmission and reflection tensor. We derive tensorial transmission and reflection coefficients, and an expression for the incident field in presence of the multi-layered medium and in absence of the scatterer.

Throughout this thesis we will use the Fourier transformation along an arbitrary coordinate ξ, defined as

$$f(k_\xi) = \mathcal{F}_\xi \left[f(\xi) \right](k_\xi) = \int_{-\infty}^{\infty} d\xi \, f(\xi) e^{-jk_\xi \xi}, \tag{2.1}$$

and with inverse

$$f(\xi) = \mathcal{F}_{k_\xi}^{-1} \left[f(k_\xi) \right](\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk_\xi \, f(k_\xi) e^{jk_\xi \xi}. \tag{2.2}$$

We will only explicitly write down the Fourier transformation of a function when it is needed for clarity. When we write k as the argument of a function, its Fourier transform is intended.

2.2.1 Waves in a homogeneous medium

Let us start from the time (t) dependent Maxwell's equations [85] in the space-time domain as

$$\begin{aligned} \nabla \times \tilde{\mathbf{H}}(\mathbf{x}, t) &= \tilde{\mathbf{J}}(\mathbf{x}, t) + \partial_t \tilde{\mathbf{D}}(\mathbf{x}, t) \\ \nabla \times \tilde{\mathbf{E}}(\mathbf{x}, t) &= -\partial_t \tilde{\mathbf{B}}(\mathbf{x}, t), \end{aligned} \tag{2.3}$$

with electric field $\tilde{\mathbf{E}}$, magnetic flux density $\tilde{\mathbf{B}}$, magnetic field $\tilde{\mathbf{H}}$, electric flux density $\tilde{\mathbf{D}}$, and an electric current density $\tilde{\mathbf{J}}$, which can act as a source. The current density can be divided in a source current density $\tilde{\mathbf{J}}^s$ and a contrast current density $\tilde{\mathbf{J}}^c$, which is induced by the electric field. We assume absence of the source current density, $\tilde{\mathbf{J}}^s = 0$ and therefore we omit the ^{c} superscript in the contrast current density. When we assume illumination by a time-harmonic source with angular frequency ω, all fields will have an $\exp(j\omega t)$ time-dependence, which will be left out in the notation. We can write Maxwell's equations in the space-frequency domain as

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{x},\omega) &= \mathbf{J}(\mathbf{x},\omega) + j\omega\mathbf{D}(\mathbf{x},\omega) \\ \nabla \times \mathbf{E}(\mathbf{x},\omega) &= -j\omega\mathbf{B}(\mathbf{x},\omega). \end{aligned} \tag{2.4}$$

In the absence of free charge, we also have

$$\begin{aligned} \nabla \cdot \mathbf{D}(\mathbf{x},\omega) &= 0 \\ \nabla \cdot \mathbf{B}(\mathbf{x},\omega) &= 0. \end{aligned} \tag{2.5}$$

Since a harmonic time dependence is assumed in a linear system, the ω dependence can be ignored in the notation, since ω is constant throughout.

In a homogeneous isotropic dielectric layer n we have $\varepsilon_r(\mathbf{x}) = \varepsilon_{rb,n}$, which we use in $\mathbf{B} = \mu_0\mathbf{H}$ and $\mathbf{D} = \varepsilon_0\varepsilon_r(\mathbf{x})\mathbf{E}$. After eliminating the magnetic field and the electric flux density, Eq. (2.4) becomes

$$\nabla \times \nabla \times \mathbf{E} = -j\omega\mu_0\mathbf{J} + k_0^2\varepsilon_r(\mathbf{x})\mathbf{E}, \tag{2.6}$$

where $k_0^2 = \omega^2\mu_0\varepsilon_0$ denotes the wavenumber in vacuum and constant $\varepsilon_r(\mathbf{x}) = \varepsilon_{rb,n}$. In the absence of electric current sources, this becomes

$$\begin{aligned} \nabla^2\mathbf{E} + \varepsilon_{rb,n}k_0^2\mathbf{E} &= \mathbf{0} \\ \nabla \cdot \mathbf{E} &= \mathbf{0}, \end{aligned} \tag{2.7}$$

where the second line comes from Eq. (2.5). Solutions to the first line, which is a homogeneous Helmholtz equation, are three-dimensional plane waves

$$\mathbf{E}(\mathbf{x}) = \mathbf{E}^w e^{j\mathbf{k}_w\cdot\mathbf{x}}, \tag{2.8}$$

with complex wave vector $\mathbf{k}_w$ with $\mathbf{k}_w \cdot \mathbf{k}_w = \varepsilon_{rb,n}k_0^2$ and complex amplitude $\mathbf{E}^w$. The dot product is the short-hand notation for $\mathbf{k}_w \cdot \mathbf{x} = xk_{w,x} + yk_{w,y} + zk_{w,z}$ without complex conjugation. The second line in Eq. (2.7) eliminates one degree of freedom in the direction of $\mathbf{E}$, such that $\mathbf{E}^w \cdot \mathbf{k}_w = 0$. This means that there is a two-dimensional freedom of choice for the direction of $\mathbf{E}^w$, which is called the polarization. Since there is translational symmetry in the $x - y$ plane in the multi layered medium we distinguish it by employing the subscript $_T$ to denote the part of a vector in this transverse direction, e.g. $\mathbf{x}_T = x\hat{x}+y\hat{y}$ and $\mathbf{k}_{w,T} = k_{w,x}\hat{x} + k_{w,y}\hat{y}$. The propagation in the z direction is characterized by $\gamma_n^2 = -k_{w,z}^2 = \mathbf{k}_{w,T}^2 - \varepsilon_{rb,n}k_0^2$, with the branch cut of the square root taken just below the negative real axis and $\mathbf{k}_{w,T}^2 = \mathbf{k}_{w,T} \cdot \mathbf{k}_{w,T}$. Note that γ is imaginary for propagating waves. The two solutions corresponding to both Riemann surfaces, i.e. $+\gamma_n$ and $-\gamma_n$, determine the direction of propagation in the z direction. Now the single plane wave of Eq. (2.8) can be written as

$$\mathbf{E}(\mathbf{x}_T, z) = \begin{cases} \mathbf{E}^w e^{j\mathbf{k}_{w,T}\cdot\mathbf{x}_T - \gamma_n z} & \text{for a wave moving up} \\ \mathbf{E}^w e^{j\mathbf{k}_{w,T}\cdot\mathbf{x}_T + \gamma_n z} & \text{for a wave moving down .} \end{cases} \tag{2.9}$$

For waves propagating in the horizontal plane, $\mathbf{k}_w = \mathbf{k}_{w,T}$ both solutions are equal. In general, the electric field $\mathbf{E}$ contains plane waves in all different directions with different amplitudes. When a collection of waves is given, the electric field can be computed at height z_0. Therefore we divide the total electric field $\mathbf{E}(\mathbf{x}_T, z_0)$ in a collection of plane waves moving up $\mathbf{W}^u(\mathbf{x}_{w,T}, z_0)$ and a collection moving down $\mathbf{W}^d(\mathbf{x}_T, z_0)$. These collections can be Fourier transformed to spectral transverse wavenumber $\mathbf{k}_T \in \mathbb{R}^2$, yielding $\mathbf{W}^u(\mathbf{k}_T, z_0)$ and $\mathbf{W}^d(\mathbf{k}_T, z_0)$. Where Eq. (2.9) only contains propagating waves, this can be generalized to include evanescent waves by letting $\mathbf{k}_T \in \mathbb{R}^2$. Therefore, the total electric field can be characterized by the Fourier transform

$$\mathbf{E}(\mathbf{x}_T, z) = \left(\frac{1}{2\pi}\right)^2 \int_{\mathbf{k}_T \in \mathbb{R}^2} d\mathbf{k}_T \left[\mathbf{W}^d(\mathbf{k}_T, z) + \mathbf{W}^u(\mathbf{k}_T, z)\right] e^{j\mathbf{k}_T \cdot \mathbf{x}_T}, \tag{2.10}$$

The integral over the vectorial quantities is defined as

$$\int_{\mathbb{R}^2} d\mathbf{k}_T \mathbf{v}(x) = \int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y \; \left(v_x(k_x, k_y)\hat{x} + v_y(k_x, k_y)\hat{y} + v_z(k_x, k_y)\hat{z}\right). \tag{2.11}$$

The wave amplitudes $\mathbf{W}^u(\mathbf{k}_T, z_0)$ and $\mathbf{W}^d(\mathbf{k}_T, z_0)$ in Eq. (2.10) propagate in the z direction according to Eq. (2.9), so they can be found via

$$\begin{aligned} \mathbf{W}^d(\mathbf{k}_T, z) =& \mathbf{W}^d(\mathbf{k}_T, z_0) e^{-\gamma_n (z - z_0)} \\ \mathbf{W}^u(\mathbf{k}_T, z) =& \mathbf{W}^u(\mathbf{k}_T, z_0) e^{-\gamma_n (z_0 - z)}. \end{aligned} \tag{2.12}$$

This notation for wave propagation in the z direction will be kept throughout the rest of this thesis and by a z-propagating wave we will mean this instead of the plane wave of Eq. (2.8) and Eq. (2.9).

Since the $\mathbf{W}^u$ and $\mathbf{W}^d$ are given in terms of $\mathbf{k}_T$, we fix the remaining degree of freedom in the polarization of the plane waves with respect to $\mathbf{k}_T$ in two polarizations, a transverse electric (TE or e) and a transverse magnetic (TM or h) polarization [86, Chapter 11]. The TE polarization is defined such that the electric field lies in the transverse plane and is perpendicular to $\mathbf{k}_T$. We now look at the three components of the electric field, which together constitute the two field polarizations.

- The electric field component in the z direction is certainly h-polarized.
- The electric field component in the $\mathbf{k}_T$ direction is also h-polarized.
- The electric field component perpendicular to the $\hat{z}, \mathbf{k}_T$-plane must be e-polarized.

It is clear that for a general vector $\mathbf{V}$, V_z is certainly h-polarized. In the transverse plane we can select the h-polarized part from $\mathbf{V}_T$ by the two-dimensional projection operator $\mathcal{P}_h$ defined as

$$\mathcal{P}_h \cdot \mathbf{V}_T = \mathbf{k}_T \frac{\mathbf{k}_T \cdot \mathbf{V}_T}{\mathbf{k}_T^2} \tag{2.13}$$

and the e-polarized part by the projection operator $\mathcal{P}_e$ defined as

$$\mathcal{P}_e \cdot \mathbf{V}_T = (\mathbf{k}_T \times \hat{z}) \frac{(\mathbf{k}_T \times \hat{z}) \cdot \mathbf{V}_T}{\mathbf{k}_T^2}. \tag{2.14}$$

Rank 2 tensors, such as these projection operators can be distinguished by their caligraphic font. We can separate these projections, both polarizations can be separated in wave $\mathbf{W}^u$ through

$$\begin{aligned} \mathbf{W}^{u,h} &= \mathcal{P}_h \cdot \mathbf{W}_T^u \\ \mathbf{W}^{u,e} &= \mathcal{P}_e \cdot \mathbf{W}_T^u + \hat{z} W_z^u, \end{aligned} \tag{2.15}$$

and a similar seperation for $\mathbf{W}^d$ into $\mathbf{W}^{d,h}$ and $\mathbf{W}^{d,e}$. Since the direction of the e and h parts are fixed by $\mathbf{k}_T$, we sometimes use scalars to denote the amplitude of these waves, when this is more convenient for notation.

2.2.2 Reflection and transmission at an interface

For the transmission and reflection at a single layer interface, we take a background medium with $N = 1$, i.e. two half spaces with a single interface at $z = 0$. Let us assume that the upper half space $z < 0$ has dielectric permittivity $\varepsilon_{rb,0}$ and the half space $z > 0$ has relative dielectric permittivity $\varepsilon_{rb,1}$. When an incident wave $\mathbf{S}^d(\mathbf{k}_T, z)$ moving down in half space 0 reaches half space 1, it is partly reflected back into half space 0 as $\mathbf{R}^u(\mathbf{k}_T, z)$ and partly transmitted into layer 1 as $\mathbf{T}^d(\mathbf{k}_T, z)$. The transmission and reflection of a wave through the interface between materials with different permittivities is different for the two polarizations of a wave [85, Chapter 7].

Since the transmission and reflection depend on the polarization, using Eq. (2.15), we start by writing down the equations for the e polarized part, which is scalar, i.e.

$$E^e(\mathbf{k}_T, z) = \begin{cases} S^e(\mathbf{k}_T, z) + R^e(\mathbf{k}_T, z) & \text{when } z < 0 \\ T^e(\mathbf{k}_T, z) & \text{when } z > 0. \end{cases} \tag{2.16}$$

By ensuring Maxwell's equations hold in integral form on the interface at $z = 0$, i.e. by ensuring boundary conditions, reflection and transmission coefficients can be obtained [85] such that

$$\begin{aligned} R^e(\mathbf{k}_T, z) =& r^e(\mathbf{k}_T) S^e(\mathbf{k}_T, 0) e^{\gamma_0 z} \\ T^e(\mathbf{k}_T, z) =& t^e(\mathbf{k}_T) S^e(\mathbf{k}_T, 0) e^{-\gamma_1 z}. \end{aligned} \tag{2.17}$$

Here r^e and t^e signify the reflection and transmission coefficients from the layer interface at $z = 0$ for e polarization. For reflections in the h polarized part we have to look at the H-field, which is directed perpendicular to the $\hat{z} - \mathbf{k}_T$-plane as well for this polarization

$$H^h(\mathbf{k}_T, z) = \begin{cases} S^h(\mathbf{k}_T, z) + R^h(\mathbf{k}_T, z) & \text{when } z < 0 \\ T^h(\mathbf{k}_T, z) & \text{when } z > 0. \end{cases} \tag{2.18}$$

This leads to

$$\begin{aligned}R^h(\mathbf{k}_T, z) =& r^h(\mathbf{k}_T) S^h(\mathbf{k}_T, 0) e^{\gamma_0 z}\\ T^h(\mathbf{k}_T, z) =& t^h(\mathbf{k}_T) S^h(\mathbf{k}_T, 0) e^{-\gamma_1 z}.\end{aligned} \tag{2.19}$$

The transverse part of the electric field corresponding to the magnetic fields R^h, T^h and S^h is identical up to a constant, whereas the z part part gets a minus sign when traveling in the negative z-direction. The exact form of these coefficients be found in [85, Chapter 7].

By means of a transmission tensor $\mathcal{T}(\mathbf{k}_T)$ we can construct a single transmission coefficient that incorporates both polarizations. In a Cartesian coordinate system, the tensor is given by

$$\mathcal{T}(\mathbf{k}_T) = \frac{1}{\mathbf{k}_T^2} \begin{pmatrix} k_x^2 t^h(\mathbf{k}_T) - k_y^2 t^e(\mathbf{k}_T) & k_y k_x t^h(\mathbf{k}_T) - k_y k_x t^e(\mathbf{k}_T) & 0 \\ k_y k_x t^h(\mathbf{k}_T) - k_y k_x t^e(\mathbf{k}_T) & k_y^2 t^h(\mathbf{k}_T) - k_x^2 t^e(\mathbf{k}_T) & 0 \\ 0 & 0 & \mathbf{k}_T^2 t^h(\mathbf{k}_T) \end{pmatrix}, \tag{2.20}$$

and similarly a reflection tensor

$$\mathcal{R}(\mathbf{k}_T) = \frac{1}{\mathbf{k}_T^2} \begin{pmatrix} k_x^2 r^h(\mathbf{k}_T) - k_y^2 r^e(\mathbf{k}_T) & k_y k_x (r^h(\mathbf{k}_T) - r^e(\mathbf{k}_T)) & 0 \\ k_y k_x (r^h(\mathbf{k}_T) - r^e(\mathbf{k}_T)) & k_y^2 r^h(\mathbf{k}_T) - k_x^2 r^e(\mathbf{k}_T) & 0 \\ 0 & 0 & -\mathbf{k}_T^2 r^h(\mathbf{k}_T) \end{pmatrix}, \tag{2.21}$$

where the zz element gets a minus sign, since for a reflection the direction of propagation changes. With the aid of these tensors it is possible to give an equivalent of Eqs. (2.17) and (2.19) for the electric fields

$$\begin{aligned}\mathbf{R}^u(\mathbf{k}_T, z) =& \mathcal{R}(\mathbf{k}_T) \cdot \mathbf{S}^d(\mathbf{k}_T, 0) e^{\gamma_0 z}\\ \mathbf{T}^d(\mathbf{k}_T, z) =& \mathcal{T}(\mathbf{k}_T) \cdot \mathbf{S}^d(\mathbf{k}_T, 0) e^{-\gamma_1 z}.\end{aligned} \tag{2.22}$$

We will refer to these and similar quantities as reflection and transmission coefficients, although they are tensor-valued.

2.2.3 A multi-layered medium

To calculate the electric field throughout a multi-layered medium, such as in Figure 2.1, a similar technique is used as for a single interface. When an incident wave $\mathbf{W}_0^{i,d}(\mathbf{k}_T, z)$ is present in the upper half space, $z < z_1$, the electric field in all layers can be calculated. In every layer of the multi-layered medium waves are traveling in both directions. In layer m there exist a wave traveling up $\mathbf{W}_m^{i,u}(\mathbf{k}_T, z)$ and a wave traveling down $\mathbf{W}_m^{i,d}(\mathbf{k}_T, z)$. Similar to the half space reflection coefficients, it is possible to use the boundary conditions at the layer interfaces to match the amplitudes of all these waves. In [58, Chapter 2] [57, Chapter 4] [56, Chapter 5], [55, Chapter 8] and [87], methods are given to find scalar transmission and reflection coefficients for both polarizations in the form of t_{mn}^d and t_{mn}^u to calculate the transmitted and reflected waves moving down and up respectively in layer n

generated by a source in layer m. By applying the technique that yields Eq. (2.20), both polarizations can be combined into tensorial transmission coefficients $\mathcal{T}^d_{mn}$ and $\mathcal{T}^u_{mn}$. The indices of these transmission coefficients are clarified in Figure 2.2, m signifying the layer where the source is located and n the layer in which the wave is observed. Using these tensorial transmission coefficients, the electric field in abscence of the scatterer, including reflections from the complete layer stack, which we call the incident electric field, can be calculated in layer i as

$$\begin{aligned} \mathbf{W}^{i,d}_n(\mathbf{k}_T, z) =& e^{-\gamma_n(z-z_n)}\mathcal{T}^d_{0n} \cdot W^{i,d}_0(\mathbf{k}_T, z_1) \\ \mathbf{W}^{i,u}_n(\mathbf{k}_T, z) =& e^{-\gamma_n(z_{n+1}-z)}\mathcal{T}^u_{0n} \cdot W^{i,d}_0(\mathbf{k}_T, z_1) \\ \mathbf{E}^i(\mathbf{x}) =& \left(\frac{1}{2\pi}\right)^2 \int_{\mathbf{k}^2_T \in \mathbb{R}^2} d\mathbf{k}_T \left[\mathbf{W}^{i,d}_n(\mathbf{k}_T, z) + \mathbf{W}^{i,u}_n(\mathbf{k}_T, z)\right] e^{j\mathbf{k}_T \cdot \mathbf{x}_T}, \end{aligned} \tag{2.23}$$

for $z_{min} < z < z_{max}$, where we used Eq. (2.10) for the last line. In most cases, the incident wave, $\mathbf{W}^{i,d}_0$ is chosen to be a plane wave, a case in which $\mathbf{W}_0$ becomes a Dirac-delta distribution in the $\mathbf{k}_T$ plane. More general incident fields can also be chosen as long as they are solutions to the Maxwell equations in the upper half space, for example complex source beams and spherical waves originating from a point source. The connection between the tensorial transmission coefficients and the reflection coefficients in e.g. Eq. (2.22) is that for example $\mathcal{T}^u_{i,i}$ can be considered as the reflection coefficient from the interface between layer i and $i + 1$ for a wave moving down.

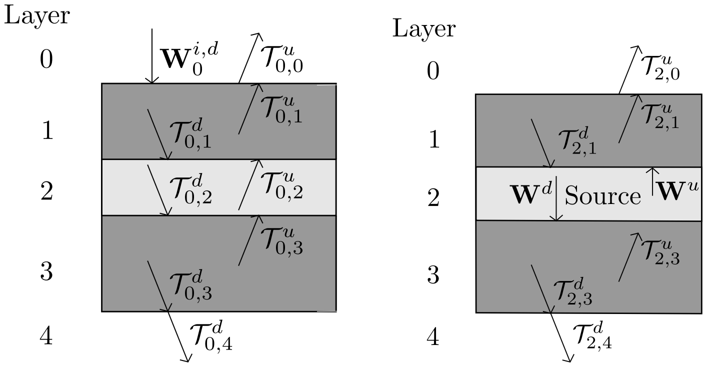


Figure 2.2: Illustration to clarify the definition of the $\mathcal{T}^u_{ij}$ and $\mathcal{T}^d_{ij}$ transmission coefficients that can be calculated with the methods in [87, 55]. In the left figure, the source of $\mathbf{W}^{i,d}_0$ is above the top layer, in the right figure the source is located in layer 2, resulting in both an upgoing $\mathbf{W}^u$ and downgoing $\mathbf{W}^d$ wave.

2.3 The Green function

2.3.1 The spatial integral equation

Until now, we have dealt with waves in the multi-layered medium, without a dielectric object present. In this section we will add a dielectric object to the formulation. From here, the focus will be on fields scattered by the dielectric object in the simulation domain. Therefore, we will sometimes refer to the layered medium as the background medium and the corresponding relative permittivity will be denoted with an additional subscript $_b$, e.g. $\varepsilon_{rb,i}$ for the relative permittivity of the background in layer i.

The incident field $\mathbf{E}^i$ in layer n is understood to contain the reflections of the multilayer medium only, therefore it obeys Eq. (2.6) where $\varepsilon_r(\mathbf{x}) = \varepsilon_{rb,n}$ for $z_n < z < z_{n+1}$ and where $\mathbf{J} = 0$. It satisfies

$$\nabla \times \nabla \times \mathbf{E}^i(\mathbf{x}) = k_0^2 \varepsilon_{rb,n} \mathbf{E}^i(\mathbf{x}). \tag{2.24}$$

We remind the reader that we assume the scattering object to be located in layer i. The electric field $\mathbf{E}(\mathbf{x})$ in layer i in presence of the scattering object satisfies

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{x}) = k_0^2 \varepsilon_r(\mathbf{x})\mathbf{E}(\mathbf{x}) = -j\omega\mu_0 \mathbf{J}(\mathbf{x}) + k_0^2 \varepsilon_{rb,i}\mathbf{E}(\mathbf{x}) + k_0^2(\varepsilon_r(\mathbf{x}) - \varepsilon_{rb,i})\mathbf{E}(\mathbf{x}), \tag{2.25}$$

where we can decompose the electric field as

$$\mathbf{E}(\mathbf{x}) = \mathbf{E}^i(\mathbf{x}) + \mathbf{E}^s(\mathbf{x}), \tag{2.26}$$

with $\mathbf{E}^s$ the scattered field. Subtracting Eq. (2.24) from Eq. (2.25) we find

$$\nabla \times \nabla \times \mathbf{E}^s(\mathbf{x}) - \varepsilon_{rb,i} k_0^2 \mathbf{E}^s(\mathbf{x}) = -j\omega\mu_0 \mathbf{J}(\mathbf{x}) + (\varepsilon_r(\mathbf{x}) - \varepsilon_{rb,i}) k_0^2 \mathbf{E}^s(\mathbf{x}). \tag{2.27}$$

As a more convenient way to describe the scattering object we define the contrast function through

$$\chi(\mathbf{x}) = \frac{\varepsilon_r(\mathbf{x})}{\varepsilon_{rb,i}} - 1. \tag{2.28}$$

Note that this contrast function is nonzero only on the scattering object, so $\chi(\mathbf{x}) = 0$ when $\mathbf{x} \notin D$. Now the $(\varepsilon_r(\mathbf{x}) - \varepsilon_{rb,i})k_0^2\mathbf{E}(\mathbf{x})$ term in Eq. (2.27) can be written in terms of a synthetic current density. We define this contrast current $\mathbf{J}^c(\mathbf{x})$ through

$$\mathbf{J}^c(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb,i}\chi(\mathbf{x})\mathbf{E}(\mathbf{x}). \tag{2.29}$$

This equation is known as the field-material interaction. Since other current density sources are absent, we drop the ^{c} subscript in the notation of the contrast current density and we write

$$\nabla \times \nabla \times \mathbf{E}^s(\mathbf{x}) - \varepsilon_{rb,i} k_0^2 \mathbf{E}^s(\mathbf{x}) = -j\omega\mu_0 \mathbf{J}(\mathbf{x}). \tag{2.30}$$

The scattered field is the field that is emitted by the contrast current and it can be calculated employing the Green tensor $\mathcal{G}$, corresponding to Eq. (2.30). Formally, the scattered field can then be calculated as

$$\mathbf{E}^s(\mathbf{x}) = \int_{\mathcal{D}} \mathbf{dx}\mathcal{G}(\mathbf{x}|\mathbf{x}') \cdot j\omega\varepsilon_0\varepsilon_{rb,i}\chi(\mathbf{x}')\mathbf{E}(\mathbf{x}'). \tag{2.31}$$

Combining this with Eq. (2.26) yields the integral equation

$$\mathbf{E}^{i}(\mathbf{x}) = \mathbf{E}(x) - \int_{\mathcal{D}} \mathbf{dx}\mathcal{G}(\mathbf{x}|\mathbf{x}') \cdot \mathbf{J}(\mathbf{x}'), \tag{2.32}$$

where $\mathbf{J}$ depends on the electric field $\mathbf{E}$ through Eq. (2.29). This is the spatial domain version of the integral equation we wish to solve. We will first deduce a Green tensor $\mathcal{G}$ for the scattering problem with a homogeneous background permittivity $\varepsilon_{rb,i}$. In the subsequent sections we will generalize this to a multi-layered background by including reflections from the layer interfaces.

2.3.2 The Green function in a homogeneous medium

We want to find a Green tensor $\mathcal{G}^{h}$ for the electric field from a current source $\mathbf{J}$ of Eq. (2.30) in layer i in absence of the layered medium. Therefore, we look at the three-dimensional spatial Fourier transform of Eq. (2.30), which results in

$$\mathbf{k}(\mathbf{k} \cdot \mathbf{E}^{s}(\mathbf{k})) + (\varepsilon_{rb,i}k_0^2 - \mathbf{k}^2)\mathbf{E}^{s}\hat{(}\mathbf{k}) = j\omega\mu_0\mathbf{J}(\mathbf{k}), \tag{2.33}$$

with $\mathbf{k} = k_x\hat{x} + k_y\hat{y} + k_z\hat{z}$. We can exploit the symmetry of the homogeneous Green tensor $\mathcal{G}^{h}$, so it suffices to calculate $\mathcal{G}^{h}$ only for a point source $\mathbf{J}$ directed along $\hat{z}$, i.e. $\mathbf{J}(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb,i}\delta(\mathbf{x})\hat{z}$. We will write $\mathbf{G}^{h}_{z} = \mathcal{G}^{h} \cdot \hat{z}$ for the electric field originating from a point source pointed in the z direction and located at the origin, since it is simple to translate the source and the Green function afterwards. We arrive at

$$\mathbf{k}(\mathbf{k} \cdot \mathbf{G}^{h}_{z}(\mathbf{k})) + (\varepsilon_{rb,i}k_0^2 - \mathbf{k}^2)\mathbf{G}^{h}_{z}(\mathbf{k}) = -\varepsilon_{rb,i}k_0^2\hat{z}. \tag{2.34}$$

Clearly, $\mathbf{G}^{h}_{z}$ lies in the plane spanned by $\mathbf{k}$ and $\hat{z}$ and therefore we will decompose it in these two directions. We write $\mathbf{k} = k_T\hat{\mathbf{k}}_T + k_z\hat{z}$ with $\hat{\mathbf{k}}_T = \mathbf{k}_T/|\mathbf{k}_T|$, and use it to decompose this equation in the transverse xy-plane and the z-direction

$$\begin{aligned} \mathbf{k}_T(\mathbf{k}_T \cdot \mathbf{G}^{h}_{z,T}(\mathbf{k}) + k_zG^{h}_{z,z}(\mathbf{k})) + (\varepsilon_{rb,i}k_0^2 - \mathbf{k}_T^2 - k_z^2)\mathbf{G}^{h}_{z,T}(\mathbf{k})) =& \mathbf{0}_T \\ k_z(\mathbf{k}_T \cdot \mathbf{G}^{h}_{z,T}(\mathbf{k}) + k_zG^{h}_{z,T}(\mathbf{k})) + (\varepsilon_{rb,i}k_0^2 - \mathbf{k}_T^2 - k_z^2)G^{h}_{z,z}(\mathbf{k})) =& -\varepsilon_{rb,i}k_0^2. \end{aligned} \tag{2.35}$$

This set of equations is solved by

$$\begin{aligned} \mathbf{G}^{h}_{z,T} &= -\frac{\mathbf{k}_T k_z}{\mathbf{k}^2 - \varepsilon_{rb,i}k_0^2} \\ G^{h}_{z,z}(\mathbf{k}) &= \frac{\varepsilon_{rb,i}k_0^2 - k_z^2}{\mathbf{k}^2 - \varepsilon_{rb,i}k_0^2}. \end{aligned} \tag{2.36}$$

Because of the translation symmetry in the homogeneous background, we can now write the general expression for the Green function as

$$\mathcal{G}^h_{i,j}(\mathbf{k}) = \begin{cases} \dfrac{-k_i k_j}{\mathbf{k}^2 - \varepsilon_{rb,i}k_0^2} & \text{if } i \neq j \\ \dfrac{\varepsilon_{rb,i}k_0^2 - k_i^2}{\mathbf{k}^2 - \varepsilon_{rb,i}k_0^2} & \text{if } i = j, \end{cases} \tag{2.37}$$

where $i, j \in \{x, y, z\}$.

Because the translation symmetry in the z direction will be absent in a multi-layered background, we will perform a Fourier transformation to the spatial z coordinate. We employ the identity

$$\int_{-\infty}^{\infty} dk_z \frac{e^{jk_z z}}{\mathbf{k}^2 - \varepsilon_{rb,i}k_0^2} = \pi \frac{e^{-\gamma|z|}}{\gamma},$$

with $\gamma = \sqrt{k_x^2 + k_y^2 - \varepsilon_{rb,i}k_0^2}$ and the branch cut of the square root chosen just below the negative real axis. We can explicitly write all components of the Green tensor, i.e.

$$\begin{aligned}
G^h_{x,x}(k_x, k_y, z|z') &= (\varepsilon_{rb,i}k_0^2 - k_x^2)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{x,y}(k_x, k_y, z|z') &= (-k_x k_y)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{x,z}(k_x, k_y, z|z') &= (-jk_x\gamma)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{y,x}(k_x, k_y, z|z') &= (-k_y k_x)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{y,y}(k_x, k_y, z|z') &= (\varepsilon_{rb,i}k_0^2 - k_y^2)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{y,z}(k_x, k_y, z|z') &= (-jk_y\gamma)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{z,x}(k_x, k_y, z|z') &= (-jk_x\gamma)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{z,y}(k_x, k_y, z|z') &= (-jk_y\gamma)\frac{e^{-\gamma|z-z'|}}{2\gamma} \\
G^h_{z,z}(k_x, k_y, z|z') &= \left(\gamma^2 - \frac{2\delta(z-z')}{\gamma}\right)\frac{e^{-\gamma|z-z'|}}{2\gamma}.
\end{aligned} \tag{2.38}$$

Once the Green function has been obtained, it can be used to represent the scattered field as

$$\mathbf{E}^{s,h}(\mathbf{k}_T, z) = \int_{z_{min}}^{z_{max}} dz'\, \mathcal{G}^h(\mathbf{k}_T, z|z') \cdot \mathbf{J}(\mathbf{k}_T, z'). \tag{2.39}$$

Note that the contrast current density $\mathbf{J}(x, y, z)$ is nonzero only for $z_{min} \leq z \leq z_{max}$. Since the Green tensor has an absolute value in the exponential function it is convenient to decompose the integration domain according to where $z - z'$ is positive and to where it is negative. This yields two waves, the wave $\mathbf{W}^u(\mathbf{k}_T, z)$ propagates upwards and has its sources below z, and the wave propagating downwards, $\mathbf{W}^d(\mathbf{k}_T, z)$, has its sources above layer z. We write

$$\mathbf{E}^{s,h}(k_x, k_y, z) = \mathbf{W}^u(k_x, k_y, z) + \mathbf{W}^d(k_x, k_y, z) \tag{2.40}$$

with

$$\begin{aligned} \mathbf{W}^u(k_x, k_y, z) &= \begin{cases} 0 & \text{if } z > z_{max} \\ \int_z^{z_{max}} dz'\, \mathcal{G}(k_x, k_y, z|z') \cdot \mathbf{J}(k_x, k_y, z') & \text{if } z \leq z_{max} \end{cases} \\ \mathbf{W}^d(k_x, k_y, z) &= \begin{cases} 0 & \text{if } z < z_{min} \\ \int_{z_{min}}^{z} dz'\, \mathcal{G}(k_x, k_y, z|z') \cdot \mathbf{J}(k_x, k_y, z') & \text{if } z \geq z_{min}. \end{cases} \end{aligned} \tag{2.41}$$

2.3.3 The Green function in a multi-layered medium

It is possible to incorporate the reflections from the layer above and the layer below layer i, in which the simulation domain is located, in the Green function.

In Section 2.2.3, the reflection coefficients for a source in layer 0 are introduced. When a source is located in another layer i, (parts of) a wave can bounce multiple times between the stack of layers above z_i and the stack below z_{i+1}, before it travels away through the layers above and below the interfaces of layer i. To include this effect we calculate effective transmission and reflection coefficients $\mathcal{T}_{i,n}^{u,\mathrm{eff}}(\mathbf{k}_T)$ and $\mathcal{T}_{i,n}^{d,\mathrm{eff}}(\mathbf{k}_T)$, for transmission from a source in layer i observed in layer n, including the bouncing between the interfaces of layer i. Note that, between each bounce in layer i, the wave has traveled distance $2d_i$ and is reflected against both interfaces of layer i, with reflection coefficients $\mathcal{T}_{ii}^u$ and $\mathcal{T}_{ii}^d$ respectively. This is sketched in Figure 2.3.

Assume a wave with e polarization is generated in layer i and bounces back and forth within layer i with scalar reflection coefficients $r_e^d(\mathbf{k}_T) = \mathcal{T}_{ii}^d \cdot \mathcal{P}_e$ against the stack of layers above layer i and similar r_e^u for the reflection from the stack of layers below layer i. Employing these reflection coefficients from the whole stacks above and below layer i ensures that we incorporate all reflections at layer interfaces that are not adjacent to layer i. The effective reflection coefficient can be calculated through

$$\begin{aligned} r_e^{u,\mathrm{eff}}(\mathbf{k}_T) =& r_e^u(\mathbf{k}_T) \sum_{n=0}^{\infty} \left(r_e^u(\mathbf{k}_T) r_e^d(\mathbf{k}_T) e^{-2\gamma_i d_i} \right)^n \\ =& r_e^u(\mathbf{k}_T) \frac{1}{1 - r_e^u(\mathbf{k}_T) r_e^d(\mathbf{k}_T) e^{-2\gamma_i d_i}}, \end{aligned} \tag{2.42}$$

where we used the geometric series to find an analytic expression for the sum. Similarly, an effective reflection coefficient for h polarization can be found. To work with full vector waves such as in Eq. (2.12), this can be generalized to find effective transmission tensors

$\mathcal{T}_{i,n}^{d,\text{eff}}$

$$\mathcal{T}_{i,n}^{d,\text{eff}} = \mathcal{T}_{i,n}^{d} \cdot \\ \begin{pmatrix} \frac{k_x^2 r_h^{d,\text{eff}}(\mathbf{k}_T) - k_y^2 r_e^{d,\text{eff}}(\mathbf{k}_T)}{\mathbf{k}_T^2} & \frac{k_x k_y (r_h^{d,\text{eff}}(\mathbf{k}_T) - r_e^{d,\text{eff}}(\mathbf{k}_T))}{\mathbf{k}_T^2} & 0 \\ \frac{k_y k_x (r_h^{d,\text{eff}}(\mathbf{k}_T) - r_e^{d,\text{eff}}(\mathbf{k}_T))}{\mathbf{k}_T^2} & \frac{k_y^2 r_h^{d,\text{eff}}(\mathbf{k}_T) - k_x^2 r_e^{d,\text{eff}}(\mathbf{k}_T)}{\mathbf{k}_T^2} & 0 \\ 0 & 0 & r_h^{d,\text{eff}}(\mathbf{k}_T) \end{pmatrix}, \tag{2.43}$$

and similarly for $\mathcal{T}_{i,n}^{u,\text{eff}}$, where $\mathcal{T}_{i,n}^{u}$ and $\mathcal{T}_{i,n}^{d}$ define a transmitted wave moving up or down, respectively. The wave generated by sources in layer i will generate an upward-traveling reflection $\mathbf{K}^u(\mathbf{k}_T, z)$ at z_i, the interface between layer $i-1$ and layer i.

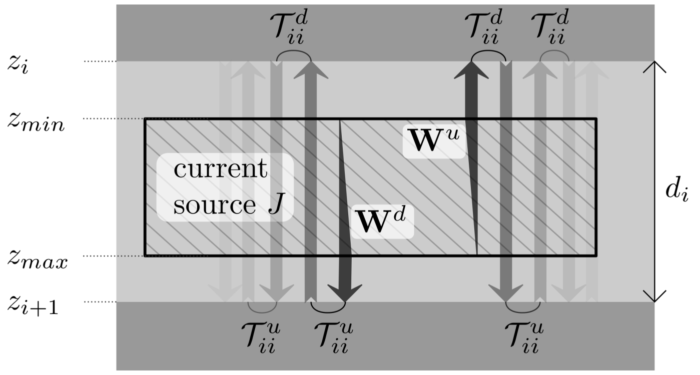


Figure 2.3: The waves, $\mathbf{W}^u$ and $\mathbf{W}^d$, generated by the current source $\mathbf{J}$, are reflected multiple times between z_i and z_{i+1}. The sum of all waves propagating up yields $\mathbf{K}^u(k)$ and the sum of all waves propagating down yields $\mathbf{K}^d(k)$.

To calculate the complete scattered field, the contributions of the reflections have to be added to the scattered field in the homogeneous Green function of Eq. (2.39). We will focus on the downward-directed reflection $\mathbf{K}^d(\mathbf{k}_t, z)$. The downward-directed wave in layer i is composed of the wave moving up $\mathbf{W}^u(\mathbf{k}_T, z_{i+1})$ propagated to the layer interface and the wave moving down $\mathbf{W}^d(\mathbf{k}_T, z_i)$ propagated to the layer interface with $i+1$ and reflected back up again. This yields

$$\begin{aligned} \mathbf{K}^d(\mathbf{k}_T, z) = e^{-\gamma_i(z_i - z)} \mathcal{T}_{ii}^{d,\text{eff}}(\mathbf{k}_T) \cdot [&\mathbf{W}^u(\mathbf{k}_T, z_{max}) e^{-\gamma_i(z_{i+1} - z_{max})} \\ &+ \mathcal{T}_{ii}^{u}(\mathbf{k}_T) \cdot \mathbf{W}^d(\mathbf{k}_T, z_{min}) e^{-\gamma_i(d_i + z_{min} - z_i)}]. \end{aligned} \tag{2.44}$$

A similar expression can be written down for the up-directed wave $\mathbf{K}^u(\mathbf{k}_T, z)$. The scattered electric field at z can now be calculated by adding the homogeneous contributions $\mathbf{W}$ from Eq. (2.41) and the effective reflecting waves $\mathbf{K}^u$ and $\mathbf{K}^d$, i.e.

$$\mathbf{E}^s(\mathbf{k}_T, z) = \mathbf{W}^u(\mathbf{k}_T, z) + \mathbf{W}^d(\mathbf{k}_T, z) + \mathbf{K}^d(\mathbf{k}_T, z) + \mathbf{K}^u(\mathbf{k}_T, z) \tag{2.45}$$

By using the transmission coefficients $\mathcal{T}_{i,n}(\mathbf{k}_T)$, it is possible to propagate the scattered waves to other layers. This makes it possible to treat problems with scattering objects in multiple layers.

The Green tensor for a layered medium computes the scattered electromagnetic field from the contrast current source. Eqs (2.40) and Eq. (2.41) are employed to compute the scattering from the homogeneous-medium Green funtion $\mathcal{G}^h$. From these results, the reflections from the multilayered medium can be computed in Eq. (2.44) and added to the result with the homogeneous Green tensor in Eq. (2.45). We call the concattenation of all these operations on the contrast current source the result of the layered medium Green tensor $\mathcal{G}$.

2.4 A spatial spectral integral equation

We conclude this chapter by defining a spatial spectral integral equation which will be the basis of the rest of this thesis. After stating the equations, we will relate this method to some other important integral equation and spectral methods.

It is the combination of Eq. (2.45), which is given in the spectral domain and Eq. (2.29), which is given in the spatial domain, that defines the integral equation. When Eq. (2.45) is summarized into a single multi-layered Green tensor $\mathcal{G}(\mathbf{k}_t, z|z')$, then the integral equation can be written as

$$\begin{aligned}\mathbf{E}^i(\mathbf{x}) =& \mathbf{E}(x) - \mathcal{F}^{-1}_{\mathbf{k}_T}\left[\int_{z_{min}}^{z_{max}} dz'\ \mathcal{G}(\mathbf{k}_x, z|z') \cdot \mathcal{F}_{\mathbf{x}'_T}\left[\mathbf{J}(\mathbf{x}'_T, z')\right](\mathbf{k}_T)\right](\mathbf{x}_T) \\ \mathbf{J}(\mathbf{x}) =& j\omega\varepsilon_0\varepsilon_{rb,i}\chi(\mathbf{x})\mathbf{E}(\mathbf{x}),\end{aligned} \tag{2.46}$$

where we explicitly wrote down the Fourier transforms that are part of the core of this algorithm.

Compared to the more conventional spatial integral equation of Eq. (2.32), this spatial spectral formulation several differences are noticable.

1. A Green tensor as a function of the $\mathbf{k}_T$ and z coordinate is used.

2. An integration is performed in the z direction only.

3. The convolution in the transverse plane has become a pointwise multiplication.

4. Fourier transformations are added in the transverse direction.

Point by point we will briefly compare these aspects to the literature.

An exact expression for the Green function in the spatial domain does not exist. Through the use of Sommerfeld integrals [54], the spatial domain integral equation can be calculated as explained in [58, Chapter 2] [57, Chapter 4] [56, Chapter 5], [55, Chapter 8]. In our spatial spectral formulation these tedious Sommerfeld integrals are avoided.

The second point, the remaining integral in the z-direction can be carried out efficiently by using the method by Gohberg and Koltracht [72]. This is not the only method, convolutions such as this one could also be carried out efficiently using CGFFT [37] in one direction. However CGFFT is less flexible in the sense that an equidistant sampling is needed in the

z direction and calculating FFTs. In principle the Gohberg method can be extended simply to a non-equidistant sampling in the z direction, which can be advantageous when the electric field contains singularities. It should be added that non-equidistantly sampled extensions of the CGFFT method exist as well, such as AIM [42]. However, it is doubtful whether such a complicated construction yields a faster method than the simple recursive z convolution of the Gohberg method.

The third point, the fact that the convolution in the transverse plane becomes a pointwise multiplication is also exploited in e.g. CGFFT [37] in a homogeneous background medium and in pVIM [70, 71, 88] for a periodically repeating scatterer in a multi-layered medium. For a multi-layered medium the Green function contains poles, that are hard to discretize with an equidistant sampling. In the case of CGFFT the poles were absent and in the case of pVIM the poles were present, but because of the periodic nature of the problem the spectral domain falls apart into discrete modes, and the chance that a mode is located exactly at a pole is negligible. This however, is not the case for finite scatterers and the poles in the Green function are an important problem. A significant part of this thesis will be devoted to an efficient representation that evades these poles.

To address the fourth point, Fourier transformations are also present in CGFFT and pVIM. There is a subtle difference between the FFTs in CGFFT, where the FFTs are used simply as a way to improve the speed of the spatial convolution with the Green function in a problem that is discretized exclusively in the spatial domain. Similar to that, in pVIM the discretization is exclusive in the spectral domain and FFTs are used as a way to improve the speed of the spectral convolution in the field-material interaction. In the present method the Fourier transformation should transform accurately and fast between a discretization in the spectral domain and a discretization in the spatial domain.

2.5 Two-dimensional scattering

When a plane wave is incident on a structure with translation symmetry in the y direction, the scattering problem can be written as a two-dimensional one. We will use two-dimensional formulations as testcases for the mathematical methods we develop. However, two-dimensional methods are also useful for practical application because of the greatly reduced complexity and computation time compared to full three-dimensional problems.

From the Green tensor in Eq. (2.38), we can deduce the Green function for the special cases in two dimensions for TE (e) and TM (h) polarization, when we assume translation symmetry in the y direction i.e. χ, $\mathbf{J}$ and $\mathbf{E}$ are functions of x and z only, independent of y. This symmetry also implies that $k_y = 0$, since the electric field does not depend on y. From $k_y = 0$ in Eq. (2.38) and Eq. (2.43) we can deduce that

$$\mathcal{G}_{2D}(k_x, z|z') = \begin{pmatrix} G_{xx}(k_x, 0, z|z') & 0 & G_{xz}(k_x, 0, z|z') \\ 0 & G_{yy}(k_x, 0, z|z') & 0 \\ G_{zx}(k_x, 0, z|z') & 0 & G_{zz}(k_x, 0, z|z') \end{pmatrix}. \tag{2.47}$$

For TE polarization, the incident field points in the y direction. Clearly the two-dimensional Green function of Eq. (2.47) yields a scattered field pointing in the y direction for a contrast current density pointing in the y direction. Therefore all quantities are pointing in the y direction. Only $\mathcal{G}_{y,y}$ generates the electric field and the TE-polarized Green function is given by the scalar function

$$G_{TE}(k_x, z|z') = \varepsilon_{rb,i} k_0^2 \frac{e^{-\gamma_i|z-z'|}}{2\gamma_i}. \tag{2.48}$$

With a similar argument we observe that for TM polarization, where $\mathbf{H}$ is pointing purely in the y direction, $\mathbf{E}$ and $\mathbf{J}$ point exclusively in the xz-plane. So for TM polarization the Green tensor can be written as the two-dimensional tensor

$$\mathcal{G}_{TM}(k_x, z|z') = \left(\begin{array}{cc} G_{xx}(k_x, 0, z|z') & G_{xz}(k_x, 0, z|z') \\ G_{zx}(k_x, 0, z|z') & G_{zz}(k_x, 0, z|z') \end{array} \right). \tag{2.49}$$

Chapter 3

Solution strategy

3.1 Normal-vector fields

3.1.1 Accuracy and discontinuous functions

The purpose of this work is to present a fast iterative solver, that should work well for scattering from discontinuous objects. To be fast, the result should converge well to the desired accuracy with respect to all simulation parameters that influence the calculation time. One of the most important of such parameters is the spectral range k_{max} at which the spectral domain $\mathbf{k}_T \in [-k_{max}, k_{max}]^2$ is truncated. Especially for discontinuous scatterers, this truncation is important, since the spectral domain representation of the then discontinuous electric field and contrast function converge slowly to zero as $O(1/k_{max})$, whereas we would like a faster convergence of the final result. We recap the equations that we want to solve in the electric field integral equation (EFIE) in Eqs. (2.45), (2.29):

$$\begin{aligned} \mathbf{E}^s(k_x, k_y, z) =& \int dz' \mathcal{G}(k_x, k_y, z|z') \cdot \mathbf{J}(k_x, k_y, z') \\ \mathbf{J}(x, y, z) =& j\omega\epsilon_0\epsilon_{rbi}\chi(x, y, z)\mathbf{E}(x, y, z). \end{aligned} \tag{3.1}$$

In the first line a truncation to k_{max} does not yield many problems. A truncation of the contrast current density does not have a large effect on the calculated electric field, since radiation only occurs for $\mathbf{k}_T^2 < \varepsilon_{rb,i}k_0^2$, for larger $\mathbf{k}_T$ a truncation will produce an error in the electric field locally only. It is important that the contrast current density is correct for small $\mathbf{k}_T$, since this error would radiate all over the scatterer. However, the truncation of $\mathbf{J}$ or $\mathcal{G}$ at k_{max} both do not influence this directly. The local error of a truncated discontinuity will only generate a radiating error when it is close to another discontinuity in the dielectric scatterer, so both local errors can 'mix'.

There is an inconvenience in the field-material interaction (second line), i.e. when χ exhibits a discontinuity, also $\mathbf{E}$ exhibits a discontinuity at the same position. It was shown by Lifeng Li in [80, 81] and others [83, 82] that multiplying two functions with a spatial discontinuity at the same position represented in the spectral domain by Floquet modes gives rise to truncation errors at low k_x and of $O(1/k_{max})$. This problem is not unique for representations using Fourier modes; in the next section we show that a slow convergence is also observed with a *continuous* spectral-domain representation.

3.1.2 Multiplication of discontinuous functions in the spectral domain

Suppose two pulse functions $f_1(x) = U(x+b) - U(x-a)$ and $f_2(x) = U(x+a) - U(b-X)$ are to be multiplied, with U the unit (Heaviside) step function and with a and b real numbers, with $0 \leq a << b$, where the b cutoff is only included to make the Fourier integrals converge. In the spatial domain this multiplication simply yields $m(x) = f_1(x)f_2(x) = U(x+a)U(a-x)$ without problems. In the spectral domain these functions are represented by

$$\begin{aligned} f_1(k) &= -\frac{j}{k}\left(-e^{-jka} + e^{jkb}\right) \\ f_2(k) &= -\frac{j}{k}\left(e^{jka} - e^{-jkb}\right). \end{aligned}$$

The spectral representation of their spatial product can be represented by a convolution

$$\begin{aligned} m(k) &= \int_{-\infty}^{\infty} dk'\, f_1(k-k')f_2(k') \\ &= -\int_{-\infty}^{\infty} dk'\, \frac{1}{k'(k-k')}\left(-e^{-ja(k-k')} + e^{jb(k-k')}\right)\left(e^{jak'} - e^{-jbk'}\right). \end{aligned} \tag{3.2}$$

Since it just represents a pulse function, this integral evaluates to $-(j/k)[-\exp(-jak) + \exp(jak)]$. We will not analytically solve this integral, since we are interested in the numerical convergence, not the result. Several features should be noted about this integral.

- The denominator has zeros, but the numerator has them at the same positions. The integrand is therefore well-behaved.
- For $k' \to \pm\infty$, the integrand is bounded by $1/k'^2$, so it converges.
- The value of b should not influence the result of the exact integral directly, since the product of $f_1(x)$ and $f_2(x)$ is governed by the choice of a.

Now we are interested to see what happens when a approaches zero. In that case the pulse-functions no longer overlap, so their product is simply zero. This holds for the convolution when $k_{max} \to \infty$ in

$$\begin{aligned} &m(k) = 0 = \int_{-k_{max}}^{k_{max}} dk'\, \frac{-\left(-1 + e^{jb(k-k')}\right)\left(1 - e^{-jbk'}\right)}{k(k-k')} \\ &= \int dk'\, \frac{1 - \cos bk' + \cos b(k-2k') + \cos b(k-k') - j\{\sin bk' - \sin b(k-2k') + \sin b(k-k')\}}{k'(k-k')}. \end{aligned} \tag{3.3}$$

However, the convergence rate of this integral to large k_{max} is poor. The numerator is a periodic function, that averages to 1 through a period of length $2\pi/b$ and whose frequency can be chosen very large by varying b, since b does not influence the result. The parameter b is merely chosen to make the spectral respresentation continuous. Therefore, it is the denominator that governs the convergence of this integral for $a = 0$. Note that for nonzero a the numerator in Eq. (3.2) averages to zero and therefore the integral converges faster.

Since we truncate the integral at k_{max}, we would like to see a fast convergence of the integral, but this cannot be expected because the convergence is governed by the denominator which decays as $1/k_{max}^2$, so its integral converges as $1/k_{max}$ over the whole range of k as is illustrated in Figure 3.1. In the figure it is clearly visible that there is a significant contribution in $m(k)$ around $k = 0$, which converges as $O(1/k_{max})$, which is not fast enough for an efficient algorithm. Clearly, this is only the case for $a = 0$. For nonzero a the convergence is much faster, since the convergence is not governed by the denominator only, but also by the numerator in Eq. (3.2), which averages at zero then. We conclude that Lifeng Li's results [81] for a Fourier series representation can be generalized to a continuous spectral representation: the multiplication of functions with spatial discontinuites at the same position in a continuous spectral-domain representation yields poor convergence as a function of the spectral domain truncation parameter k_{max}.

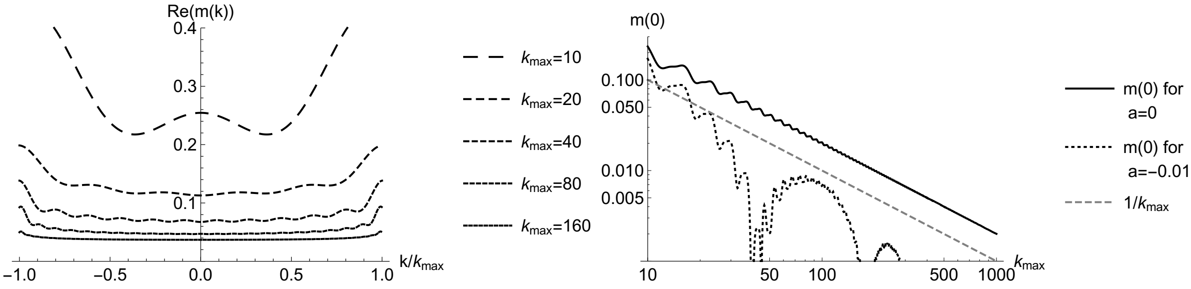


Figure 3.1: Illustration of the poor convergence of $m(k)$ for $a = 0$ and $b = 3$. In (a) over the whole range from $-k_{max}$ to k_{max} for $Re[m(k)]$ and in (b) the convergence for $m(0)$ as a function of k_{max}. The integrals have been computed numerically.

3.1.3 Projections

The poor convergence in the preceding section can be circumvented through a reformulation [83, 82] such as using normal-vector fields [84, 89, 71]. The key observations are that the electric field $\mathbf{E}$ normal to the boundaries of dielectric objects is discontinuous and that the electric flux density $\mathbf{D}$ tangential to these boundaries is also discontinuous. On the other hand, the electric field tangential to the boundaries is continuous as is the electric flux density normal to the boundaries.

By writing $\mathcal{P}_D$ as a projection operator selecting only the components of vectors normal to the object boundaries and $\mathcal{P}_E$ a projection operator selecting only components parallel to the object boundary with $\mathcal{P}_D + \mathcal{P}_E = Id$ we can construct the continuous auxiliary field

F through

$$\mathbf{F} = \mathcal{P}_E \cdot \mathbf{E} + \frac{1}{\varepsilon_0}\mathcal{P}_D \cdot \mathbf{D}. \tag{3.4}$$

Note that there is much freedom in choosing these projection operators, as they are only fixed around the object boundary. Following the normal-vector field theory developed in [70, 71, 90, 91], the electric field can be calculated with the operator $\mathcal{C}_\varepsilon$ defined by

$$\mathbf{E} = \mathcal{C}_\varepsilon \cdot \mathbf{F} = \left(\mathcal{P}_E + \frac{1}{\varepsilon_0}\frac{1}{1+\chi}\mathcal{P}_D\right) \cdot \mathbf{F} = \mathbf{F} - \frac{1}{\varepsilon_0}\frac{\chi}{1+\chi}\mathcal{P}_D \cdot \mathbf{F}, \tag{3.5}$$

where, although formally equal, the rightmost definition is easier to implement, since $\chi/(1+\chi)$ is zero outside the support of the contrast. Similarly, the contrast current density **J** can be calculated through the operator $[\chi\mathcal{C}_\varepsilon]$

$$\mathbf{J} = [\chi\mathcal{C}_\varepsilon] \cdot \mathbf{F} = j\omega\varepsilon_0\varepsilon_{rb,i}\left(\chi\mathcal{P}_E + \frac{1}{\varepsilon_0}\frac{\chi}{1+\chi}\mathcal{P}_D\right) \cdot \mathbf{F} \tag{3.6}$$

We rewrite the EFIE from Eq. (3.1), to combination of terms that individually converges better in the spectral domain, i.e.

$$\begin{aligned} &\mathbf{E}^i(\mathbf{r}_T, z) = \mathcal{C}_\varepsilon(\mathbf{r}_T, z) \cdot \mathbf{F}(\mathbf{r}_T, z) - \\ &\mathcal{F}^{-1}_{\mathbf{k}_T}\left[\int_{z_{min}}^{z_{max}} dz'\, \mathcal{G}(\mathbf{k}_T, z|z')\mathcal{F}_{\mathbf{r}'_T}\left[[\chi\mathcal{C}_\varepsilon](\mathbf{r}'_T, z') \cdot \mathbf{F}(\mathbf{r}'_T, z')\right](\mathbf{k}_T, z')\right](\mathbf{r}_T, z). \end{aligned} \tag{3.7}$$

3.1.4 Example: normal-vector fields on an aligned brick

As an illustration we will show how we have implemented the normal-vector fields for a rectangular bar aligned with the Cartesian coordinate system. Let the bar be centered around the origin and have sides in the x, y, and z-direction of length $2L_x$, $2L_y$, and $2L_z$ respectively.

We only use the normal-vector fields in the xy-plane. In the z-direction the discontinuity can be incorporated in the discretization, since we use a purely spatial discretization in that direction. We base the projection operators on the normal-vector field $\mathbf{n}(x, y)$ defined through

$$\mathbf{n}(x, y) = \begin{cases} \mathbf{n}(x, y) = \hat{\mathbf{y}} & \text{if } |x|/L_x < |y|/L_y \\ \mathbf{n}(x, y) = \hat{\mathbf{x}} & \text{if } |y|/L_y < |x|/L_x \end{cases}. \tag{3.8}$$

Now the projection operator $\mathcal{P}_D$ can in general be written as

$$\frac{1}{\varepsilon_0}\mathcal{P}_D(x, y) = \begin{pmatrix} n_x^2(x, y) & n_x(x, y)n_y(x, y) \\ n_x(x, y)n_y(x, y) & n_y^2(x, y) \end{pmatrix}. \tag{3.9}$$

Note that for this particular normal-vector field the off-diagonal components are zero. When the normal-vector field has been chosen, (3.5) and (3.6) can be used to calculate the field-material interactions. In Figure 3.2, we show an example of the $[C_\varepsilon](x, y)$ and $[\chi C_\varepsilon](x, y)$ operators. Since only the diagonal components are nonzero for this choice of normal vector field, we only show their magnitude in the xy plane. The light gray areas in the xx and yy components corresponds to regions where $\mathbf{D}$ is used in auxiliary field F_x and F_y respectively. The advantage of choosing this particular normal vector field is that it can be composed from two-dimensional step functions. These step functions, $H_{\mathbf{r}_1,\mathbf{r}_2}(x, y)$, are defined as 1 left of a line through coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$ and 0 right of the line, or

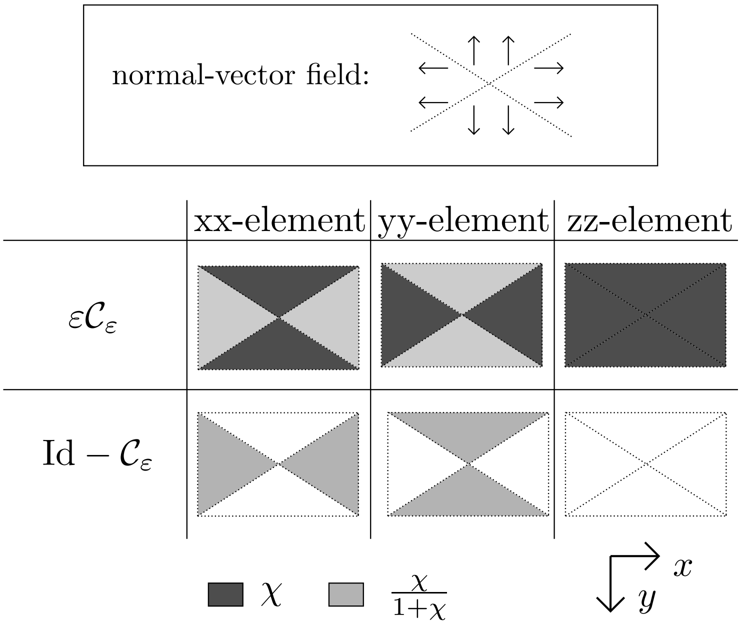


Figure 3.2: The diagonal elements of the $\varepsilon\mathcal{C}_\varepsilon$ and the $\mathcal{C}_\varepsilon$ operator for the normal-vector field proposed in Eq. (3.8). Note that the other elements of these tensors are equal to zero.

$$H_{\mathbf{r}_1,\mathbf{r}_2}(x, y) = \begin{cases} 1 \text{ if } \hat{z} \times (\mathbf{r}_2 - \mathbf{r}_1) \cdot \mathbf{x} > \hat{z} \times (\mathbf{r}_2 - \mathbf{r}_1) \cdot \mathbf{r}_1 \\ 0 \text{ if } \hat{z} \times (\mathbf{r}_2 - \mathbf{r}_1) \cdot \mathbf{x} < \hat{z} \times (\mathbf{r}_2 - \mathbf{r}_1) \cdot \mathbf{r}_1. \end{cases} \tag{3.10}$$

Using only step functions is of significance, because it is relatively easy to accurately approximate such a function in an arbitrary discretization. Although more continuous normal-vector fields than the one in Eq. (3.8) can be used as well, these fields will often exhibit singularities at the corners of the bar, which are much harder to discretize. In Appendix A we show how to discretize $H_{\mathbf{r}_1,\mathbf{r}_2}$.

3.2 Discretization in the z direction

We carry out a spatial discretization in the z direction that employs triangular or piecewise-linear (PWL) functions Λ_n as expansion functions, which are defined as

$$\Lambda_n(z) = \begin{cases} 1 - \frac{|z-n\Delta-z_{min}|}{\Delta} & \text{if} \quad |z-n\Delta-z_{min}| < \Delta \\ 0 & \text{if} \quad |z-n\Delta-z_{min}| > \Delta \end{cases}. \tag{3.11}$$

For notational convenience, we will assume that $z_{min} = -\Delta$ and $z_{max} = N_z\Delta$ throughout this section, so $n = 1$ coincides with $z = 0$. We choose a piecewise-linear discretization because the field $\mathbf{F}$ is a continuous function. When we expand $\mathbf{F}(\mathbf{k}_T, z)$ in these PWL functions we find

$$\mathbf{F}(\mathbf{k}_T, z) \approx \sum_{n=1}^{N_z} \mathbf{F}_n(\mathbf{k}_T)\Lambda_n(z). \tag{3.12}$$

We use Dirac-delta testing functions in the z direction at the points $z_{min} + n\Delta$ to find the coefficients $\mathbf{F}_n(\mathbf{k}_T)$, since it was observed that this leads to a well-conditioned linear system [92], i.e.

$$\mathbf{F}_n(\mathbf{k}_T) = \int_{\Delta}^{N_z\Delta} dz\ \delta(z-n\Delta)\mathbf{F}(\mathbf{k}_T, z). \tag{3.13}$$

We use the same method to generate $\mathbf{E}^i_n(\mathbf{k}_T)$ from $\mathbf{E}^i(\mathbf{k}_T, z)$. By writing the discretized contrast current as $\mathbf{J}_n(\mathbf{k}_T) = \mathbf{J}(k_T, z_n)$, the homogeneous Green function convolution in z in Eq. (2.39) can be rewritten as

$$\begin{aligned}
\mathbf{E}^{s,h}_n(\mathbf{k}_T) =& \int_{z_{min}}^{z_{max}} dz' \sum_{n'=1}^{N_z} \mathcal{G}^h(\mathbf{k}_T, n\Delta|z') \cdot \mathbf{J}_{n'}(\mathbf{k}_T)\Lambda_{n'}(z') \\
=& \hat{z}J_{z,n}(\mathbf{k}_T) \\
&+ \sum_{n'=2}^{n} \int_{(n'-1)\Delta}^{n'\Delta-} dz'\mathcal{G}^h(\mathbf{k}_T, n|z') \cdot \mathbf{J}_{n'}(\mathbf{k}_T)\Lambda_{n'}(z') \\
&+ \sum_{n'=1}^{n-1} \int_{n'\Delta+}^{(n'+1)\Delta} dz'\mathcal{G}^h(\mathbf{k}_T, n|z') \cdot \mathbf{J}_{n'}(\mathbf{k}_T)\Lambda_{n'}(z') \\
&+ \sum_{n'=n+1}^{N_z} \int_{(n'-1)\Delta}^{n'\Delta-} dz'\mathcal{G}^h(\mathbf{k}_T, n|z') \cdot \mathbf{J}_{n'}(\mathbf{k}_T)\Lambda_{n'}(z') \\
&+ \sum_{n'=n}^{N_z-1} \int_{n'\Delta+}^{(n'+1)\Delta} dz'\mathcal{G}^h(\mathbf{k}_T, n|z') \cdot \mathbf{J}_{n'}(\mathbf{k}_T)\Lambda_{n'}(z').
\end{aligned}$$

Here we used the notation $\int_a^{b-} = \lim_{\xi\uparrow b}\int_a^{\xi}$ and similarly $\int_{a+}^{b} = \lim_{\xi\downarrow a}\int_{\xi}^{b}$, to avoid the delta function in the zz-component of the homogeneous Green tensor Eq. (2.38). This zz component is taken into account separately in the $\hat{z}J_{z,n}(\mathbf{k}_T)$ term. By taking into account the z-dependence of $\mathcal{G}^h$ and Λ the above can be written in the form

$$\begin{aligned} E_n^{s,h} =& \hat{z} J_{z,n}(\mathbf{k}_T) \\ &+ \sum_{n'=2}^{n} \mathbf{a}_n^{u,m}(\mathbf{k}_T) e^{-\gamma\Delta(n-n')} + \sum_{n'=1}^{n-1} \mathbf{a}_n^{u,e}(\mathbf{k}_T) e^{-\gamma\Delta(n-n'-1)} \\ &+ \sum_{n'=n+1}^{N_z} \mathbf{a}_n^{d,e}(\mathbf{k}_T) e^{-\gamma\Delta(n-n'-1)} + \sum_{n'=n}^{N_z-1} \mathbf{a}_n^{u,m}(\mathbf{k}_T) e^{-\gamma\Delta(n-n')} \end{aligned} \tag{3.14}$$

where we define $\mathbf{a}_n^{\alpha,\beta}(\mathbf{k}_T)$ with $\alpha = u$ or d (up or down) and $\beta = m$ or e (middle or end) by

$$\begin{aligned} \mathbf{a}_n^{u,m}(\mathbf{k}_T) &= \int_{-\Delta}^{0} dz' \; \mathcal{G}(\mathbf{k}_T, 0|z') \cdot \mathbf{J}_n(\mathbf{k}_T)\Lambda_0(z') \\ \mathbf{a}_n^{u,e}(\mathbf{k}_T) &= \int_{-\Delta}^{0} dz' \; \mathcal{G}(\mathbf{k}_T, 0|z') \cdot \mathbf{J}_{n-1}(\mathbf{k}_T)\Lambda_{-1}(z') \\ \mathbf{a}_n^{d,m}(\mathbf{k}_T) &= \int_{0}^{\Delta} dz' \; \mathcal{G}(\mathbf{k}_T, 0|z') \cdot \mathbf{J}_n(\mathbf{k}_T)\Lambda_0(z') \\ \mathbf{a}_n^{d,e}(\mathbf{k}_T) &= \int_{0}^{\Delta} dz' \; \mathcal{G}(\mathbf{k}_T, 0|z') \cdot \mathbf{J}_{n+1}(\mathbf{k}_T)\Lambda_1(z') \end{aligned} \tag{3.15}$$

This set of expressions can be summarized as

$$\mathbf{a}_n^{u/d,\alpha} = h^\alpha(\mathbf{k}_T) \begin{pmatrix} k_0^2 - k_x^2 & -k_x k_y & \pm i k_x \gamma \\ -k_x k_y & k_0^2 - k_y^2 & \pm i k_y \gamma \\ \pm i k_x \gamma & \pm i k_y \gamma & k_0^2 + \gamma^2, \end{pmatrix} \cdot \begin{cases} \mathbf{J}_n(\mathbf{k}_T) & \text{if } \alpha = m \\ \mathbf{J}_{n\pm 1}(\mathbf{k}_T) & \text{if } \alpha = e, \end{cases} \tag{3.16}$$

, where

$$\begin{aligned} h^m(\mathbf{k}_T) &= \int_0^\Delta dz' \, (1 - \frac{z}{\Delta}) \frac{e^{-\gamma z}}{2\gamma k_0^2} = \frac{e^{-\gamma\Delta} - 1 + \Delta\gamma}{2\Delta\gamma^3 k_0^2} \\ h^e(\mathbf{k}_T) &= \int_0^\Delta dz' \, \frac{z}{\Delta} \frac{e^{-\gamma z}}{2\gamma k_0^2} = \frac{(e^{\gamma\Delta} - 1 - \Delta\gamma) e^{-\gamma\Delta}}{2\Delta\gamma^3 k_0^2}. \end{aligned} \tag{3.17}$$

We can interpret the second and third terms in the right hand side of Eq. (3.14) as waves traveling upwards and the fourth and fifth as waves traveling downwards. When we write the results of these sums as $\mathbf{W}_n^u(\mathbf{k}_T)$ and $\mathbf{W}_n^d(\mathbf{k}_T)$, this equation reduces to

$$\mathbf{E}_n^{s,h}(\mathbf{k}_T) = \hat{z} J_{z,n}(\mathbf{k}_T) + \mathbf{W}_n^u(\mathbf{k}_T) + \mathbf{W}_n^d(\mathbf{k}_T) \tag{3.18}$$

where the $\mathbf{W}_n(\mathbf{k}_T)$ can be calculated in a recursive manner by employing the relations

$$\begin{aligned} \mathbf{W}_n^u(\mathbf{k}_T) =& \mathbf{W}_{n-1}^u(\mathbf{k}_T) e^{\gamma\Delta} + \mathbf{a}_n^{u,e}(\mathbf{k}_T) + \mathbf{a}_n^{u,m}(\mathbf{k}_T) \\ \mathbf{W}_n^d(\mathbf{k}_T) =& \mathbf{W}_{n+1}^d(\mathbf{k}_T) e^{\gamma\Delta} + \mathbf{a}_n^{d,e}(\mathbf{k}_T) + \mathbf{a}_n^{d,m}(\mathbf{k}_T) \end{aligned} \tag{3.19}$$

as I. Gohberg and I. Koltracht pointed out in [72]. This removes the need for repeatedly computing the full sum in Eq. (3.14). To arrive at the complete scattered field $\mathbf{E}_n^s(\mathbf{k}_T)$, the reflections should be added according to Eq. (2.45).

3.3 Discretization in the transverse direction

The main challenge for the formulation in Chapter 2 is the discretization in the transverse (x and y) directions. The complete integral equation consists of both a contrast multiplication and a Green-tensor convolution. The contrast multiplication has to be performed in the spatial domain, whereas the Green-tensor convolution is handled more efficiently in the spectral domain. Since both the spatial and the spectral domain are used, it is important to be able to use fast methods to transform between the domains, e.g. based on FFTs. However, FFTs are not applicable directly to continuous functions. Hence, there is a need for a discretization that allows for an efficient Fourier transformation between the two domains and is parsimonious in memory consumption. As an example, consider the PWL discretization in Section 3.2. By definition, the PWL functions have a discontinuous derivative, even when a $C^{\infty}(\mathbb{R})$ function is discretized with the aid of PWL functions, its continuity is reduced to C^1 and its Fourier transform will decay only according to the $C^1(\mathbb{R})$ continuity in the discretization instead of the superior decay found for most $C^{\infty}(\mathbb{R})$ functions.

To efficiently transform between the spatial and the spectral domain with continuous functions, we use a Gabor frame as a discretization, which will be explained in Chapter 4. A Gabor frame is efficient since it exhibits an exponential convergence to continuous functions and can use analytical frame functions in both domains. However, this is not the only option, as we will show also that a higher-order (higher than 1) Hermite-spline-based discretization can be used to efficiently discretize functions. A high-order Hermite-spline-based approach can exhibit a high-order polynomial convergence in the approximation of a function and its derivatives. A third option would be to use Slepian functions [93] in a manner similar to [94]. On that case an orthogonal basis is employed that is localized in both a spatial square and a spectral-domain disk simultaneously. However, we did not pursue this option further.

To efficiently discretize the Green tensor in the spectral domain is challenging, since the Green tensor exhibits branch cuts along which there can be high-frequency oscillations and in a multilayer medium it can also exhibit poles. The high-frequency oscillations can in principle be countered by using a finer discretization wherever the number of oscillations is larger, as is presented in Chapter 5. However, when we go to a multi-layered medium this approach of finer sampling cannot be applied, because of the poles that can be encountered in the reflection coefficients Eq. (2.42). Since the poles can be present directly on the real $\mathbf{k}_T$-plane, convergence of the Fourier integrals can only be guaranteed by somehow mitigating these poles.

In principle, it is possible to analytically remove the poles from the reflection coefficients, perform the Fourier transform and then add the Fourier transform of the poles separately. However, this requires the pole locations to be known, and these are not readily available. Both location and strength of the poles have to be found numerically, which we prefer to avoid.

We have chosen to represent the electric field and the contrast current density on a complex path through the spectral domain in Chapter 6. Such a path can be chosen such that the oscillations along the branch cuts and the poles are kept at a considerable distance from the path on which we represent the contrast current density and the electric field. A major part of this work is to show that an efficient and fast transformation to and from such a spectral path is available.

Although we make a choice for a particular spectral path, other options are also available. The main advantage of the path of our choice is a fast transformation to and from the spatial domain, since most information is contained in regions with a $\mathbf{k}_T$-independent shift into the complex plane, for which an ordinary Fourier transformation can be used. However, this method has the disadvantage that the spectral path is broken up into three parts in two dimensions and nine parts in three dimensions, each with their individual discretization, and which all have to be connected. When a fast, more flexible method is available to transform directly to a wider range of possible complex-plane paths, this might lead to a more efficient procedure.

3.4 Iterative solvers

In general, a discretized equation can be written as a set of coefficients $\mathbf{F}_m$ that approximate the field $\mathbf{F}(x, y, z)$ through

$$\mathbf{F}(x, y, z) \approx \sum_{m=1}^{M} \mathbf{F}_m b_m(x, y, z). \tag{3.20}$$

Here the $b_m(x, y, z)$ are the functions used to span the function space in which the components of $\mathbf{F}(x, y, z)$ reside, e.g. Gabor frames in the x, y directions and PWL functions in the z direction. We can define the scalar product between two sets of coefficients through

$$\langle \{\mathbf{F}_m\}, \{\mathbf{H}_n\} \rangle = \sum_{m=1}^{M} \sum_{n=1}^{M} \mathbf{F}_m \cdot \overline{\mathbf{H}}_n \int_{\mathcal{D}} dx\, dy\, dz\; b_m(x, y, z) \overline{b_n(x, y, z)}, \tag{3.21}$$

where $\mathcal{D}$ is the computational domain spanned by the support of the basis functions. The discretized integral equation Eq. (3.7) can be written in discretized form as an $M \times M$ matrix $\mathcal{A}$, and the resulting linear system can be written as

$$\mathbf{E}^i_m = \sum_{m=1}^{M} \mathcal{A}_{mn} \cdot \mathbf{F}_n, \tag{3.22}$$

where $\mathbf{E}^i_m$ is a set of coefficients describing the incident field. This is a matrix equation, that can be solved by many techniques. The most obvious choice would be to directly invert $\mathcal{A}_{mn}$. However, since this matrix is typically very large, M will often be much larger than 10^6. This is not practical since the computation time for direct matrix inversion scales as $O(M^3)$ in computation time. Iterative techniques are more suitable, since they only need matrix-vector (matvec) multiplications, which can be calculated faster. Naively implemented, the calculation time of such a multiplication scales as $O(M^2)$, but by using efficient techniques this can be scaled down for some problems. We achieve this in $O(M \log M)$ operations through the mixed spatial-spectral formulation of the problem, when the basis functions allow for fast operations depending on FFTs.

Many different iterative techniques for calculating the solution to Eq. (3.22) are available. Note that the calculation of a scalar product (Eq. (3.21)) is important to be able to measure how far a proposed solution has converged. Important aspects for the choice of a technique are the required number of iterations for convergence, the need for the adjoint of $\mathcal{A}$, and the memory requirements. Some iterative techniques do not even converge for more complicated scattering problems, especially when the contrast is high. It is out of the scope of this thesis to give a broad overview of all the different techniques in existence, but we will mention the ones important for our application. Since we found programming the adjoint of $\mathcal{A}$ very time consuming, we will focus on methods that do not employ the adjoint.

One of the first techniques to be developed is the generalized minimal residual (GMRES) [95, 96]. The idea behind this method is to find a small linear subspace of the complete space spanned by the $\{b_m(x, y, z)\}$, in which the solution resides. With each iteration a projection vector is added to the linear subspace until we expect that the solution can be represented in the subspace up to a certain accuracy level. Each new vector is chosen through the Arnoldi iteration, i.e. by letting $\mathcal{A}$ work on the previous vector and projecting out all previous vectors. This allows for a projection of $\mathcal{A}$ to a small linear subspace, this projection is then inverted and yields the approximate solution. Since the matrix in the projected subspace is much smaller, its inversion is calculated much faster. There are several advantages to this technique: it does not need the adjoint of $\mathcal{A}$ and it is guaranteed to yield the correct solution in M iterations in infinite precision. The downside is that this representation needs a copy of all the projection vectors onto the linear subspace, which are $N_i M$ numbers, with N_i the number of iterations. When a large number of iterations is needed, the memory requirement will become too large.

A variation of the GMRES method is the restarted GMRES [96]. Here GMRES is used with a smaller number of iterations, so it has only converged partly to the solution. The partly converged solution is then used as a starting point from which to start GMRES once more. Each time GMRES is started, a better solution is obtained. This method has the advantage of using less memory, since it does not save the projection vectors for all iterations. The downside is that sometimes more iterations are needed and that convergence in M or less steps is not guaranteed anymore. However, in finite precision the orthogonality of the subspace in which $\mathcal{A}$ is projected gets lost after a large number of iterations and then a restart can even be advantageous.

Another popular method is based on Conjugate Gradient (CG): the stabilized bi-Conjugate Gradient method BiCGstab(ℓ) [39, 96, 97, 98]. Since CG requires the adjoint of $\mathcal{A}$, we did not consider it because of the extra effort of assembling the adjoint. At each iteration in BiCGstab(ℓ) a set of 2ℓ vectors are calculated on which the error of the approximated solution is minimized. In the following iteration a new set of vectors is calculated that is conjugate to the old ones with respect to $\mathcal{A}$. Advantages of this method are that it only needs ℓM memory space, and that it is easy to monitor the convergence while running, as opposed to GMRES. However, convergence is again not guaranteed, although for low contrast this is generally not a problem.

Chapter 4

The Gabor frame

For the discretization in the transverse direction a method is needed that is efficient both in the spatial and spectral domains. Additionally a fast way of transforming between these domains should exist, since these transformations are part of the core of the algoritm. In this chapter we summarize some important results about the Gabor frame and explain how it is implemented in our algorithm, showing the applicability in this case. We start with a definition of the Gabor basis, followed by the Gabor frame which often needs much fewer coefficients than the Gabor basis to represent the same function accurately and two methods to calculate the dual window function are presented. Then we illustrate a fast algorithm to calculate the Gabor coefficients of a function and show how the Fourier transform of a function represented by Gabor coefficients is calculated. We end with some numerical examples. This chapter is not meant as a thorough treatment of the Gabor basis and Gabor frames. An example of such a treatment can be found in [99, 100].

4.1 The Gabor basis

The Gabor basis was first introduced by Dennis Gabor in [101]. For its definition we follow [102] to define the basis functions as

$$g^b_{mn}(x) = g(x - mX)e^{jKnx}, \tag{4.1}$$

where there is a freedom of choice for the window function $g(x)$. We use the popular Gaussian window function given by

$$g(x) = 2^{\frac{1}{4}} e^{\left(-\pi \frac{x^2}{X^2}\right)}. \tag{4.2}$$

The $2^{1/4}$ normalization is customary in the literature. We have used the superscript ^{b} to distinguish the Gabor basis from the Gabor frame that will be defined later on. The spacing of the window functions for the Gabor basis equals $X = \frac{2\pi}{K}$. It is this specific choice of X and K that yields a basis for $\mathcal{L}^2(\mathbb{R})$ [103]. This is analogous to the discrete Fourier transformation, where the product of the spectral and spatial sampling rates also equals $2\pi/N$. Since this set of functions $g_{mn}(x)$ forms a basis for $\mathcal{L}^2(\mathbb{R})$ we can write

$$f(x) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f_{mn} g^b_{mn}(x), \tag{4.3}$$

for every $f(x) \in \mathcal{L}^2(\mathbb{R})$ with Gabor coefficients f_{mn}. These Gabor coefficients can be calculated via

$$f_{mn} = \int_{-\infty}^{\infty} dx\, f(x) \eta^b_{mn}(x), \tag{4.4}$$

where the dual basis $\eta^b_{mn}(x)$, which is also a Gabor basis, is defined in terms of the dual window function $\eta(x)$ as

$$\eta^b_{mn}(x) = \eta(x - mX) e^{jKnx}. \tag{4.5}$$

For the particular choice of Eq. (4.2), the dual window $\eta^b(x)$ is given by [102, 99]

$$\eta^b(x) = \frac{1}{X 2^{1/4}} \left(\frac{K_0}{\pi}\right)^{\frac{-3}{2}} \sum_{n+\frac{1}{2} \geq |x/X|}^{\infty} (-1)^n \exp\left(\pi \frac{x^2}{X^2} - \pi(n + \frac{1}{2})^2\right), \tag{4.6}$$

with $K_0 = 1.85407468\cdots$, the complete elliptic integral for modulus $\sqrt{1/2)}$, which is plotted in Figure 4.1. Note that the Fourier transform of Eq. (4.1) forms a similar Gabor basis in the spectral domain, hence the same constructions can be used and it can be shown that the dual window function in the spectral domain is just a scaled version of Eq. (4.6). The main advantage of the Gabor basis is that functions can be written in a mixed spatial and spectral representation. Every basis function is localized in the spatial domain around mX and in the spectral domain around nK. This behaviour differs considerably from e.g. the triangle function of a PWL discretization, which is localized in the spatial domain, but very spread out in the spectral domain.

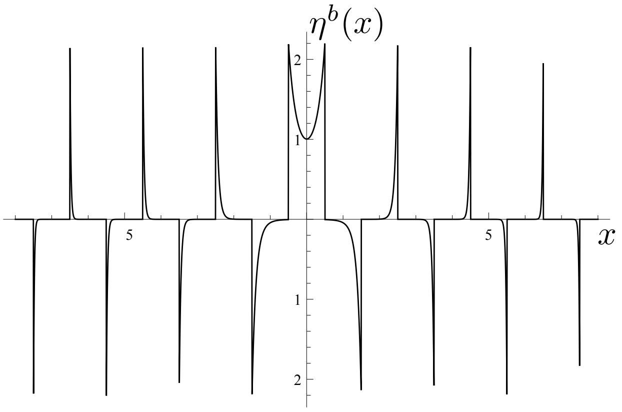


Figure 4.1: The dual window $\eta^b(x)$ of Eq. (4.6) for $X = 1$.

4.2 Gabor frames with oversampling

4.2.1 The Gabor frame

As can be seen from Figure 4.1, the dual window function does not decay rapidly for large x in the spatial domain. All peaks have the same maximum amplitude, but at large x, they become narrower, at a rate $1/x$. The same dual window function is applicable in the spectral domain, since the spectral dual window is a stretched version of the dual window Eq. (4.6) in Figure 4.1. Because of this, many significantly contributing Gabor coefficients are produced by the integral in Eq. (4.4), even when $f(x)$ is itself localized and rapidly decaying both spatially and spectrally. Hence, many Gabor coefficients contribute significantly, and the convergence of the sums in Eq. (4.3) is poor, which results in a need for a relatively large number of basis functions compared to other discretization schemes. The Balian Low theorem [104, 105, 106, 107] proves that a Gabor basis with exponential decay does not exist.

However, when oversampling is used, i.e. when the spatial or the spectral sampling of the function is denser than $X \cdot K = 2\pi$, the Balian Low theorem does not hold and a more convenient dual window function can be chosen. Although this oversampled set of functions does not form a basis, it forms a frame, i.e. a set of functions that does span the whole space, but has redundancy.

The frame is defined through

$$g_{mn}(x) = g(x - \alpha m X)e^{j\beta K n x}, \tag{4.7}$$

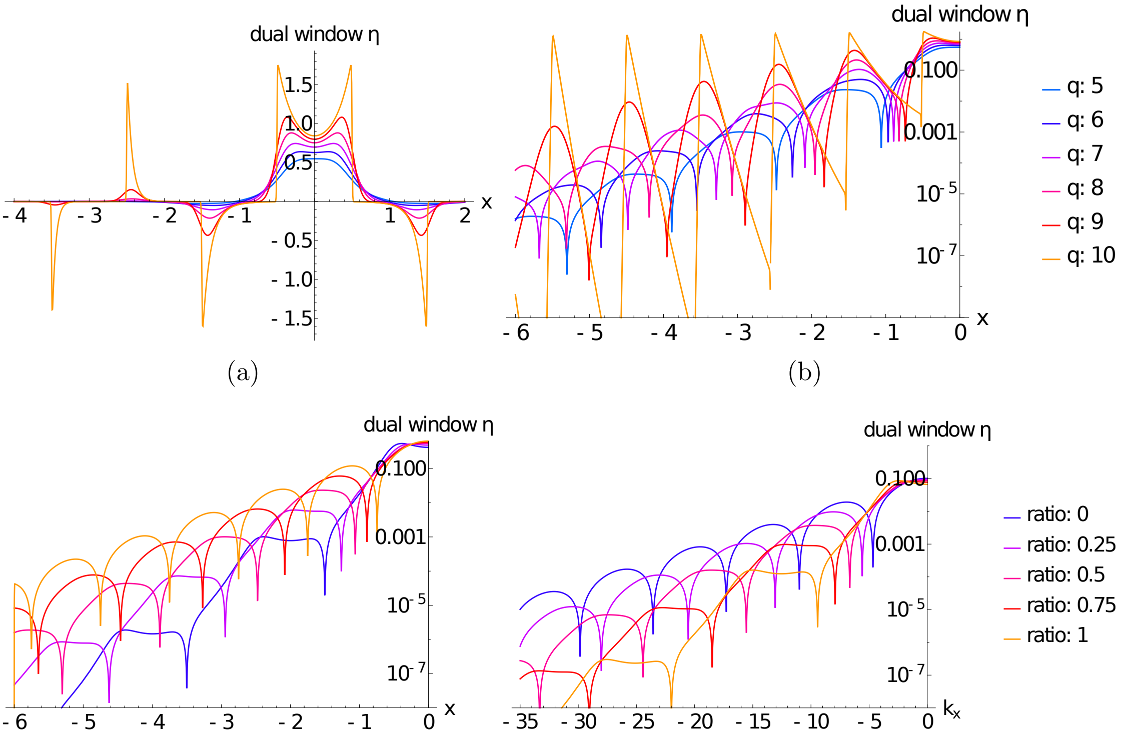
with α defining the spatial oversampling and β defining the spectral oversampling such that $\alpha\beta < 1$. In practice it is convenient to choose rational oversampling, i.e. $\alpha\beta = q/p$ with $q, p \in \mathbb{N}$ [100]. When $\alpha\beta = 1$ the frame is critically sampled and Eq. (4.7) then coincides with the equation for the Gabor basis Eq. (4.1) apart from a shift in $X = 2\pi/K$. When $\alpha\beta > 1$, the frame is undersampled and a stable representation is not possible. For $\alpha\beta < 1$, the frame is oversampled, which means that the representation in Gabor coefficients is not uniquely determined. However, this non-uniqueness has the advantage of a freedom to choose the Gabor coefficients in better converging ways. To fix this choice, we calculate Gabor coefficients in a manner similar to Eq. (4.4), with a dual window. The dual frame function $\eta(x)$ for an oversampled frame can be chosen more localized and smooth than $\eta^b(x)$ in Eq. (4.6) for a better decay in both the spectral and spatial domain. Although Gabor coefficients are not uniquely determined for a function in an oversampled frame, the choice of the dual frame fixes this non-uniqueness. Note that a Wilson basis [108] can be constructed from an oversampled Gabor frame [109]. A Wilson basis, which is closely related to the Gabor frame, does not have redundancy. However, the localization of the basis functions is slightly poorer than the Gaussian basis. Therefore the Gabor frame is our preferred choice.

Analogous to the dual Gabor basis in Eq. (4.5), the Gabor frame has a dual frame given by

$$\eta_{mn}(x) = \eta(x - \alpha m X)e^{j\beta K n x}, \tag{4.8}$$

but since the Gabor coefficients are not uniquely determined, there is freedom of choice for the dual window function. The specific choice of this window function is critical for a good decay of the Gabor coefficients at large m and n-indices. For calculations we prefer a fast exponential decay of the dual window function, both in the spatial and in the spectral domain, so we can apply a reasonable cut off in the Gabor coefficients.

In Figure 4.2(a & b), several dual window functions are shown for an increasing oversampling q, equally divided over both the spatial and spectral domains, $\alpha = \beta = \sqrt{q/p}$. Clearly, the dual window with the highest oversampling ($q = 5$) decays faster. In Figure 4.2(c & d), the effect is shown of different ratios of oversampling in $\alpha = (q/p)^r$ and $\beta = (q/p)^{1-r}$. Clearly, when the spatial oversampling increases, the dual window converges faster in the spatial domain. However, its Fourier transform converges slower. This indicates that, there is a trade-off between them. For that reason we choose $\alpha = \beta$ throughout the rest of this thesis, since it is a reasonable choice that works well in both domains.



(a) (b)

(c) (d)

Figure 4.2: The dual window $\eta(x)$ for different oversampling ratios q/p. All plots with $p = 10$, $X = 1$. Plot (a) and (b) with increasing q and $\alpha = \beta = \sqrt{q/p}$. Plot (c): the dual window with $q = 5$ and the oversampling divided between α and β through ratio r, such that $\alpha = (q/p)^r$, $\beta = (q/p)^{(1-r)}$, and (d) the Fourier transform of the dual window in (c).

4.2.2 Calculation of the dual window through the Zak transform

To be able to compute the dual window function we now give a short summary of the work by Bastiaans in [100, 110], which is the most well-known method to calculate the dual window. We will only present the steps used in the calculation, more details can be found in [100]. The Zak transformation of a function $w(x)$ is defined as

$$\tilde{w}(x, \omega; X) = \sum_{m=-\infty}^{m=\infty} w(x + mX) e^{jm\omega X}, \tag{4.9}$$

where $\tilde{w}(x, \omega; X)$ denotes the Zak transform. Again X signifies the window width, which is considered constant. The inverse Zak transformation is

$$w(x + mX) = \frac{X}{2\pi} \int_0^{2\pi/X} d\omega \, \tilde{w}(x, \omega; X) e^{-i\omega X}. \tag{4.10}$$

The Zak transform is periodic in the frequency ω with period $\frac{2\pi}{X}$ and quasi periodic in x with period X. Hence, all the information about the function is contained in the $x - \omega$ rectangle $[0, X] \times [0, 2\pi/X]$ with area 2π. In case of rational oversampling by a factor $q/p \leq 1$ we can calculate the dual window by using the $(p \times q)$ matrix-function W with entries $w_{sr}(x, y)$, given by

$$w_{sr}(x, y) = \tilde{\eta} \left((x + s)\frac{\alpha p X}{q}, (y + \frac{r}{p})\frac{K}{\alpha}; \alpha X \right). \tag{4.11}$$

This is the Zak transform of the dual window $\eta(x)$. In a similar way we create the $(q \times p)$ matrix G from the Zak transform of the window function g. If one requires that Eqs. (4.3) and (4.4) hold, one can show that this boils down to the requirement

$$\frac{\alpha X}{q} G \cdot W^* = \mathbb{1}_q, \tag{4.12}$$

with the asterisk denoting the conjugate transpose. Consequently, when we decide to employ a certain window function g, we can find matrices W that satisfy this relation. For critical sampling, the only solution is $\eta^b(x)$ from (4.6). Note that this matrix equation is underdetermined in the case of oversampling. Hence, there is some freedom in the choice of the solution, but a popular choice [100, 99] is the generalized Moore-Penrose pseudo inverse to calculate W. It can be shown that this pseudo inverse is the optimum solution in the sense of minimizing the $\mathcal{L}^2$ norm of η [99, Chapter 8][111, 112]. Other criteria to calculate the pseudo inverse of this matrix yield dual window functions that can be optimal in other respects, e.g. [99, Chapter 8] [113].

4.2.3 Direct calculation of the dual window

The computation of the dual window by means of the Zak transformation, as explained in the preceeding section, requires the evaluation of integrals as in Eq. (4.10). Since we prefer to avoid computing numerical integrals we continue by explaining how the dual window can be computed in a more direct manner. We start by giving an approximate method for critical sampling, i.e. we choose $\alpha = \beta = 1$. Subsequently we will give an approximation for oversampling in Section 4.2.4 and an exact method for critical sampling in Section 4.2.5, followed by a short discussion about this method and the resemblence with the Zak-transform method of the preceeding section.

We observe that the inverse Gabor transform of a function $f(x)$, Eq. (4.3) can be seen as a set of shifted window functions $g(x - mX)$ with a periodic modulation $\mathring{f}_m(x)$ with period X

$$f(x) = \sum_{m,n\in\mathbb{Z}} g(x - mX) f_{mn} e^{jKnx} = \sum_{m\in\mathbb{Z}} g(x - mX)\mathring{f}_m(x). \tag{4.13}$$

This defines the set of periodic functions $\mathring{f}_m(x)$ as

$$\mathring{f}_m(x) = \sum_{n\in\mathbb{Z}} f_{mn} e^{jKnx}. \tag{4.14}$$

We observe that $\mathring{f}_m(x)$ is defined as a Fourier series. For good convergence in the summation over n we need smooth functions $\mathring{f}_m(x)$. When the $\mathring{f}_m(x)$ are found, the Gabor coefficients f_{mn} can be calculated and this implicitly defines the dual window function. Because these functions are periodic, we can divide the non-periodic functions $f(x)$ and $g(x)$ in Eq.(4.13) into intervals $[\frac{-X}{2}, \frac{X}{2}]$. First we divide $f(x)$ to obtain a sum of functions f_j defined on different unique subdomains of $\mathbb{R}$.

$$f_j : \left[\frac{-X}{2}, \frac{X}{2}\right] \to \mathbb{R} \big| f_j(x) = f(x + jX), \tag{4.15}$$

We can write this as a vector $\underline{f}$ with elements f_j. Now each element $f_j(x)$ covers one part of the domain of $f(x)$ with width X.

For the set of (shifted) window functions in (4.13) we can use the matrix

$$G_{mn}(x) = g(x - (n - m)X), \tag{4.16}$$

which allows writing (4.13) as

$$\underline{f}(x) = \underline{\underline{G}}(x) \cdot \underline{\mathring{f}}(x). \tag{4.17}$$

Since the vector space is infinite-dimensional, this may be difficult to solve formally. In practice, we can truncate this to a finite size, so this becomes a matrix equation with a finite matrix and we can find our approximate $\mathring{f}_m(x)$ as

$$\underline{\mathring{f}}(x) = \underline{\underline{G}}^{-1}(x) \cdot \underline{f}(x). \tag{4.18}$$

We can use this $\underline{\underline{G}}^{-1}(x)$, to construct $\eta(x)$ since this function is able to generate $\underline{\mathring{f}}(x)$, which is only a Fourier transformation away from the Gabor coefficients. To see this, we work out f_{mn}, i.e.

$$\begin{aligned}
f_{mn} &= \int_{-\frac{X}{2}}^{\frac{X}{2}} dx\ e^{\frac{2\pi j}{X}xn} \mathring{f}_m(x) \\
&= \int_{-\frac{X}{2}}^{\frac{X}{2}} dx\ e^{\frac{2\pi j}{X}xn} \sum_{i\in\mathbb{Z}} G^{-1}_{mi}(x) f_i(x) \\
&= \sum_{i\in\mathbb{Z}} \int_{-\frac{X}{2}}^{\frac{X}{2}} dx\ e^{\frac{2\pi j}{X}xn} G^{-1}_{mi}(x) f(x-iX) \\
&= \sum_{i\in\mathbb{Z}} \int_{iX-\frac{X}{2}}^{iX+\frac{X}{2}} dx\ e^{\frac{2\pi j}{X}xn} G^{-1}_{mi}(x+iX) f(x) \\
&= \int_{-\infty}^{\infty} dx\ e^{\frac{2\pi j}{X}xn} \eta(x-mX) f(x),
\end{aligned}$$

where on the last line we used the definition of the Gabor transformation (4.4). Now we can write the dual window function $\eta(x)$ as

$$\eta(x-mX) = \sum_{i\in\mathbb{Z}} G^{-1}_{mi}(x+iX) \mathbb{1}_{[(i-\frac{1}{2})X,(i+\frac{1}{2})X]}(x), \tag{4.19}$$

with $\mathbb{1}_{[a,b]}$ equal to 1 on the interval $[a,b]$ and 0 everywhere else. If we approximate this by taking a finite matrix for G, we arrive at a good approximation for the dual window.

4.2.4 Approximation of η for oversampling

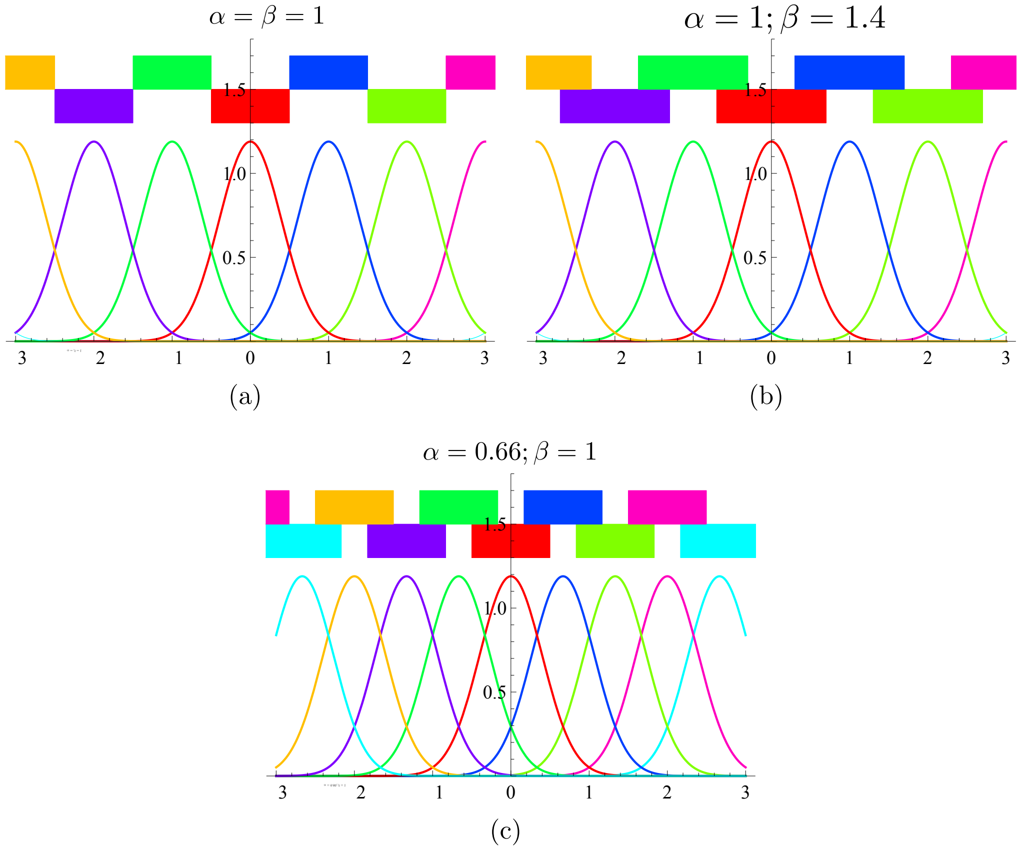
Now that we have found a method to compute the dual window function without oversampling in the preceding section, we explain how this can be generalized to an oversampled Gabor frame. The main principle behind the calculation is the same as in the previous section. The main property that changes when there is oversampling is that the $\mathring{f}_m(x)$ functions 'overlap'. The spacing between the window functions is smaller than the period of the periodic functions, as can be seen in Figure 4.3. Here we plotted window functions with a spacing αX and above each window we drew a rectangle with a width corresponding to the period of a periodic function in the same color. For critical sampling there is no overlap, so the periodic function has to contain all information of $f(x)$ over its complete period. With critical sampling it can therefore happen that the periodic function needs to be discontinuous when $f(x)$ is continuous, just because $\mathring{f}_m(-X/2) \neq \mathring{f}_m(X/2)$. A discontinuous periodic function will not have a rapidly converging Fourier series in Eq. (4.14). This is an intuitive explanation of the Balian Low theorem. It should also be clear that with oversampling a smooth transition can be chosen between the different periodic functions, and, therefore, the coefficients can decay much faster, which yields a better behaved dual window function.

We will first look at the case where $\alpha < 1$ and $\beta = 1$, where Eq.(4.16) changes to

$$G_{mn}(x) = g(x-(m+\alpha n)X), \tag{4.20}$$

Square matrices will yield an ill conditioned matrix that is hard to invert accurately. Therefore it is better, although not absolutely necessary, to avoid square matrices. If we want to make an approximation of the matrix $G_{mn}(x)$ in a certain range of x, we need more m values than n values, since index m results in a X-sized step in $g(x)$ and n results in an αX sized step, as can be observed from Eq. (4.20). Therefore more n values are needed by at least a factor $\frac{1}{\alpha}$. Now for the matrix inversion we again use the Moore-Penrose pseudo-inverse.

This method is more flexible in the sense that we can now choose $\alpha \in \mathbb{R}$. It looks as if this method is only capable of taking into account spatial oversampling α and no spectral oversampling β. However, it is possible to approximate the same dual windows that are found in Section 4.2.2. From Figure 4.3 one can conclude that a dual window obtained by using the Zak transformation method with α and β can be found both by a transformation $g(x) \to g(x/\beta)$, $\eta(x) \to \eta(\beta x)$, and $\alpha \to \alpha/\beta$ using the approximate method.



(a)

(b)

(c)

Figure 4.3: A few shifted window functions, with rectangles above the graphs signifying the length of one period of $\overset{\circ}{f}_m(x)$. In (a) for critical sampling there is no overlap. For (b) there is overlap because the period, $1/\beta K$ is larger, for (c) there is overlap because the window functions have been shifted closer together at αX.

This method is also more flexible in the sense that we can use it for exact results for a finite number of window functions, or window functions that are not all equal, although we do not investigate this in this thesis.

4.2.5 Exact η for critical sampling - connection with the Zak transform

It is instructive to see what happens when the $G_{mn}(x)$ matrix is not truncated in the direct method of Section 4.2.3. Without oversampling this will be solved in a formal manner. A result equivalent to Eq. (4.12) is found in the end and we can identify a Zak transformation in the process.

Without truncation the basic steps are the same, but the problem is the inversion of the operator $G(x)$ in Eq. (4.18), which we will no longer consider as a finite matrix. We have to come up with a way to formulate the inverse of this operator. If we define (dropping the x dependence) $G_{mn}(x) = w_{m+n}$, which is allowed since G is Toeplitz, we can define the inverse $G^{-1}_{i+\ell}(x)$ as $v_{i+\ell}$

$$\delta_{mk} = \sum_{\ell\in\mathbb{Z}} G_{m\ell}(x) G^{-1}_{\ell k}(x) = \sum_{\ell\in\mathbb{Z}} w_{m-\ell} v_{\ell-k}. \tag{4.21}$$

Now we can carry out the inverse Fourier transformation of $w_k = \int_{-\pi}^{\pi} d\eta \; \tilde{w}(\eta) \; e^{-j\eta k}$ and its inverse $\tilde{w}(\eta) = (1/2\pi) \sum_k w_k e^{j\eta k}$ to find

$$\begin{aligned} \delta_{mk} =& \sum_{\ell\in\mathbb{Z}} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{d\eta d\eta'}{4\pi^2} \, \tilde{w}(\eta) \, v(\eta') \; e^{-j(m-\ell)\eta - j(\ell-k)\eta'} \\ =& \int_{-\pi}^{\pi} \frac{d\eta}{2\pi} \, \tilde{w}(\eta) \, \tilde{v}(\eta) \; e^{-j(m-k)\eta}. \end{aligned} \tag{4.22}$$

This can only be true when $\tilde{w}(\eta)\tilde{v}(\eta) = 1$, or

$$\tilde{v}(\eta) = \frac{1}{\tilde{w}(\eta)}. \tag{4.23}$$

From this we can calculate $G^{-1}(x)$. In the integrals in Eq. (4.22) we recognize the Zak transform of Eq. (4.10). Although $\tilde{v}$ and $\tilde{w}$ are written as function of a single coordinate we emphasize that the x-dependence was left out in Eq. (4.21). The resemblance between Eq. (4.23) and Eq. (4.12) for $p = q = \alpha = 1$ (no oversampling) is now apparent. Since we would usually choose a window function with exponential spatial decay, its Fourier transformation will decay, but there is no guarantee that $\tilde{w}(\eta)$ has no zeros. Because this method is similar to the Zak-transform based approach, the existence of $1/\tilde{w}$ boils down to the same conditions as were discussed in [100] for the existence of the function η.

We conclude that our approach is close to the articles of Bastiaans [114, 102, 100, 110], which uses Zak-transforms. For completeness it should be added that an approach based on functional analysis was developed by Wexler, Raz and Daubechies [115, 116, 117]. An important difference between the methods is that the method developed by Bastiaans tries to find an expression to directly evaluate the dual window functions for each x argument, whereas the methods developed by Daubechies lead to a Gabor expansion of the dual window. The difference is that a pseudo-inverse of a matrix has to be computed for each x-argument with the approach that we employ, whereas the functional-analysis method requires only one pseudo-inverse to find the Gabor coefficients of the dual window. Afterwards, the dual window can be evaluated as a function of x by evaluating Eq. (4.3). In principle, the functional-analysis approach evaluates the dual window faster, but the evaluation of Eq. (4.3) is still relatively slow. Therefore, evaluation through an interpolation is worthwile for both methods and a difference in performance is only observed in the initialization time. In practice the computation of the interpolation coefficients with the method we described throughout this chapter can be evaluated in about a second (on an Intel i7-4600U CPU), which is satisfactory. Therefore, we did not further pursue the functional-analysis based approach.

4.3 Computational aspects

4.3.1 Fast transform to and from Gabor coefficients

When we are given a Gabor frame with window function $g(x)$ and dual window $\eta(x)$ we can calculate the Gabor coefficients of a function f using the dual frame functions $\eta_{mn}(x)$ by means of Eq. (4.4). Unfortunately calculating all these integrals is a tedious procedure.

Suppose we would like to calculate Gabor coefficients for $m \in \{-N_x, \cdots, N_x\}$ and $n \in \{-N_k, \cdots, N_k\}$. We write down a slightly different version of the periodic functions of Eq. (4.14) with the spectral oversampling included, i.e.

$$\mathring{f}_m(x) = \sum_{\ell=-\infty}^{\infty} f_{m\ell} e^{j\beta\ell K x}. \tag{4.24}$$

We can insert this in Eq.(4.4) to find

$$\begin{aligned}\mathring{f}_m(x) &= \int_{-\infty}^{\infty} dx' \sum_{\ell=-\infty}^{\infty} f(x')\eta(x'-m\alpha X)e^{j\beta\ell K(x-x')} \\ &= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} dx'\, f(x')\eta(x'-m\alpha X)\delta(x-x'-\frac{2\pi n}{\beta K}) \\ &= \sum_{n=-\infty}^{\infty} f(x+n\frac{2\pi}{\beta K})\eta(x-m\alpha X+n\frac{2\pi}{\beta K}),\end{aligned}$$

where we have written the summation over ℓ with complex exponentials as a sum of delta functions, by using the identiy $\frac{2\pi}{\beta K} = \frac{X}{\beta}$. Since we choose $\alpha\beta \in \mathbb{Q}$ or $\alpha\beta = \frac{p}{q} < 1$, Eq.

(4.24) can be rewritten as

$$\mathring{f}_m(x) = \sum_{n=-N_x}^{N_x} f(x+\alpha X\frac{p}{q}n)\eta(x-\alpha Xm+n\alpha X\frac{p}{q}). \tag{4.25}$$

Since we want to use the inverse of Eq. (4.24) to calculate the coefficients f_{mn} from $\mathring{f}_m(x)$ using FFTs, we do not need to know the $\mathring{f}_m(x)$ for arbitrary x, but only for certain discrete $x_i = i\frac{2\pi}{\beta K N_k} = i\frac{\alpha X}{N_k}\frac{p}{q}$, where N_k signifies the highest spectral coefficient.

After substituting all this in Eq. (4.25)

$$\mathring{f}_m(x_i) = \sum_{n=-N_x}^{N_x} f\left(\alpha X(\frac{1}{N_k}\frac{p}{q}i+\frac{p}{q}n)\right)\eta\left(\alpha X(\frac{i}{N_k}\frac{p}{q}-m+n\frac{p}{q})\right),$$

we notice that the summation over n is similar to a discrete convolution evaluated at m except for a factor p/q. When we write $n=\ell+qk$ with $\ell \in \{1,\cdots,q\}$ the sum becomes

$$\mathring{f}_m(x_i) = \sum_{\ell=1}^{q}\sum_{k=-\infty}^{\infty} f\left(\alpha X(\frac{1}{N_k}\frac{p}{q}i+\frac{\ell p}{q}+pk)\right)\eta\left(\alpha X(\frac{p}{q}\frac{i}{N_k}-m+\frac{\ell p}{q}+pk)\right).$$

When we now take $m=r+ps$ with $r\in\{0,\cdots,p-1\}$ this reduces to a true discrete convolution in k and s

$$\mathring{f}_{r+ps}(x_i) = \sum_{\ell=1}^{q}\sum_{k=-\infty}^{\infty} f\left(\alpha X(\frac{1}{N_k}\frac{p}{q}i+\frac{\ell p}{q}+pk)\right)\eta\left(\alpha X(\frac{p}{q}\frac{i}{N_k}+\frac{\ell p}{q}-r+p(k-s))\right), \tag{4.26}$$

which can be calculated efficiently via FFTs. Since we are only interested in a finite number of Gabor coefficients, the summation over k can be restricted to $k\in\{-N_x/q,\cdots,N_x/q\}$, since it is of no use to sample the function outside the range where the Gabor coefficients accurately represent the function f. When we have found the values $\mathring{f}_m(x_i)$, we can easily calculate the Gabor coefficients f_{mn} by using an FFT to approximate the inverse of Eq.(4.24) as

$$f_{mn} = \int_0^{2\pi} dx\mathring{f}_m(x)e^{-i\beta nKx}. \tag{4.27}$$

For this method to yield accurate results, it is important that the function $f(x)$ does not contain components with a spatial frequency higher than included by the number of modulation frequencies, i.e. f_{mn} is negligible for $|n|>N_k$.

To return from Gabor coefficients to a representation in the spatial or spectral domain, we can reverse all steps from Eq. (4.27) back to Eq. (4.24) to generate the values of the function sampled at x_i. Most inversions are trivial. However, care has to be taken with the $\eta_{mn}(x)$ function, since $\eta(x)$ was calculated through a pseudo inverse, the matrix is badly conditioned. But since $\eta_{mn}(x)$ is the pseudo inverse of $g_{mn}(x)$, the frame $g_{mn}(x)$ can be used instead.

It is possible to calculate an interpolation on points between the sample points as well by using the Gabor frame equivalent to Eq. (4.3), i.e.

$$f(x) = \sum_{m,n} f_{mn} g_{mn}(x), \tag{4.28}$$

which requires performing a sum over all coefficients for each point x. Since Gabor coefficients are used, this is an interpolation with exponential decay in the spectral domain and can be useful when high precision is required at only a few evaluation points. However, it is often more efficient to use the FFT-based algorithm, since it provides the entire range of values at an equidistant lattice and points in between can be interpolated e.g. through a Hermite spline.

4.3.2 Gabor representation of the spectral domain

Having defined the Gabor coefficients and transformations for spatial functions, we would like to define a similar frame for the corresponding spectral representations i.e. Fourier transforms. It turns out that the Fourier transform of the frame functions itself forms a Gabor frame very similar to the spatial Gabor frame. Therefore it is our frame of choice in the spectral domain. In this section we will use hats to emphasize when a function or its Gabor coefficients are defined in the spectral domain. Since we will show in this section that Gabor coefficients in the spectral and the spatial domain are very closely related and that their transformation is fast, we will not distinguish between spatial and spectral coefficients after this section and the tilde notation will be dropped for coefficients as well as functions.

When we use the Gaussian window function of Eq. (4.2), we can use its Fourier transform to form a frame for spectral functions. We will call this function $g(k)$ and for a Gaussian window $g(x)$ it is a Gaussian window as well, i.e.

$$\hat{g}(k) = 2^{\frac{1}{4}} T \exp\left(\frac{-\pi}{K^2} k^2\right). \tag{4.29}$$

For the frame we can then write

$$g_{mn}(k) = g(k - m\beta K) e^{-ikn\alpha X}. \tag{4.30}$$

The similarity between this function and Eq. (4.7) is obvious. Now we can find a dual window $\eta(k)$ for the spectral domain as well with the constructions of one of the preceding sections.

Of course we need a way to proceed from the spectral Gabor coefficients to the spatial Gabor coefficients. So we calculate the Fourier transform of $g_{mn}(x)$ giving

$$\hat{g}_{mn}(k) = \hat{g}(k - n\beta K) e^{-jk\alpha mX} \; e^{2\pi j \beta \alpha mn} = g_{nm}(k) e^{2\pi j \beta \alpha mn}.$$

This allows us to go from spatial coefficients f_{mn} of a function $f(x)$ to spectral coefficients $\hat{f}_{mn}$ via the transformation

$$\hat{f}_{mn} = f_{nm} e^{-2\pi j \beta \alpha mn}, \tag{4.31}$$

that can be used in e.g.

$$\hat{f}(k_x) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \hat{f}_{mn} \hat{g}^b_{mn}(k_x). \tag{4.32}$$

The numerical implementation of this Fourier transformation is simple. The advantage is that we can now use functions spatially as well as spectrally in our algorithm, since coefficients are identical except for a factor and a transposition. For this reason we will drop the hat notation for spectral Gabor coefficients.

4.3.3 Numerical examples

To acquire a feeling for the capabilities of Gabor frame representations, we will discuss numerical examples in this section. We begin by showing the performance of Gabor frames with different oversampling in representing three simple functions: the constant function $c(x)$, the Heaviside step function $H(x)$, and a modulated and shifted Gaussian pulse $g_t(x)$ given by

$$g_t(x) = e^{(x-\frac{2}{3})^2+\frac{3}{2}jx}. \tag{4.33}$$

First we discretize these functions using the Gabor frames of Figure 4.2, with $X = 1$, $p = 10$, $q \in \{5, 6, \cdots, 10\}$, and $\alpha = \beta = \sqrt{q/p}$. The coefficients are restricted to $m, n \in \{-3, \cdots, 3\}$ in Eq. (4.7). On the first row in Figure 4.4 we show the original functions. On the second row we show how these functions are approximated by the Gabor basis ($q = 10$) and two Gabor frames ($q = 5$, $q = 8$). Clearly, with more oversampling, lower q, the range for x on which the approximation holds is smaller. Also it is clear that the Gabor basis ($q = 10$) performs much poorer than the frames. On the bottom row, the error in these approximations is shown. Clearly, the error somewhat depends on the function that is approximated, although all oversampled frames perform well in general. The largest oversampling does not automatically yield the best approximation. This can be partly explained by the fact that a larger oversampling yields a smaller spectral window spacing βK, and therefore decreases the interval of the spectral domain that is included in the representation.

From Figure 4.4 it is clear that using a frame with $q = 2$, $p = 3$ with $\alpha = \beta = \sqrt{2/3}$ contains enough oversampling to yield good results. We prefer to choose low p and q since some algorithms, such as the fast Gabor transform of Section 4.3.1, are less efficient for large p and q.

Figure 4.5 illustrates how Gabor expansions depend on the number of included Gabor coefficients. Here the number of spatial coefficients is defined by M_x through $m \in \{-M_x, \cdots, M_x\}$ in Eq. (4.7) and similarly the spectral range of coefficients through $n \in \{-N_x, \cdots, N_x\}$. In the first column the constant function $c(x) = 1$ is sampled with an increasing number of spatial coefficients M_x. In the spatial domain, the approximation is accurate over a range with increasing number of samples. When these sets of Gabor coefficients are transformed to the spectral domain, the coefficients should approximate a

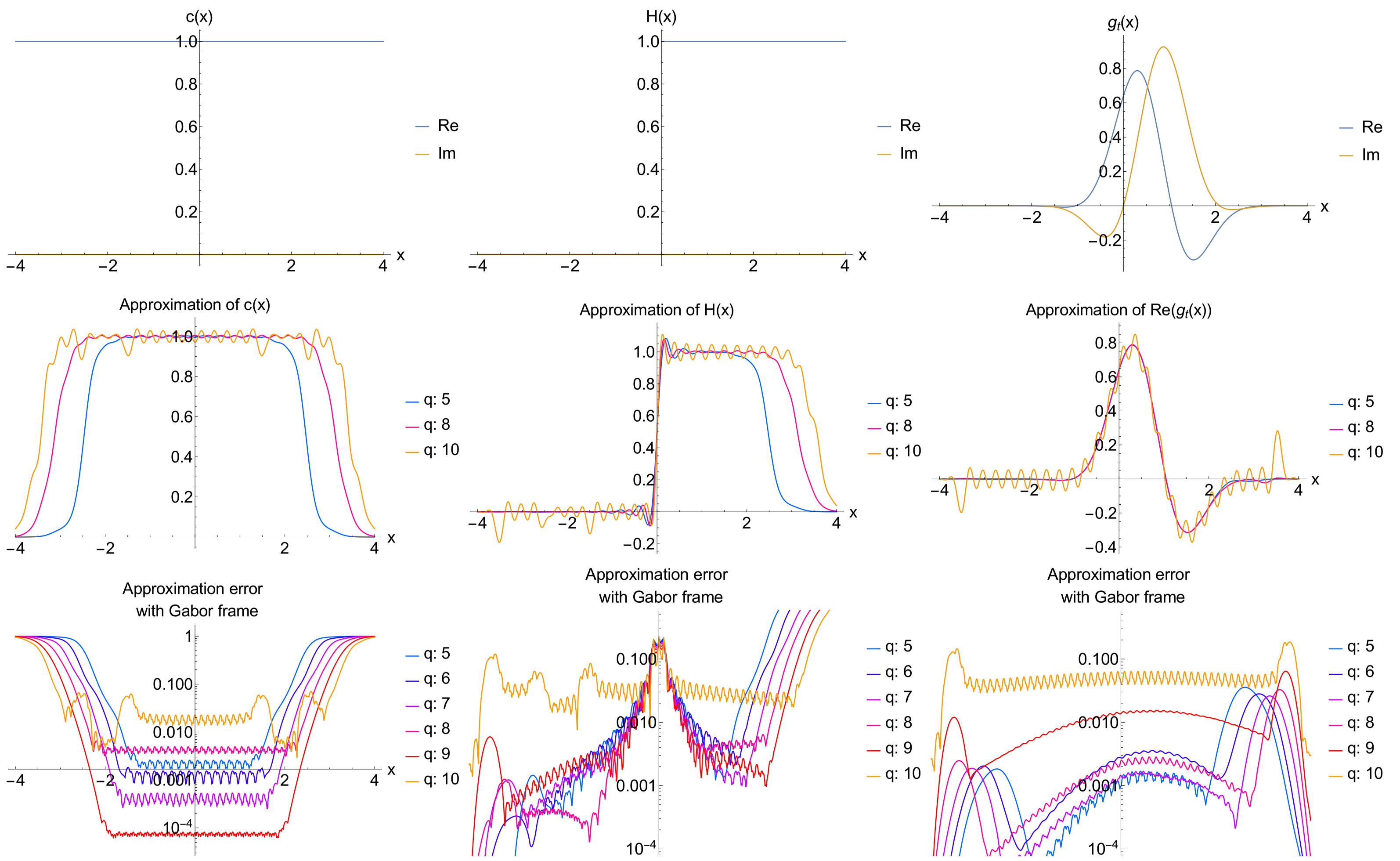
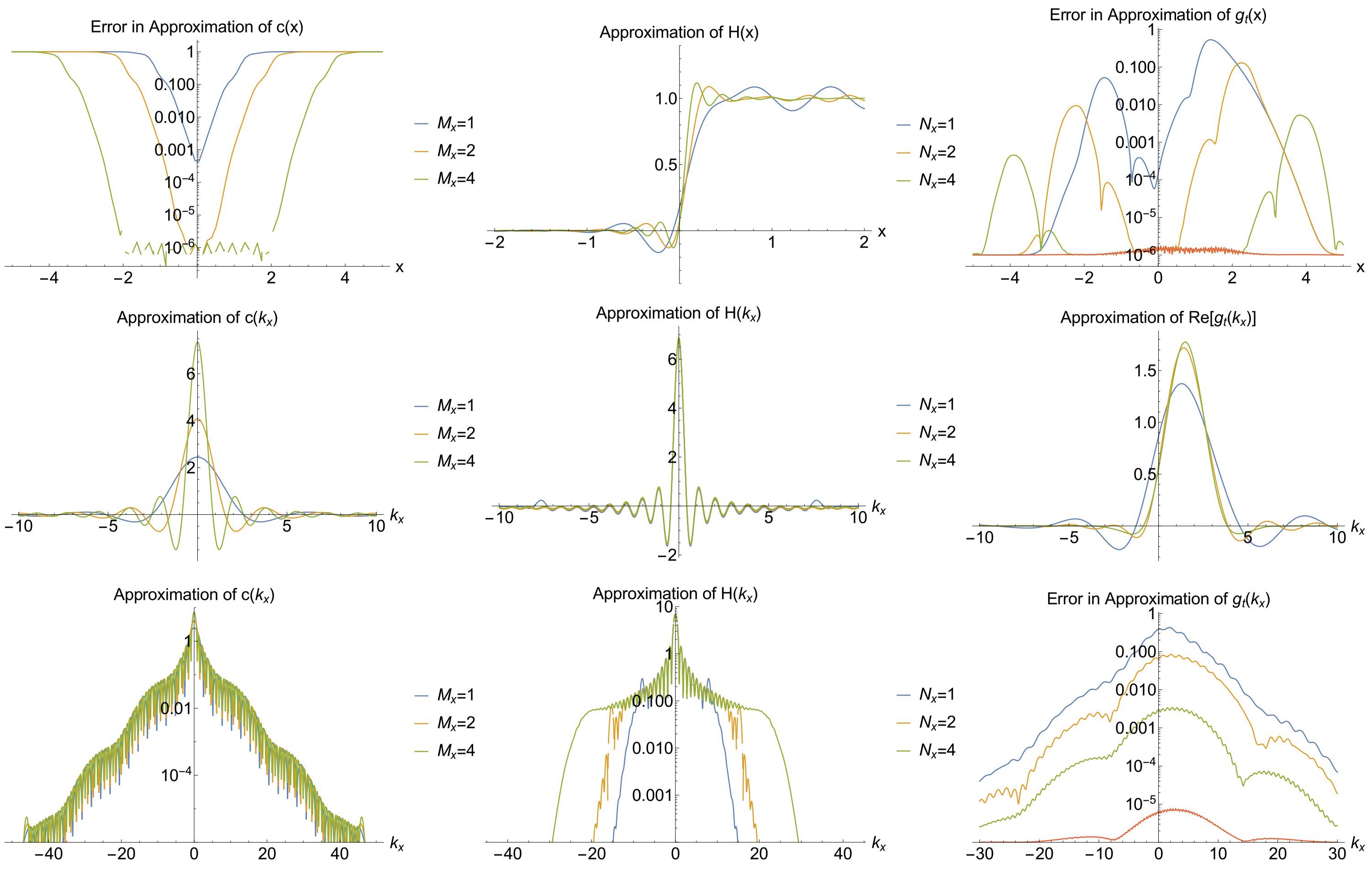


Figure 4.4: The effect of oversampling on the representation of a constant function $c(x)$, a Heaviside step function $H(x)$ and a modulated Gaussian pulse $g_t(x)$. The first row contains plots of the functions themselves, the second row contains plots of the functions approximated by a truncated Gabor frame and the third row shows the error of these approximations.

Dirac-delta distribution. For an increasing number of spatial coefficients, the distribution becomes narrower around $k_x = 0$. However, the decay in the spectral domain remains the same for large k_x.

In the second column in Figure 4.5, the Heaviside step function is sampled with an increasing number of spectral samples N_x. With an increasing number of spectral samples, the transition from 0 to 1 becomes narrower. However, similar to the Gibbs phenomenon, the oscillations do not decrease in amplitude.

In the third column, the modulated Gaussian pulse $g_t(x)$ in Eq. (4.33) is sampled over an increasingly long range. Since this function has an effectively finite support, the function can be well represented by a finite number of coefficients. Clearly, with increasing M_x, the approximation becomes better and both the representation in Gabor coefficients and in Fourier-transformed coefficients converge well to the function.


Error in Approximation of c(x)

Approximation of H(x)

Error in Approximation of $g_t(x)$

Approximation of c(k_x)

Approximation of H(k_x)

Approximation of Re[$g_t(k_x)$]

Approximation of c(k_x)

Approximation of H(k_x)

Error in Approximation of $g_t(k_x)$

Figure 4.5: Illustration of the effect of the number of samples in both the spatial and spectral domain. Here M_x defines the range of the m-index, i.e. $m \in \{-M_x, \cdots, M_x\}$ and N_x defines the range of the n index, i.e. $n \in \{-N_x, \cdots, N_x\}$ in Eq. (4.28). The first of the three columns shows the constant function $c(x) = 1$, for different spatial samplings M_x. The second column shows the Heaviside step function for different spectral samplings N_x. And the third column a Gaussian Eq. (4.33) for different spatial samplings.

Chapter 5

The Gabor frame as a discretization for the 2D transverse-electric scattering-problem domain integral equation[1]

5.1 Abstract

We apply the Gabor frame as a projection method to numerically solve a 2D transverse-electric-polarized domain-integral equation for a homogeneous medium. Since the Gabor frame is spatially as well as spectrally very well convergent, it is convenient to use for solving a domain integral equation. The mixed spatial and spectral nature of the Gabor frame creates a natural and fast way to Fourier transform a function. In the spectral domain we employ a coordinate scaling to smoothen the branch cut found in the Green function. We have developed algorithms to perform multiplication and convolution efficiently, scaling as $O(N \log N)$ on the number of Gabor coefficients, yielding an overall algorithm that also scales as $O(N \log N)$.

5.2 Introduction

Numerical modeling of electromagnetic scattering by dielectric objects is important in many fields. To efficiently design and optimize many types of optical structures, accurate and fast numerical schemes are required to characterize their optical properties. We are interested in modeling the electromagnetic scattering from a finitely-sized dielectric structure that is illuminated by an incoming field. This type of problem includes electromagnetic band gap materials and other metamaterials [119] that can be built using dielectric objects, dielectric diffractive optics [120] such as dielectric binary diffractive lenses [121], and photonic integrated circuits [24, 25] such as grating couplers and polarization converters.

[1] This chapter was first published as the article [118].

Many numerical methods (solvers) that have been developed can solve this type of problem. In general, a trade-off between flexibility, accuracy and speed of a numerical method is made. A class of solvers that does not exploit symmetries includes local methods, such as finite difference time domain (FDTD)[26], finite element method (FEM)[32], and finite integration technique (FIT)[122]. The flexibility of these methods has led to a widespread use of these approaches in commerical software.

Here, we are interested in the scattering from a dielectric object of finite size in a homogeneous background medium. We choose to employ a global method, an integral equation approach. With this approach it is possible to exploit the translation symmetry of the homogeneous background medium. Because of this symmetry, the Green function is a function of only the distance between source and observer. As a consequence, the spatial-domain convolution between the contrast current density and the Green function can be performed in the spectral domain, where it is simply a pointwise multiplication. This pointwise multiplication can in principle be performed much faster than a convolution and therefore it can lead to an efficient algorithm.

Historically, the Conjugate Gradient Fast Fourier Transform (CGFFT) [37] was pioneering the exploitation of a spectral representation for a fast convolution. Even though CGFFT is a spatial method, with the use of FFTs the convolution is performed in a discrete spectral domain, which is fast. Improvements on this method are the Adaptive Integral Method (AIM) [42] and pre-corrected FFT (pFFT) [43], which use a meshing to accurately describe the scatterer and a grid of multipoles that approximate the radiation from the mesh at a large distance. On this grid FFTs can be used again for efficiency.

The translation symmetry can also be exploited by using a spectral discretization of the problem. Although we are interested in aperiodic problems, it is fascinating to see that for periodic problems several algorithms exist that are based on a spectral representation. A key ingredient is that periodicity inherently leads to a discrete set of modes as a basis for the spectral domain. This is exploited in Rigorous Coupled Wave Analysis (RCWA) [74, 73], the C-method [76, 123] and the Periodic Volume Integral Method (PVIM) [71].

Based on these periodic solvers it is possible to calculate the scattering by aperiodic scatterers as well. For example by using Perfectly Matched Layers (PMLs), where a scatterer of finite size is placed inside a box with absorbing sides that can be periodically repeated and solved with a periodic solver [124]. However, we are looking for a method that applies a discretization directly in the spectral domain.

Inspired by the success of CGFFT, AIM and pFFT, we understand the need for some sort of equidistant grid in the discretization. To achieve more flexibility, we intend to build an algorithm similar to meshless methods, so our discretization should be continuous in the spatial domain as well. Also inspired by the succes of the aforementioned periodic spectral methods, we would like to use a spectral discretization. The method should work for aperiodic scatterers, so we need to find a discretization that is continuous in the spectral domain as well.

A discretization with a Gabor frame [102, 100, 99] fits these requirements. We explicitly choose the Gabor frame over the Gabor basis, because within the Gabor frame the convergence of the discretization is much better than the Gabor basis, where the convergence is hindered by the Balian Low theorem [106, 107]. The Gabor frame consists of a set of window functions defined on a uniform grid that are modulated by a set of equidistant frequencies. By a Fourier transformation, the spatial Gabor frame yields a spectral Gabor frame as well. Because the Gabor frame employs window functions, it has the advantage of being localized spatially as well as spectrally. Moreover, a transformation from the spatial to the spectral domain is almost trivial. Because there is a discrete translation symmetry in the equidistant window functions, several fast methods exist to perform calculations on functions represented by a Gabor frame based on the FFT [99, 100, 125].

The Gabor frame has been used before to solve diffraction problems by line or surface scatterers as a source for Gaussian Beams [126, 127, 128, 129]. These articles show that this method can be very useful for efficient calculation of the scattered field at a large distance (in terms of wavelengths), owing to an efficient long-distance approximation. However, to solve a domain integral equation we need to calculate the field at short distances as well, so this approximation is not beneficial. Related to the Gabor frame is the Wilson basis [109]. There are some reports on using the Wilson basis [130] as a discretization for electric fields or to solve a problem using projection methods [131], but none of them mentions an optimized numerical structure that uses this type of basis and testing functions. Optimization is vital for this method to become competitive compared to other numerical methods in the sense of calculation time.

Here we show a first version of an algorithm that uses the Gabor frame in a domain integral equation. We use the Gabor frame in only one direction of the two-dimensional transverse-electric (TE) scattering problem and use a spatial discretization in the other direction just as in [71]. This also allows us to compare the accuracy for both discretizations.

5.3 The Domain Integral Equation

5.3.1 Problem Formulation

Consider a dielectric object described by a permittivity function in the x-z plane, i.e. $\varepsilon_r(x,z)$, illuminated by an incident electromagnetic field with the electric field polarized in the y direction, i.e. TE polarization. The permittivity function is different from 1 in the region bounded by $x \in [-W, W]$, $z \in [0, d]$ and we define the contrast function,

$$\chi(x,z) = \varepsilon_r(x,z) - 1, \tag{5.1}$$

where the background medium is vacuum. Owing to the two-dimensional configuration and the polarization of the electric field, the total electric field has only a y-component, $\mathbf{E}(x,z) = E(x,z)\hat{\mathbf{y}}$. The scattering setup is shown in Fig. 5.1.

We define the contrast current density by the relation $J = j\omega\varepsilon_0\chi E$ and will make the distinction between the incoming field E^i, the scattered field E^s and the contrast current

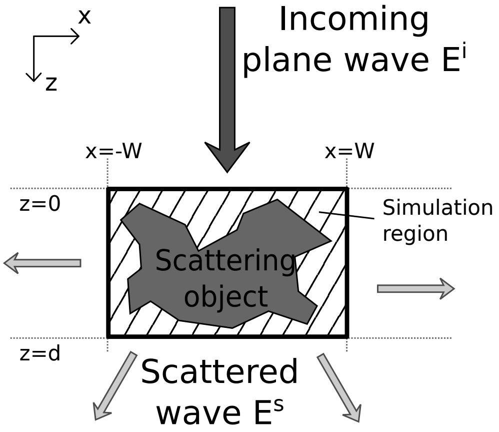


Figure 5.1: The 2D scattering problem

densities[2] J^i and J^s they induce in the dielectric scatterer

$$\begin{aligned} E(x,z) =& E^i(x,z) + E^s(x,z), \\ J^{i/s}(x,z) =& j\omega\varepsilon_0\chi(x,z)E^{i/s}(x,z). \end{aligned} \tag{5.2}$$

The Fourier transformation in the x direction is defined by

$$\hat{\varphi}(k_x) = \mathcal{F}_x[\varphi(x)](k_x) = \int_{-\infty}^{\infty} dx\, \varphi(x)e^{-jk_x x}, \tag{5.3}$$

and its inverse

$$\varphi(x) = \mathcal{F}_{k_x}^{-1}[\hat{\varphi}(k_x)](x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk_x\, \hat{\varphi}(k)e^{jk_x x}. \tag{5.4}$$

We will use k_x as a variable for all spectral functions and x as a variable for spatial functions, dropping the hat for convenience.

5.3.2 The integral form of the 2D TE scattering problem

The Maxwell equations [85], assuming time convention $e^{j\omega t}$, are given by

$$\begin{aligned} \nabla \times \mathbf{H} = j\omega\mathbf{D} \\ \nabla \times \mathbf{E} = -j\omega\mathbf{B}. \end{aligned} \tag{5.5}$$

[2] The formulation with J^i, the current induced by the incoming field, is somewhat uncommon in the literature. It has the advantage that both left and right hand side in the integral equation, Eq. (5.8), decay in a similar fashion at the end of the simulation region, where the Gabor frame ends. With a traditional formulation, the result of the Green function convolution decays differently from the incident field, that is then discretized directly in the spatial domain. This would lead to intolerable artifacts at the end of the simulation domain.

In a dielectric material we have $\mathbf{B} = \mu_0\mathbf{H}$ and $\mathbf{D} = \varepsilon_0(1+\chi)\mathbf{E}$. The spectral domain Green function for the 2D transverse electric problem can be written as

$$G(k_x, z|z') = \frac{e^{-\gamma|z'-z|}}{2\gamma}, \tag{5.6}$$

where $\gamma^2 = k_x^2 - k_0^2$ and $k_0^2 = \omega^2\varepsilon_0\mu_0$.

We can express the scattered field by the integral representation[3]

$$E^s(k_x, z) = \frac{1}{j\omega\varepsilon_0}\int_0^d dz' G(k_x, z|z')k_0^2 J(k_x, z'), \tag{5.7}$$

which together with Eq. (5.2) can be written as

$$\begin{aligned} -\, k_0^2\chi(x,z)\mathcal{F}_{k_x}^{-1}\left[\int_0^d dz'\, G(k_x, z|z')J^i(k_x, z')\right](x) = \\ -J^s(x,z) + k_0^2\chi(x,z)\mathcal{F}_{k_x}^{-1}\left[\int_0^d dz'\, G(k_x, z|z')J^s(k_x, z')\right](x), \end{aligned} \tag{5.8}$$

with the advantage that the contrast current density J is compactly supported, since the contrast χ has a finite support. Now the left-hand side depends on the known incident field $E^i(k_x, z)$ and the right-hand side can be viewed as an operator working on the scattered current, which we want to calculate. With the Fourier transforms we emphasize that we want to do the convolution in the x-direction in the spectral domain. In the z direction the convolution is done spatially.

5.3.3 Discretization along the z direction

Along the z direction we will use piecewise-linear functions Λ_n as expansion functions. These expansion functions are given by

$$\Lambda_n(z) = \begin{cases} 1 - \frac{|z-n\Delta|}{\Delta} & \text{if} \quad |z-n\Delta| < \Delta \\ 0 & \text{if} \quad |z-n\Delta| > \Delta \end{cases}, \tag{5.9}$$

with Δ the step size in the z discretization. This discretization is convenient, because the electric field is continuous and the contrast density is continuous in regions where χ is continous. When the contrast current density $J(k_x, z)$ is expanded into the expansion functions, we obtain

$$J(k_x, z) \approx \sum_{n=0}^{N_z} J_n(k_x)\Lambda_n(z), \tag{5.10}$$

where N_z is the total number of expansion functions in the z direction. We use Dirac-delta testing functions in the z direction to find the coefficients[4]$J_n(k_x)$, since it was observed

[3] Note that the Green function is defined slightly different from Chapter 2

that this leads to a well-conditioned linear system for a similar formulation for periodic scattering problems [71]. Now we have set up the sets of testing and expansion functions, they will be used on the integral equation Eq. (5.8). Because the Green function is semi-separable in z, it is advantageous to write

$$\begin{aligned} E_n^s(k_x) &= \int_0^{n\Delta} dz' \sum_{n'=0}^{n+1} G(k_x, n\Delta|z') k_0^2 J_{n'}(k_x) \Lambda_{n'}(z') + \\ &+ \int_{n\Delta}^{d} dz' \sum_{n'=n-1}^{N_z} G(k_x, n\Delta|z') k_0^2 J_{n'}(k_x) \Lambda_{n'}(z') \\ &= K_n^u(k_x) + K_n^d(k_x). \end{aligned}$$

Here the integral is split along z' in the two integrals $K_n^u(k_x)$ and $K_n^d(k_x)$. There exists a recursive algorithm to find these two integrals using the fact that $J_n(k_x)$ is nonzero only on the support of the contrast source. We will only perform the calculation for K_n^u, since K_n^d is similar. Working this out we find

$$\begin{aligned} K_{n+1}^u(k_x) = K_n^u(k_x) e^{-\gamma\Delta} \quad & - J_n(k_x) \int_0^{\Delta} dz' k_0^2 \Lambda_0(z') \frac{e^{-\gamma(\Delta - z')}}{2\gamma} \\ & - J_{n+1}(k_x) \int_0^{\Delta} dz' k_0^2 \Lambda_1(z') \frac{e^{-\gamma(\Delta - z')}}{2\gamma} \\ = K_n^u(k_x) e^{-\gamma\Delta} \quad & + J_n(k_x) I_m^u(k_x) + J_{n+1}(k_x) I_e^u(k_x), \end{aligned} \tag{5.11}$$

where I_m^u and I_e^u are introduced for the result of the integrals over z'. With this method we can calculate $K_n^{u/d}(k_x)$ for all n in only N_z steps. This yields an algorithm of linear complexity in N_z.

5.3.4 Far-field intensity

In the end we are interested in the far field due to the contrast source. Since the electric field at a large distance is needed, we use a different method than in the previous section to calculate this.

We start again from the integral representation in Eq. (5.7). Using the discretization in the z direction this is written as

$$E^s(k_x, z) = \frac{j}{\omega\varepsilon_0} \int_{-\Delta}^{\Delta} dz' \sum_{n=0}^{N_z} k_0^2 J_n(k_x) \frac{e^{-j\sqrt{k_0^2 - k_x^2}|n\Delta + z' - z|}}{2\sqrt{k_0^2 - k_x^2}} \Lambda(z'). \tag{5.12}$$

By taking the inverse Fourier transformation in the x direction and by changing to polar coordinates $x = R\cos\varphi$, $z = R\sin\varphi > 0$ in the limit $R \to \infty$, we obtain

$$E^s(R, \varphi) = \frac{j}{2\omega\varepsilon_0} \sum_{n=0}^{N_z} k_0^2 J_n(-R\cos\varphi) e^{jk_0 n\Delta \sin\varphi} \frac{2 - 2\cos(\Delta\sin\varphi)}{\Delta\sin^2\varphi} (I_0(R) + jH_0(R)),$$

[4] $J_n(k_x)$ is the contrast current density at $z = n\Delta$ and should not be mistaken for a Bessel function

with I_0 the Bessel function of the first kind of order zero and H_0 the Struve function of order zero.

The field strength depends on the distance and owing to energy conservation the field strength will decay as $1/\sqrt{R}$. Therefore, we will employ the scattering strength $S(\varphi)$ as a function of the angle φ, defined as

$$S(\varphi) = \lim_{R \to \infty} R|E^s(R, \varphi)|^2. \tag{5.13}$$

5.4 Discretization in the x direction via a Gabor frame

To discretize Eq. (5.8) in the x-direction we will employ a Gabor frame. We begin with a short description of the way we have implemented the Gabor frame. Then we list the mathematical operations needed to apply on functions represented using a Gabor frame and explain their implementation. We end by discussing the uniqueness of the numerical solution.

5.4.1 Definition

Following the exposition in [99] to introduce the Gabor frame and its properties, we show how we have implemented the Gabor frame. A Gabor transformation makes extensive use of a window function $g(x)$, for which we choose here the Gaussian

$$g(x) = 2^{\frac{1}{4}} e^{\left(-\pi \frac{x^2}{X^2}\right)}, \tag{5.14}$$

with X the parameter that defines the width of the window function. The Gabor frame is defined by

$$g_{mn}(x) = g(x - m\alpha X) e^{jn\beta K x}, \tag{5.15}$$

with the modulation step size K, defined by $KX = 2\pi$, and α and β two constants defining the oversampling, obeying $\alpha\beta \leq 1$. Function values of a function $f(x)$ that is represented by Gabor coefficients f_{mn} can be calculated by

$$f(x) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f_{mn} g_{mn}(x) = \sum_{m,n} f_{mn} g_{mn}(x). \tag{5.16}$$

Here the m sum is over spatial windows and the n sum is over the modulation frequencies, in practice these sums are truncated. To calculate the coefficients f_{mn} of a given function $f(x)$ it is common to write

$$f_{mn} = \int_{-\infty}^{\infty} dx\, f(x) \eta^*_{mn}(x), \tag{5.17}$$

where we introduced the dual window function $\eta(x)$ that forms the dual frame to $g_{mn}(x)$. Here * denotes complex conjugation. The indices have a meaning similar to Eq. (5.15), i.e.

$$\eta_{mn}(x) = \eta(x - m\alpha X) e^{jn\beta K x}. \tag{5.18}$$

Several methods to calculate a dual window function $\eta(x)$ are described in [99, 100].

5.4.2 The choice for the Gabor frame

From Eq. (5.17) it is clear that the dual window function $\eta(x)$ is very important to achieve an efficient representation of a function. The Balian Low theorem states that a good dual window only exists when oversampling is employed ($\alpha\beta < 1$ in Eq. (5.17)) [106, 107]. The dual window of an oversampled Gabor frame can be chosen in a way that it decays exponentially both spatially and spectrally [99, 100, 125]. We use the dual window obtained through the Moore-Penrose inverse, since it is the one most commonly used and it exhibits good convergence.

After the construction of a dual window, it is possible to calculate the Gabor coefficients of functions by Eq. (5.17). There are faster ways to calculate Gabor coefficients for rational oversampling $\alpha\beta = q/p$ with $p > q \in \mathbb{N}$. In that case it is possible to use FFTs in $O(pn \log n)$ operations, with n the total number of Gabor coefficients [125, 100]. This method is faster, but it uses an equidistant sampling on the function from which it calculates the Gabor coefficients, so it only works well for sufficiently smooth functions.

When functions have discontinuities, we should evaluate the integrals in Eq. (5.17), but we found that this is a very slow method. Another option is to approximate this result by using oversampling in the FFT method to calculate the coefficients f_{mn} of $f(x)$ for a broader range of n and later discard the extra coefficients. When more modulation frequencies are used, the function is sampled on a finer lattice and therefore the approximation error is smaller.

5.4.3 Building blocks

To solve an integral equation using the method outlined in Section 5.3.2, we need methods to manipulate functions represented by Gabor coefficients. It is of course important that these algorithms scale well, $O(N \log N)$ or better, where N denotes the number of Gabor coefficients.

From Eq. (5.11) and Eq. (5.2) we see that the following operations on Gabor coefficients are required

- Addition
- Multiplication by a scalar
- Fourier transformation and inverse Fourier transformation
- Multiplication by a set of Gabor coefficients
- Convolution with a set of Gabor coefficients
- Inner product

The first two operations are trivial and can be performed per coefficient. The convolution can be calculated using a combination of a Fourier transformation, a spectral multiplication and an inverse Fourier transformation. In the following we explain how to calculate the Fourier transform, the multiplication, and inner product on sets of Gabor coefficients.

5.4.4 Fourier transform

A very convenient property of the Gabor frame functions is that their Fourier transform also yields a Gabor frame. We define a spectral frame function $\hat{g}_{nm}(k)$ by

$$\hat{g}_{nm}(k) = \mathcal{F}_x[g_{mn}(x)](k) = \hat{g}(k - n\beta K)e^{-jm\alpha Xk}e^{-2\pi j\alpha\beta mn}, \tag{5.19}$$

with the spectral window function $\hat{g}(k) = \mathcal{F}_x[g(x)](k)$.

On the level of Gabor coefficients, the Fourier transform $\hat{f}(k)$ of the function $f(x)$ can be represented by

$$\hat{f}(k) = \sum_{m,n} \hat{f}_{mn}\hat{g}_{mn}(k), \tag{5.20}$$

where

$$\hat{f}_{mn} = f_{nm}e^{2\pi j\alpha\beta mn}. \tag{5.21}$$

With this definition of a spectral Gabor frame, the operation of Fourier transformation is computationally very easy and can be done in $O(N)$ steps, where N is the total number of Gabor coefficients. Of course it would have been possible to use a different Gabor frame for the spectral representation, but this one has the advantage that a Fourier transform can be obtained from the simple relation Eq. (5.21). When the spectral frame is not simply the Fourier transform of the spatial frame, such a simple relationship does not exist.

5.4.5 Multiplication of two sets of Gabor coefficients

Although there exists an easy and very fast way to calculate a Fourier transform of a function, with Gabor coefficients it is the multiplication of functions that takes most of the time. When we multiply function $v(x)$ and $w(x)$ their product, $f(x)$ can be written as

$$f(x) = v(x)w(x) = \sum_{m,n}\sum_{k,l} v_{mn}w_{kl}g_{mn}(x)g_{kl}(x). \tag{5.22}$$

When we define the functions $A_j(x) = g(x)g(x - j\alpha X)$, we can compute its Gabor transform $A_{j,pq}$, where the x-dependence in $A_j(x)$ is replaced with Gabor indices pq. These coefficients are useful for the calculation of the product

$$g_{mn}(x)g_{kl}(x) = \sum_{r,s} A_{k-m;r-m,s-n-l}\; g_{rs}(x)e^{2\pi j\alpha\beta m(n+l-s)}.$$

Now we would like to calculate the Gabor coefficients f_{mn} of $f(x)$ directly from v_{mn} and w_{mn}. We can use $A_{j;pq}$ in Eq. (5.22) to obtain

$$f(x) = \sum_{r,s} \sum_{m,n} \sum_{k,l} A_{k-m;r-m,s-n-l}\ v_{mn}\, w_{kl}\, g_{rs}(x) e^{2\pi j\alpha\beta m(n+l-s)}. \tag{5.23}$$

After two variable substitutions, $a = n + l$ and $b = k - m$

$$f_{rs} = \sum_{b=-B}^{B} \sum_{m,a} T_{b;m,a}\ A_{b;r-m,s-a} e^{2\pi j\alpha\beta m(a-s)} \tag{5.24}$$

is obtained, with

$$T_{b;m,a} = \sum_{n} v_{mn} w_{m+b,a-n}, \tag{5.25}$$

and B defining the truncation number in the b summation. The $T_{b;m,a}$ can be calculated efficiently using FFTs, because they are calculated from a discrete convolution in n. B does not need to be large, because $A_b(x)$ decays quickly when the overlap between the window functions is small. For example with $\alpha = \beta = \sqrt{2/3}$ oversampling and the window function as in Eq. (5.14) we find $|A_{\pm 3}(x)|_{\mathcal{L}^2} < 10^{-5}|A_0(x)|_{\mathcal{L}^2}$. In our case a sum of 5 terms ($B = 2$) would already be enough for an accuracy much better than 10^{-4} in this sum. So T is only needed for a small number of b values. The number of significant m and a combinations is of the order N, the total number of coefficients of the input functions.

The exponential factor in Eq. (5.24) has only p different values for a $q/p = \alpha\beta$ oversampled Gabor frame. We can exploit this, by carrying out the m summation for every different value of $s \in \{1 \dots p\}$ individually. Then we can use that the m and a summation can be written as a convolution and therefore this summation can be done efficiently using FFTs as well. The total complexity scales like $NpB \log N$, with N the total number of Gabor coefficients.

5.4.6 Inner products

The $\mathcal{L}^2(\mathbb{R})$ inner product of functions $f(x)$ and $h(x)$ can in principle be calculated from

$$< f, h >= \int_{-\infty}^{\infty} dx\ f(x)h^*(x) = \sum_{k,l,m,n} f_{kl}h^*_{mn} \int_{-\infty}^{\infty} dx\ g_{kl}(x)g^*_{mn}(x). \tag{5.26}$$

Because of the spectral and spatial translation symmetry, the inner product between the frame functions $g_{mn}(x)$ and $g_{kl}(x)$ can be further simplified to

$$\begin{aligned} < g_{kl}, g_{mn} > &= \int_{-\infty}^{\infty} dx\ g(x - \alpha kX)g^*(x - \alpha mX)e^{j(l-n)\beta Kx} \\ &= e^{2\pi j\alpha\beta k(l-n)} \int_{-\infty}^{\infty} dx\ g(x)g^*(x - \alpha(m-k)X)e^{i(l-n)\beta Kx} \\ &= e^{2\pi j\alpha\beta k(l-n)} \mathcal{M}_{k-m,l-n}, \end{aligned} \tag{5.27}$$

so it can be used to calculate the inner product

$$< f, h >= \sum_{k,l,m,n} f_{kl} h^*_{mn} e^{2\pi j \alpha\beta k(l-n)} \mathcal{M}_{k-m,l-n}. \tag{5.28}$$

Here we see the same exponential $e^{2\pi j\alpha\beta k(l-n)}$ as in the multiplication. Because of the rational oversampling there are only p unique sequences for the exponential as a function of $(l - n)$. Now the discrete convolution of h_{mn}, with $e^{2\pi j\alpha\beta k(l-n)}\mathcal{M}_{k-m,l-n}$, can be done using FFTs again for each of these p unique discrete convolution kernels.

The methods described in the previous subsections are exact, except for the multiplication and inner products of coefficients, which converge exponentially with the number of terms included.

5.4.7 The discrete formulation

By discretizing Eq. (5.8) as described in Sections 5.3.3 and 5.4, our problem can be written as

$$A \circ J^s = -(1 + A) \circ J^i, \tag{5.29}$$

with

$$J^{i/s}(k_x, z) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} \sum_{\ell=0}^{N_z} J^{i/s}_{mn,\ell}\, g_{mn}(k_x)\Lambda_p(z).$$

The operation $A\circ$ then represents the Green function G convolution and contrast $k_0^2\chi$ multiplication and an subtraction of J, as in the right-hand side of Eq. (5.8).

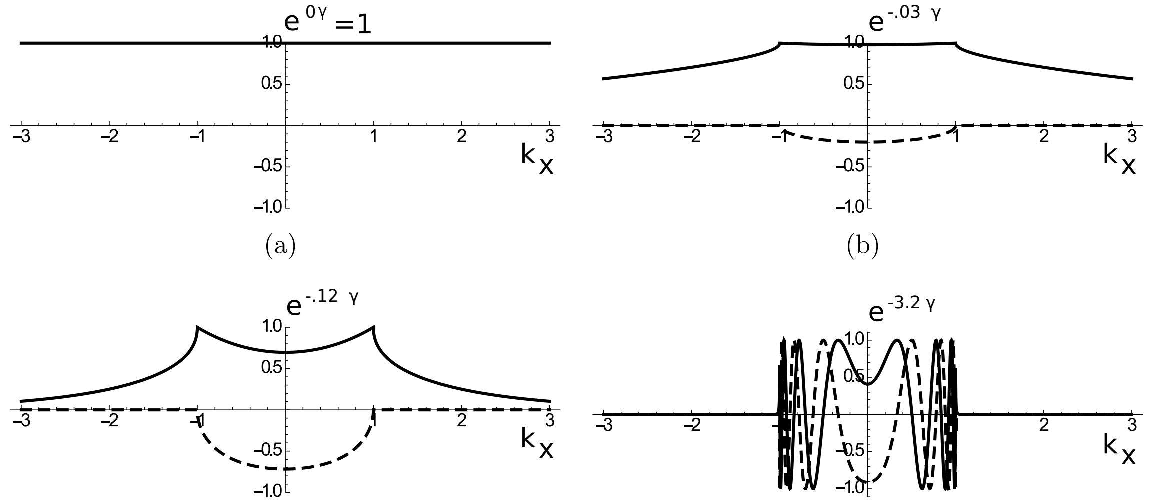
When we view J^s and $A \circ J^i$ as lists of numbers, then the operator A can be viewed as a matrix and we can write $\underline{\underline{A}} \cdot \underline{j} = \underline{b}$.

Normally one would continue by inverting this matrix, but this matrix is not invertible because of the oversampling in the Gabor frame. It is for example possible to write a function that is zero everywhere, with non-vanishing coefficients; in other words the matrix $\underline{\underline{A}}$ defined through a Gabor-frame representation has a nontrivial null-space. However, since we make a fixed choice for a certain dual window η (see Section 5.4.1), the representation of every function in the Gabor frame becomes unique. We used GMres to solve the overall linear system.

5.5 Singular behaviour of the Green function

When we calculate the Gabor coefficients of $K_n^{u/d}(k_x)$ for every n recursively according to Eq. (5.11), there can be severe discontinuities in the derivative of $K_n^{u/d}(k_x)$ around $k_x = \pm k_0$. The main reason is the first term in Eq. (5.11)

$$K^u(k_x, z) = K^u(k_x, z - \Delta)e^{-\gamma\Delta} + \cdots = K^u(k_x, 0)e^{-n\Delta\gamma} + \ldots, \tag{5.30}$$



(a) (b) (c) (d)

Figure 5.2: The spectrum (real part: solid, imaginary part: dashed) of a wave ($k_0 = 1$) emanating from an electric-field point source at $z = 0$, $\exp -\gamma z$. In (a) through (d) we subsequently plotted the spectrum at $z = 0$, $z = 0.03\lambda$, $z = 0.12\lambda$, $z = 3.2\lambda$. It can be clearly seen how the smoothness deteriorates for large z near $k_x = \pm k_0$.

with $z = n\Delta$ the height where we would like to evaluate the field. In the present discussion we ignore everything related to current density.

In Fig. 5.2(a) we show a spectral representation of an electric field due to a point source in the plane of the source. In Fig. 5.2(b) we show the field of this point source propagated one Δ step in the z direction. The discontinuous derivative is clearly visible, but does not look severe. After propagating further in the z-direction (Fig. 5.2(c)), the discontinuity in the derivative becomes more severe. After several wavelengths of z propation, oscillations start to appear in $K^u(k_x, z)$ for k_x between $-k_0$ and k_0 in Eq. (5.30), as is shown in Fig. 5.2(d).

The strong discontinuity of its derivative makes it hard to represent this function by Gabor coefficients, because a finite set of Gabor coefficients is limited in its spectral representation. It is important to note that although the function $e^{-\gamma z}$ is highly oscillatory for large z, its inverse Fourier transform $\mathcal{F}_{k_x}^{-1}[e^{-\gamma z}](x)$ is continuous, but extends to large distances in x. This implies that it should still be possible to find a good representation by Gabor coefficients that holds only in the simulation domain $x \in [-W, W]$. Intuitively, one would truncate the number of spatial Gabor coefficients. However, it is inaccurate to multiply spatially-truncated representations of $e^{-\gamma\Delta}$, as would be done in the recursive algorithm Eq. (5.11). The error results from the fact that a spectral multiplication corresponds to a spatial convolution. Spatially (the inverse Fourier transform of) $e^{-\gamma\Delta}$ does not have a bounded support. For a convolution of two functions with infinite spatial support, information about the entire support of these functions is needed. For this reason we would need the Gabor coefficients to represent the function over the entire x-axis, which takes an infinite number of coefficients.

A more convenient method is to use a coordinate scaling in the spectral domain. The goal of the scaling is to allow for an accurate and efficient representation of the field $K_n^{u/d}(k_x)$. In principle there are many choices for a scaling function that would work, but there is also a $1/\gamma$ singularity in the Green function Eq. (5.6). We will incorporate the $1/\gamma$ singularity in the Jacobian of the scaling, so it will be canceled. Scaling a function that is represented using Gabor coefficients is not a trivial task, since it is part of the core of the algorithm, and it needs to result in fast calculations.

5.5.1 The scaled coordinates

Fig. 5.3(b) shows an example for an unscaled $e^{-\gamma H}/\gamma$. The many oscillations and the two asymptotes make it difficult to represent this function using a Gabor frame. We want to make a good approximation of the Green function of waves generated at $z = 0$ that propagate all the way to $z = H$, which is the largest distance and therefore the worst case. We make use of three scaling functions s_1, s_2 and s_3, which will be optimized for different parts of the spectral domain. From this set of functions one scaling function s will be composed, which uses the most appropriate scaling function for every part of the k_x axis.

We choose the first part of the scaling function for $|k_x| < k_0$ such that the $\frac{1}{\gamma}$ factor is incorporated in the Jacobian of the scaling by using a sine transformation

$$k_x(\tau) = s_1(\tau) = k_0 \sin(c\pi\tau/2). \tag{5.31}$$

The second part of the scaling function works just outside that region, for $|k_x| > k_0$, and it can be seen as a continuation of the scaling function s_1

$$s_2(\tau) = \begin{cases} k_0 \cosh(c\pi(\tau - 1/c)/2) & \text{if } \tau > 1/c \\ -k_0 \cosh(c\pi(\tau + 1/c)/2) & \text{if } \tau < -1/c. \end{cases} \tag{5.32}$$

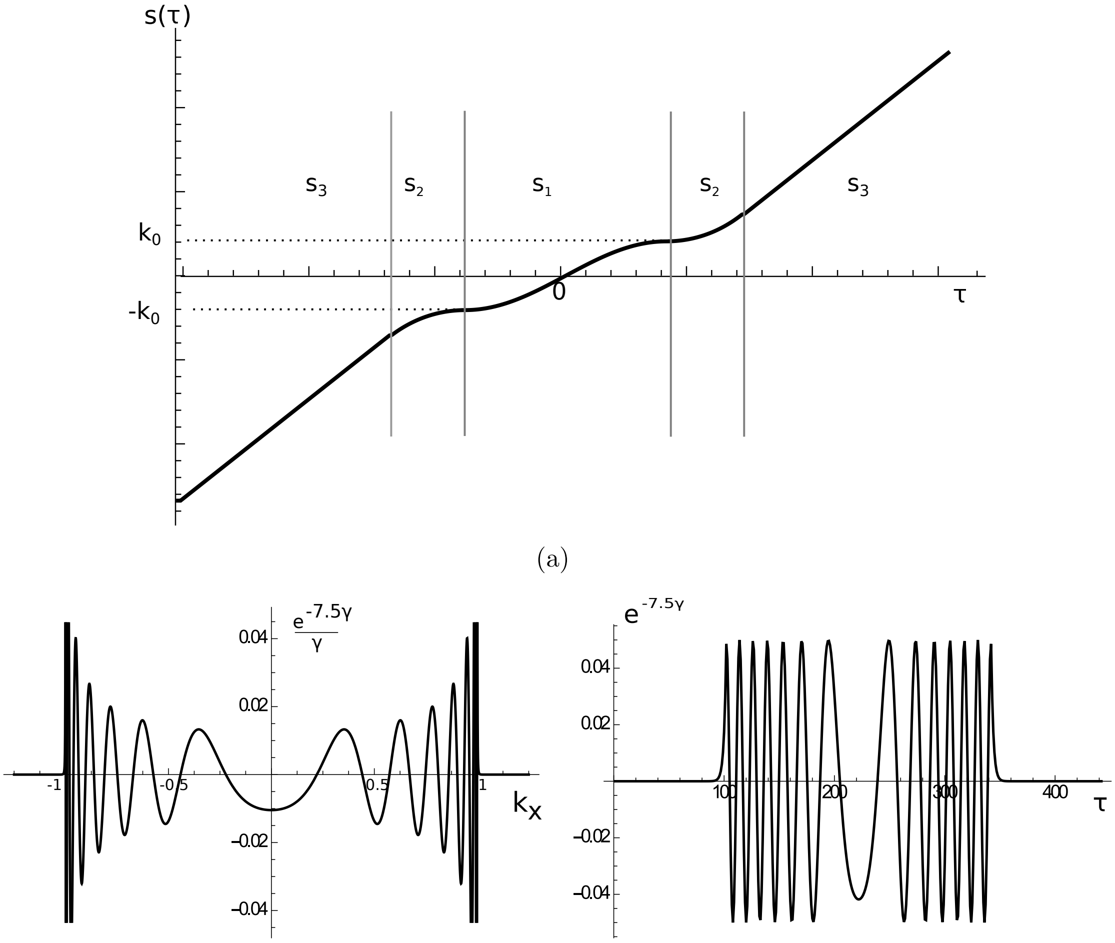
These scaling functions s_1 and s_2 will also smoothen the oscillations in $e^{-\gamma H}$ when the constant c is choosen small enough, depending on H and k_0. To decide on a value for c the $1/\gamma$ derivative is most important when H is small. When the simulation height H gets larger $e^{-\gamma H}$ becomes the governing factor to determine c. A simple plot of $e^{-\gamma H}$ as in Fig. 5.3(c) can show whether or not the choice for the constant c is good; it should show a well sampled continuous curve.

The third part of the scaling function has the constraint that the derivative of the scaling function should never exceed one. Otherwise information from the input function is lost due to a coarser sampling in the scaled coordinates. It is defined by

$$s_3(\tau) = \tau + d, \tag{5.33}$$

with d a constant that allows to shift this function to make the transition to the other scaling functions continuously differentiable. In Fig. 5.3(a) we show how to put these functions together: the scaling function is put together such that for every $k_x = s(\tau)$ the scaling function with the smallest derivative is used, and that transitions between the three scaling functions are smooth. For large values of c (small H) another region with scaling function s_3 in the middle is possible.

With this scaling an efficient representation of the electric field over the complete x axis with a limited number of k_x functions is made. Around the points $k_x = \pm k_0$ we have already compensated the $1/\gamma$ singularity in the Greens function with the Jacobian. This has the advantage that we do not have to worry about asymptotes, since the function is bounded now. For the region with scaling function s_3 we still need to incorporate the $1/\gamma$, but it is well behaved there.



(a)

(b)

(c)

Figure 5.3: (a) A typical scaling function for $k_0 = 1$. (b) The unscaled $e^{-7.5\gamma}/\gamma$. (c) The scaled $e^{-7.5\gamma}$ (the factor $1/\gamma$ disappears with the Jacobian).

5.5.2 Interpolating to and from the scaled coordinates

From equidistant to scaled

With the algorithm from [100] we can go from a spectral-domain current with equidistant k_x sampling, $J_n(k_m) = J_n(m\Delta_k)$ with $m \in \mathbb{Z}$, to Gabor coefficients $J_{n;i,j}$ and back again. What we need is to have function values $\tilde{J}_{n;m} = J_n(\tau_m)$ on the non-equidistant scaled coordinates $\tau_m = s(m\Delta_k)$ as well with $s(k_x)$ the scaling function. To obtain that, we need to interpolate the equidistant sampling in a smart manner, since linear interpolation leads to intolerably large errors. The tilde will denote functions on the scaled coordinates.

From Figs. 5.4(a) and (b) it is clear that the sampling rate is not high enough to get a good approximation of the current density with the algorithm from [100] by using a high-order interpolation only. To get a better approximation we use oversampling: we pad the Gabor coefficients with zeroes for high $|j|$ in $J_{n;i,j}$ and then transform the Gabor coefficients to an equistant sampling. In Figs. 5.4(c) and (d) it is shown that the combination of oversampling and a fifth order Hermite [132] interpolation reduces the error to a tolerable level.

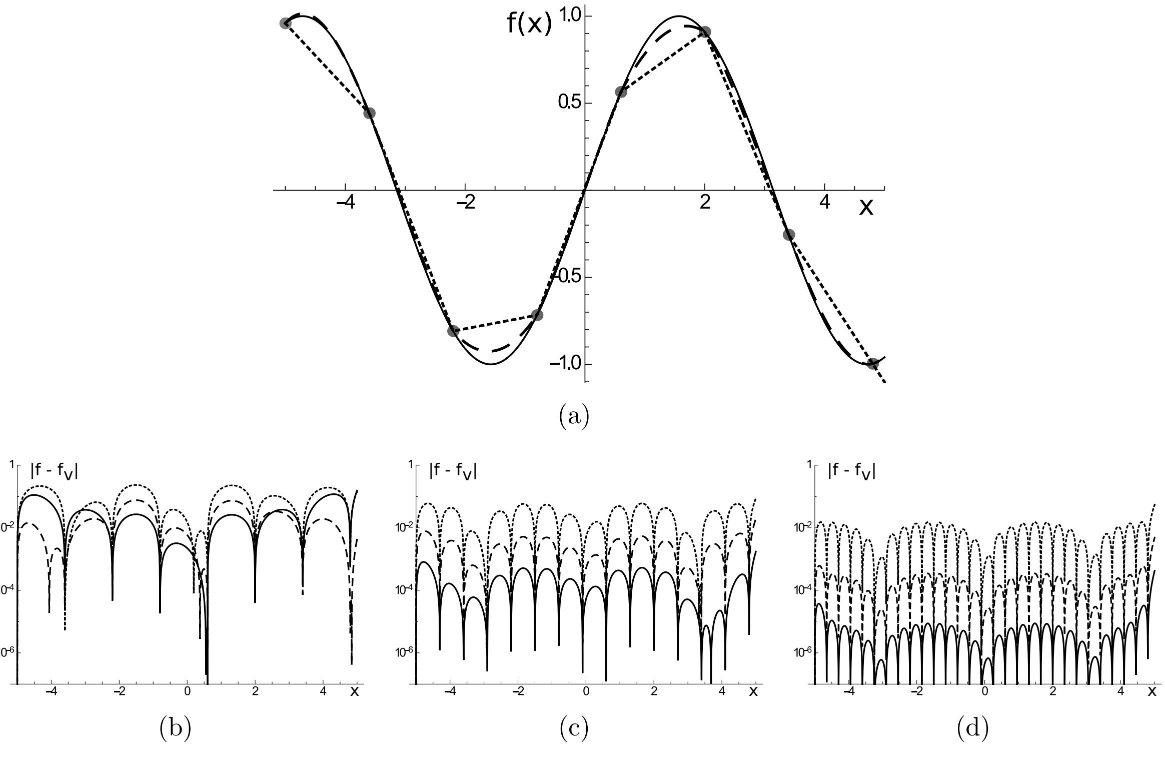


Figure 5.4: (a) Solid line: $f_V(x)$ a sine function as validation. Dotted: A linear interpolation f of the validation f_V, dashed: A third order Hermite spline interpolation f of the validation f_V. (b) The absolute error of the approximations $|f(x) - f_V(x)|$: linear interpolation (dotted), 3rd order Hermite interpolation (dashed), 5th order Hermite interpolation (solid). (c) The same for double sampling. (d) The same for quadruple sampling.

From scaled to equidistant

To go back from the scaled coordinates to equidistant coordinates, the total number of samples is reduced, especially around the singularities. We need to be careful not to throw away any important information in the process of resampling.

We transform directly from the scaled spectral coordinates to equidistant spatial coordinates on the interval $x \in [-W, W]$, since only spatially we are able to make a good approximation using a finite number of Gabor coefficients. Let us start from the inverse Fourier transform of the scaled-coordinate electric field

$$E_n(x_m) = \int_{-\infty}^{\infty} d\tau\, \mathcal{J}_s(\tau)\hat{E}_n(s(\tau))e^{js(\tau)x_m} = \int_{-\infty}^{\infty} d\tau \tilde{E}_n(\tau)e^{js(\tau)x_m}, \tag{5.34}$$

with $\mathcal{J}_s(\tau) = H_s(\tau)/\gamma$ with $H_s(\tau)$ the Jacobian of the scaling function s, chosen to incorporate the $1/\gamma$ factor from the Green function in Eq. (5.6). The x_m values are on an equidistant grid, such that this list can be used to transform to Gabor coefficients. The bounded function $\tilde{E_n}(\tau) = \mathcal{J}_s(\tau)E_n(s(\tau))$ is defined by the list of function values on τ_m through a linear interpolation

$$\tilde{E}_n(\tau) = \sum_m \Lambda_m(\tau)E_{n;m}\mathcal{J}_s(\tau), \tag{5.35}$$

with Λ the piecewise-linear function in Eq. (5.9), with $\Delta = \Delta_k$ and $E_{n;m} = E(s(m\Delta_k)) = E(\tau_m)$, a list of function values as described in Section 5.5.2. The Jacobian is chosen in such a way so as to cancel the $1/\gamma$ in the Green function Eq. (5.6). The linear interpolation is not ideal, but since we are downsampling, the error from this approximation is less of a problem than before, while upsampling.

Now we need to calculate the Fourier integral Eq. (5.34). Using an FFT would be desirable, but this is not possible because of the scaling. What we can do is approximate the integral in Eq. (5.34) by

$$E(x_m, z_n) \approx \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} e^{jn\Delta_k x_m}\int_{(n-\frac{1}{2})\Delta_k}^{(n+\frac{1}{2})\Delta_k} dk\tilde{E}_n(s^{-1}(k)). \tag{5.36}$$

Here we recognize a discrete Fourier transform with sampling distance Δ_k and function values given by the integral. The problem here is that we make the approximation

$$\int_{(l-\frac{1}{2})\Delta_k}^{(l+\frac{1}{2})\Delta_k} dk\tilde{E}_n(s^{-1}(k))e^{jkx_m} \approx e^{jl\Delta_k x_m}\int_{(l-\frac{1}{2})\Delta_k}^{(l+\frac{1}{2})\Delta_k} dk\tilde{E}_n(s^{-1}(k)),$$

which only works for $|x_m|$ small. When there are many samples in $\tilde{E}_n(k)$ in the range of one Δ_k- integral, and when $|x_m|$ is larger, this approximation breaks down. Note that the scaling function has been choosen in such a way that the integral on the right-hand side is approximated very well. The problem is that we make a poor approximation when the complex exponential is put outside the integral.

When $\tilde{E}_n(s^{-1}(k))$ is multiplied by $e^{X_\ell k}$, we should see a shift over X_ℓ in its Fourier transform. Since we can calculate the Fourier transform of $\tilde{E}_n(s^{-1}(k))$ accurately around $x_m \approx 0$ (Eq.(5.36)), we can calculate the Fourier transform around X_ℓ accurately as well by

$$E_{n;\ell}(x_m) \approx \frac{1}{2\pi} \sum_l e^{jn\Delta_k(x_m - X_\ell)} \int_{(l-\frac{1}{2})\Delta_k}^{(l+\frac{1}{2})\Delta_k} dk \tilde{E}_{n;\ell}(s^{-1}(k)) e^{jkX_\ell}, \tag{5.37}$$

which gives us approximations accurate around each X_ℓ value. To calculate from this the electric field accurately at x_m a linear interpolation is employed. First we choose ℓ such that $X_\ell < x_m < X_{\ell+1}$, then the field at x_m can be found by

$$E_n(x_m) = \frac{(x_m - X_\ell)E_{n;\ell}(x_m, z_n) + (X_{\ell+1} - x_m)E_{n;\ell+1}(x_m)}{X_{\ell+1} - X_\ell}. \tag{5.38}$$

We used this linear ℓ interpolation with 9 different spatial sampling points x_0. This was enough to get a $3 \cdot 10^{-3}$ relative error on a practical example of a contrast current. This will be enough for our purpose, but we can improve the accuracy by using higher order Hermite interpolations and by taking more X_ℓ values.

It is important to remark here that the distance between the samples that make up $\tilde{E}_n(\tau)$ is less than or equal to Δ_k. This means that for the part that is scaled using s_3, the sampling is the same for the FFT. For the parts with s_1 and s_2 there is downsampling.

To summarize, our algorithm consists of the following steps:

- Start with a list of scaled spectral samples of a function, $\tilde{E}_{n;m}$.
- Interpolate using Eq. (5.35)
- Multiply the interpolation by exponentials $E_{n;\ell}(\tau_m) = e^{jX_\ell s^{-1}(\tau_m)} E_n(\tau_m)$.
- Integrate over the scaled interpolation function to obtain a list with (a coarser) equidistant sampling for every value of ℓ (the integral in Eq. (5.37)).
- Use an FFT on each ℓ's list to get results that are a good approximation around X_ℓ.
- Interpolate between the result ℓ and $\ell + 1$ using Eq. (5.38).

5.5.3 Various representations

Now to solve the integral equation Eq. (5.8), the challenge is to multiply by the contrast current in the spatial domain and multiply by the various parts of the Green function (e.g. Eq. (5.11)) in the spectral domain.

We start from a spatial representation of the electric field $E_n(x)$, spatial because it needs to be multiplied by the contrast function Eq. (5.1) to arrive at the current distribution $J_n(x) = \chi_n(x)E_n(x)$. The Gabor coefficients of χ are calculated using the fast algorithm in [100] to approximate Eq. (5.17). Cutting off the highest spectral parts of the current distribution does not deteriorate the accuracy very much, so we might as well bound it spectrally. This means that Gabor coefficients are the most convenient representation for $J_n(x)$.

To calculate the convolution of the Green function with the current distribution we would like to use spectral domain in the x direction (see Eq. (5.11)). The choice of the scaled coordinates both compensates the divergent $1/\gamma$ in the Green function, and takes care of rapid oscillations in the Green function for large z propagation.

In Fig. 5.5 the flow of our algorithm through the various representations is represented.

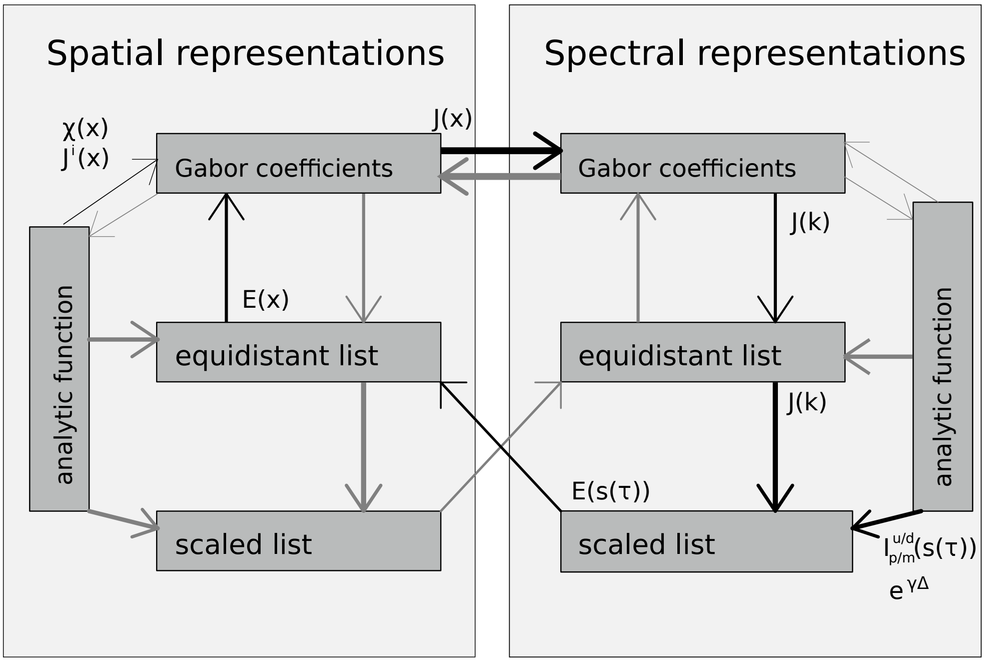


Figure 5.5: The various transformations and representations we have explained in Section 5.5. The thickness of the arrows signifies the speed of the transformation qualitatively. Thicker means faster. The black arrows signify the different transformations that are actually used in the algorithm, the grey ones are possible, but not used.

5.6 Results

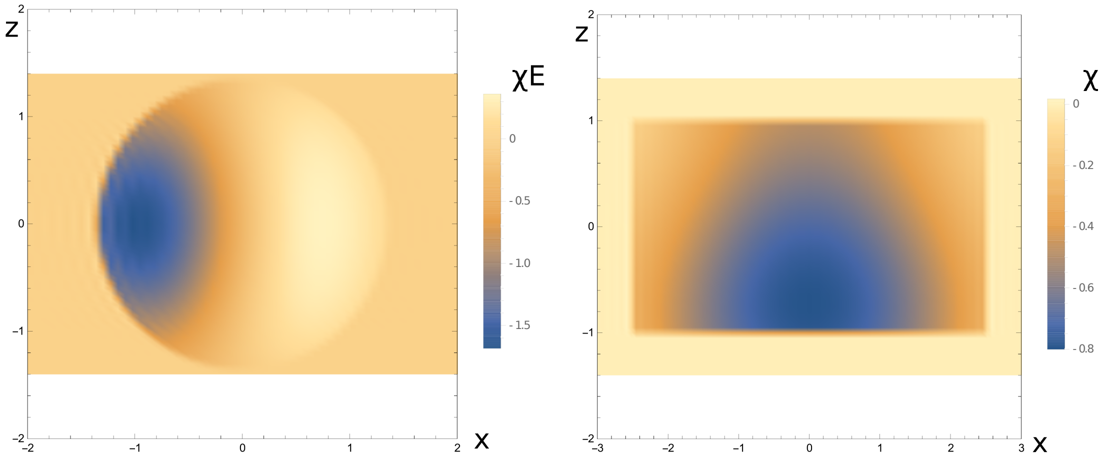
We tested the accuracy of our algorithm on a circular scatterer, a rectangular scatterer and a small grating described in Table 5.1.

The incoming field is defined as $E^i(x,z) = E_0 \exp(jk_0(x\cos\theta + z\sin\theta))$, with $E_0 = 1\,V/m$. For the circle an analytical solution exists [133], which was used to validate our results. For the rectangle we used a numerical solution calculated using the JCMWave software package as a reference [134]. In the z-direction we choose $\Delta = 0.05m$. In the x-direction we used a Gabor frame with 13 ($m \in \{-6,\ldots,6\}$) spatial windows and 7 ($n \in \{-3,\ldots,3\}$) modulation frequencies (see Eq. (5.16)). The oversampling was at

Parameters	Circle	Rectangle	Grating
ε_r of object	2	2	2
dimension of object (m)	$r = 1.35$	2.0×5.0	five blocks, 1×1.4, spacing 2
wavenumber k_0 (m^{-1})	1.45	0.7388	1.5
angle of incidence θ	270°	0°	45°
validation	Analytic solution	JCMWave	None

Table 5.1: Specification of the validation cases

$\alpha\beta = 2/3$ and $X = 0.5m$ (see Eq. (5.15)). This means that per meter there are are around 17 coefficients in the x direction and 20 in the z direction. We used 16 ℓ values in going from scaled spectral sampling to the equidistant spatial sampling (see Section 5.5.2).



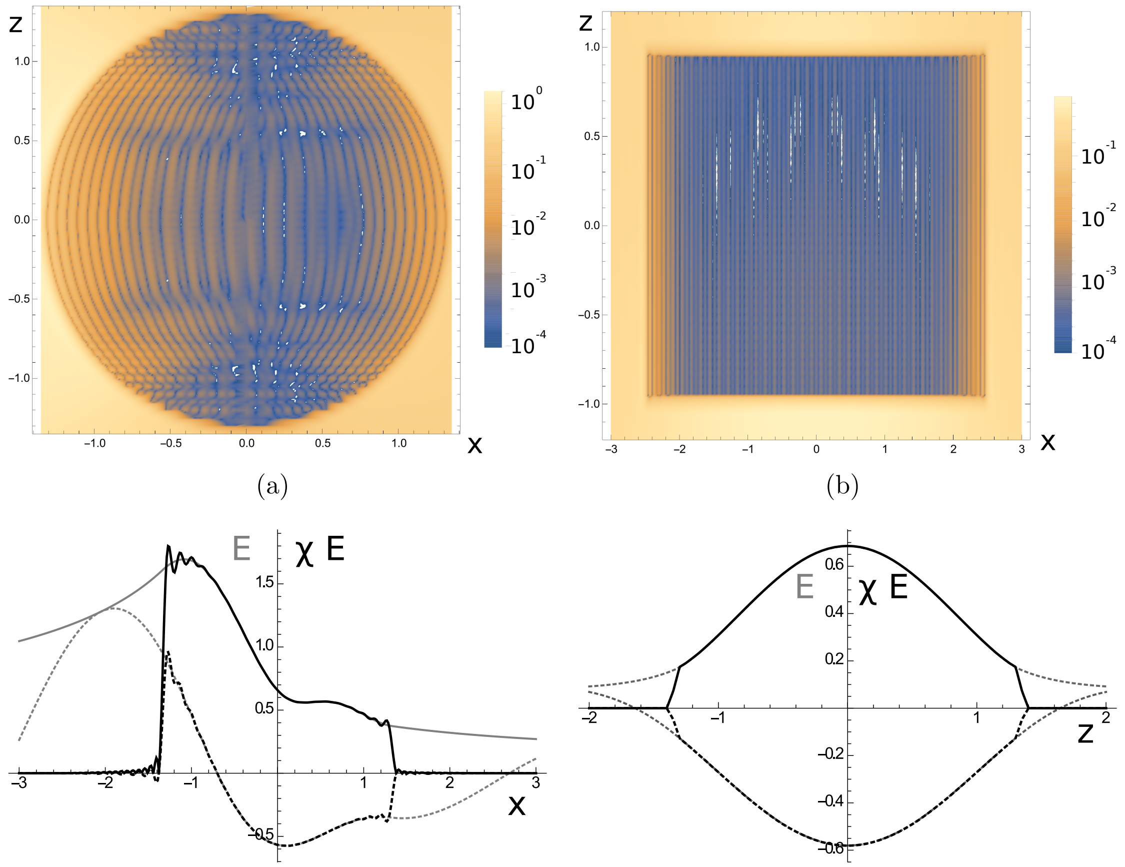
(a)

(b)

Figure 5.6: The real part of $\chi E(x, z)$ (which is proportional to the contrast current), for (a) the circle and (b) the rectangle.

In Fig. 5.6 we show the real part of the simulation results for circle and the rectangle. Figs. 5.7(a),(b) show the absolute value of the difference between the electric field E from the validation results and χE from our simulations. It can be clearly seen that in the two different directions the approximation error has a different character, because we used a different discretization scheme. Especially for the circle it looks like the error is quite large, but this is mainly due to the Gibbs ringing around the discontinuity in the contrast source as can be seen in Fig. 5.7(c). This Gibbs ringing around the discontinuities is not a large problem, since sources with a rapid spatial oscillation do not radiate. In the cut along the line $x = 0$ (Fig. 5.7(d)) the simulation and validation overlap very well.

We have also calculated the far-field scattered intensity against the angle as explained in Section 5.3.4 for the circle. Fig. 5.8(a) shows the scattering intensity versus the angle. Note that the angle of incidence is $-\pi/2$ and that the angle φ is only taken from $-\pi/2$ to $\pi/2$ because of the symmetry. In Fig. 5.8(b) we plotted the error of our simulation results (solid). We also calculated the scattering intensity from the contrast current calculated



(a) (b)

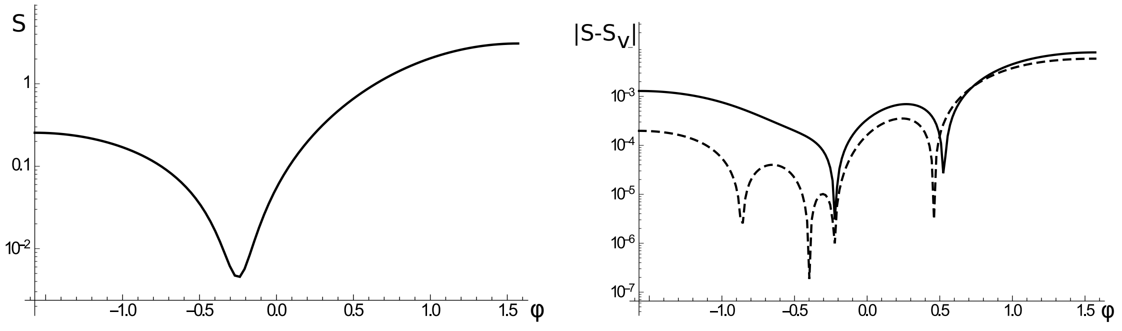
(c) (d)

Figure 5.7: In (a) and (b) we plotted the absolute value of the difference (error) between the electric field from the validation E and χE from our simulation results for the circle resp. the rectangle. At the scatterer these should coincide, because $\chi = 1$ on the scatterer. In (c) we plotted the real (solid) and imaginary (dotted) part of E from validation (grey) and χE (black) from our simulation at $z = -0.2$ for the circle. In (d) we did the same for a vertical cut along $x = 0$ for the rectangle.

using the analytic results, but discretized by our discretization. This gives an idea of the maximum reachable accuracy with the given discretization parameters.

To calculate the scattering intensity we only use the part of the contrast current with a wavenumber smaller than k_0. This means that most of the current in the spectral domain does not radiate, so it does not contribute, since there are 1.18 Gabor coefficients per unit wavevector.

The spatial sampling range in x is larger than the width of the scatterer, with this Gabor frame $W \approx 3.5$ for the circle and the rectangle. This is because the Gabor frame needs extra samples at the sides for a good approximation, because the dual window γ (Eq. (5.18)) is a few times X wider. This can be seen in Fig. 5.9(a), which has been created with a low number of spatial windows, so artifacts are visible at the edges. From Fig. 5.9(b)



(a) (b)

Figure 5.8: (a) The scattering intensity as calculated from the contrast current. (b) The absolute value of the difference between the scattering intensity S and validation data S_V, the solid line is with the contrast current from the simulation, the dashed line is with the contrast current from the exact solution, represented by the same discretization as the simulation.

it is clear that this also yields an increased error for scattering intensity. By simulation error we mean the absolute difference between the calculated scattering strength S and the validation S_V, where the absolute value of S is found in Fig. 5.8(a). By discretization error we mean the difference between the S_V and the scattering radiated by the contrast current from our validation data discretized with the indicated settings.

The Gabor coefficients seem to be quite efficient compared to piecewise-linear sampling in the z direction. When we coarsen the sampling distance in the z direction from $\Delta = 0.05$ to $\Delta = 0.1$ the error in the scattering increases by a factor of three (see Fig. 5.9(c)). While reducing the number of modulation frequencies does not affect the error in the reflection very much, it does add some (non-radiating) high-frequency noise to the contrast current (see Fig. 5.9(d)). It seems that the accuracy is limited by the z-discretization.

We can improve on the interpolation by oversampling in the x direction. We use 16 times oversampling and making that number smaller increases the error considerably (see Fig. 5.9(e)). Fig. 5.9(f) shows the approximation error from approximating an electric field (with magnitude 0.01) with 5 vs. 9 X_ℓ interpolation points. The location of these X_ℓ values is clearly visible from the dips in the error.

As an example of larger structures that our code is capable of simulating the scattering from, we have included the finite grating structure. From the current distribution in Fig. 5.10(a) it can be seen that there is a significant mutual coupling between the rectangles in this example. This object is larger than the previous two simulated objects, measuring two by one half wavelength and it shows that our algorithm is useful for larger objects as well. In Fig. 5.10(b) we can not only identify the first three diffraction orders, we also see scattering at different angles due to the mutual coupling and the finite size of the grating.

To show how our code scales to finer discretizations we increased the number of modulation frequencies, the range of n in (5.15), several times and measured the time taken for applying the operator A (Eq. (5.29)) once, as shown in Fig. 5.11. The results suggest that the scaling is even better than $O(N \log N)$, because $O(N)$ operations still dominate for this small simulation size. Therefore the $\log N$ factor will only show up for even higher numbers of Gabor coefficients, where the FFTs will in the end dominate the speed of this operator.

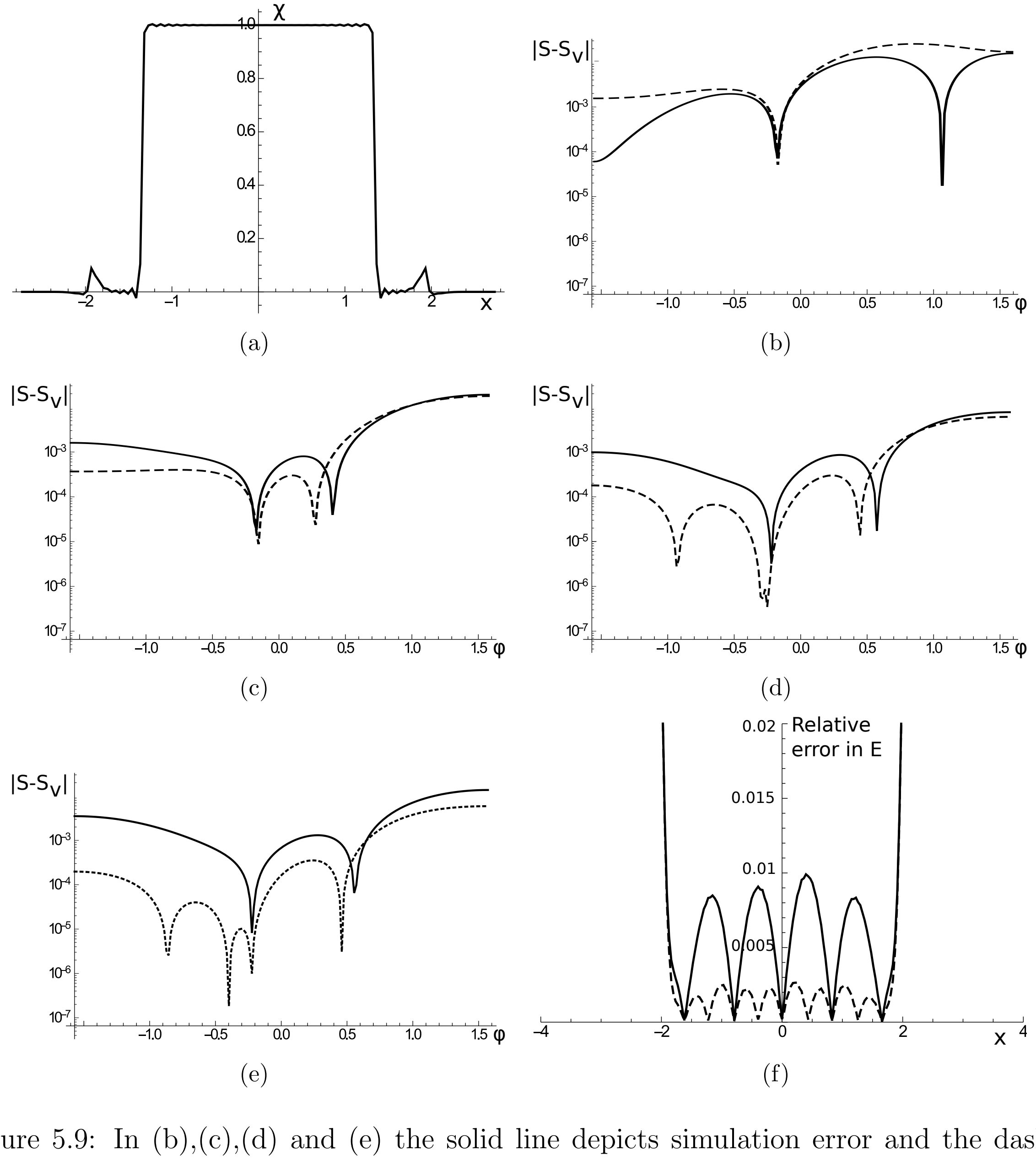
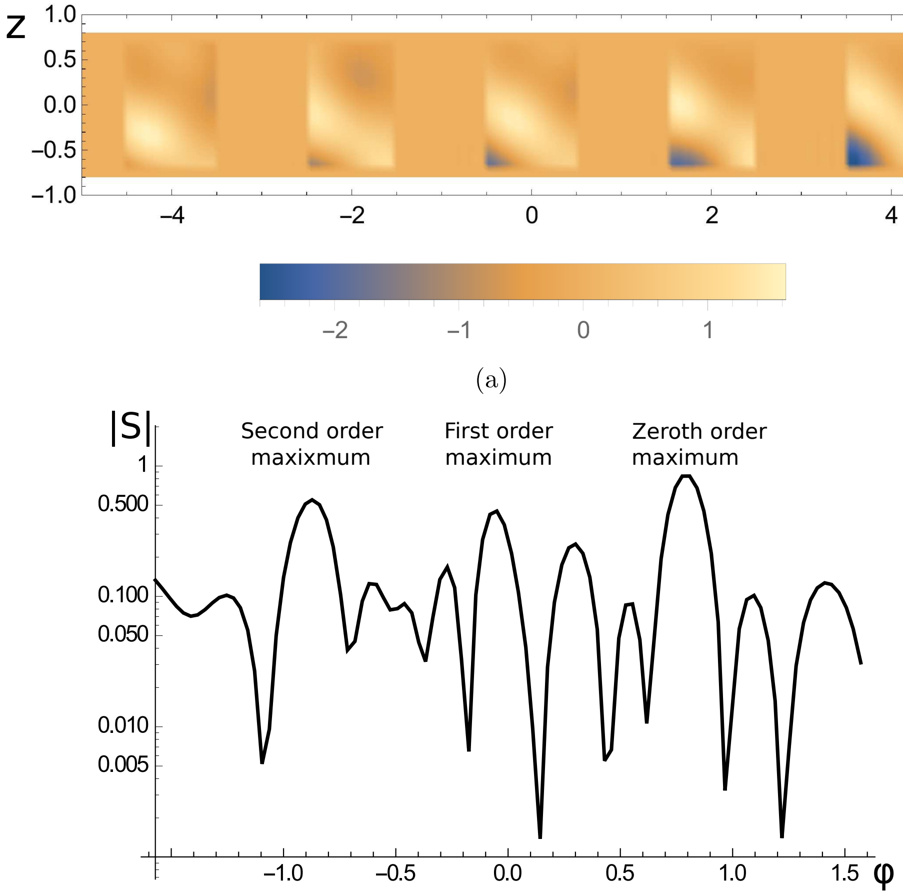


Figure 5.9: In (b),(c),(d) and (e) the solid line depicts simulation error and the dashed line discretization error. (a) The contrast function discretized with 9 windows. (b) The resulting scattering strength error. (c) The scattering strength error with $\delta = 0.1$, (d) The scattering strength error with 5 modulation frequencies. (e) The scattering strength error with 8 spatial correction points X_ℓ. (f) Plot of relative error of spectral-spatial transformation, solid: $\ell \in \{-2, \ldots 2\}$, dashed: $\ell \in \{-4, \ldots, 4\}$.



(a)

(b)

Figure 5.10: (a) The real part of the contrast current density in the grating. (b) The far field scattering intensity from the grating.

5.7 Conclusion

We have constructed an algorithm that solves a 2D domain integral equation for TE polarization in a homogeneous background discretized by a Gabor frame in one direction. The number of numerical operations of this method scales as $O(N_z N_x \log N_x)$, with N_x the number of samples in the x direction and N_z the number of samples in the z direction. The best efficiency is reached in the direction where the Gabor frame was employed. Because of the lower accuracy of the piecewise-linear discretization in the z direction, more samples per meter are needed in that direction. This clearly shows the value of using the Gabor frame as a discretization.

For every spatial discretization using Gabor frames, there exists a spectral Gabor frame where the individual coefficients are directly mapped onto each other. When we have a spatial discretization of a function with Gabor coefficients, we automatically have a discretiza-

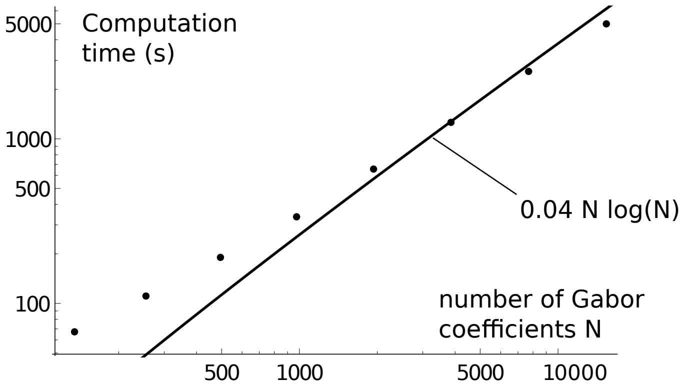


Figure 5.11: Dots indicate computation time of one matrix vector product of operator A versus the number of Gabor coefficients for an increasing number of modulation frequencies in the Gabor frame. Solid line indicates a reference of $O(N \log N)$ behavior.

tion of its Fourier transform. This is very convenient, because the Fourier transformation is exact and fast.

With other choices of expansion functions, the Fourier transformation is often a difficult step. For example a piecewise-linear discretization creates high-frequency spectral components artificially and the Fourier transformation is only an approximation. With the use of a Gabor frame, the most important approximation that we make is the discretization of the problem, the rest can be done (almost) exactly. The downside is that the Gabor frame is not very suitable for discretizing discontinuous(ly differentiable) functions, but, although the spectral Green function is not continuously differentiable, we can work around that by using a coordinate stretch. The method of coordinate stretching is an effective method to cope with the discontinuity of the derivative in $e^{\gamma|z|}$ and the branch points in $1/\gamma$ simultaneously.

For a discretization with a finer resolution in the x direction, this method scales very well as $O(N_x N_z \log N_x)$. However, a relatively large minimum number of Gabor frames is always needed for small simulation regions.

Chapter 6

A domain integral equation approach for simulating two dimensional transverse electric scattering in a layered medium with a Gabor frame discretization[1]

6.1 Abstract

We solve the 2D transverse-electrically polarized domain-integral equation in a layered background medium by applying a Gabor frame as a projection method. This algorithm employs both a spatial and a spectral discretization of the electric field and the contrast current in the direction of the layer extent. In the spectral domain we use a representation in the complex plane that avoids the poles and branch cuts found in the Green function. Because of the special choice of the complex-plane path in the spectral domain and because of the choice to use a Gabor frame to represent functions on this path, fast algorithms based on FFTs are available to transform to and from the spectral domain, yielding an $O(N \log N)$ scaling in computation time.

6.2 Introduction

For several applications in electrical engineering it is vital to have fast and accurate models to calculate the scattering of electromagnetic waves from dielectric structures of finite size. Among these are metrology for integrated circuit production [136, 137], various elements on nanophotonic chips [24, 25] and metamaterials [138]. For these applications, structures are often embedded in a host medium with multiple layers of different materials.

[1] This chapter was first published as the article [135].

Many numerical methods have already been developed for the characterization of electromagnetic scattering in a multi-layered medium, e.g. local formulations such as finite difference time domain (FDTD) [26] and the Finite Element Method (FEM) [32, 134]. Global formulations that employ a Green function exist in both a time-domain formulation and a time-harmonic formulation. In the time domain the Green function can be generalized to multilayered media as well [66, 67, 68]. Since we are interested in scattering from monochromatic light sources, we are interested in a time-harmonic formulation. Such an integral formulation requires solving a nonhomogeneous matrix equation, which can be solved efficiently with an iterative solver, especially when the matrix-vector product can be computed rapidly. One popular approach is to speed up this matrix vector product by decomposing the Green function into long-range and short-range interactions, combined with a hierarchical division of the simulation domain. Examples of such methods for a homogeneous medium are the Fast Multipole Method (FMM) [45] and the Fast Inhomogeneous Plane Wave Algorithm (FIPWA) [50, 51]. Extensions to multilayer media exists both for FMM [139] and for FIPWA [52]. Another popular approach for fast matrix vector products exploits the observation that the Green function in a layered medium exhibits a translation symmetry in the direction parallel to the layer interfaces. We will focus on a method that exploits this translation symmetry.

A domain integral formulation consists of two parts. The first part is the calculation of the contrast current density from the electric field and the contrast function. The second part is an integral over the product of the Green function and the contrast current density that yields the scattered electric field. For homogeneous media, the spatial Green function is readily obtained. Among the free-space methods that employ the Green function in the spatial domain and exploit the translation symmetry are the Conjugate Gradient Fast Fourier Transform CGFFT [37] and its enhancements, such as the Adaptive Integral Method (AIM) [42] and the pre-corrected FFT [43].

To use similar methods in stratified media, the multi-layered Green function is required. The main differences between the two-dimensional free-space Green function and the two-dimensional stratified-medium Green function are the reflections and transmissions at the layer interfaces in layered media. For stratified media, an exact expression for the Green function in the spectral domain can be derived, so in principle it is possible to calculate the Green function completely in the spatial domain via a Fourier transform. However, calculating the pertaining Fourier integral is far from trivial, since there are branch cuts and poles present in the Green function. Several methods exist to calculate these so-called Sommerfeld integrals e.g. the Discrete Complex Image Method (DCIM) [65], the steepest descent path (SDP) [59], Sommerfeld tail extrapolation [140], and a method based on a perfectly matched layer [141, 142].

An alternative approach is to consider the so-called spectral methods, in which the exact spectral-domain Green function is employed directly. For periodically repeating scattering structures, such as optical gratings, several spectral methods have already been developed. Important examples are the Rigorous Coupled Wave Analysis (RCWA), also called the Fourier Modal Method [74, 73], and some periodic-volume-integral-equation-based methods (PVIE) [92]. Then, the periodicity can be exploited because the periodicity implies a discrete spectral domain and therefore an obvious and well-performing discretization of the spectral domain exists. These methods can be adapted to solve a-periodic structures as well, for example via perfectly matched layers (PML) [124] or supercell techniques, but even then these solvers are in essence periodic.

Here we present a mixed spatial-spectral method that is completely a-periodic in nature. This also implies that the spectral domain is now continuous instead of discrete. Consequently, branch cuts and poles in the spectral Green function need to be treated carefully. We demonstrate an approach that employs a representation of the fields in the spectral-domain complex plane, that avoids the poles found in the Green function. This representation has been specifically chosen such that the Green functions consist of smooth functions with an effectively limited support, while still allowing for efficient transformations to and from the spectral domain with $O(N \log N)$ computational complexity, for N degrees of freedom. The finite support in both the spectral and the spatial domain allows for a convenient discretization in terms of Gabor frames. Owing to the Gabor frames, all operations are of $O(N \log N)$ complexity or less, thus yielding an $O(N \log N)$ scaling with the number of unknowns.

We start this paper with details about the formulation, after which we present the discretization scheme. Subsequently, the spectral complex-plane path and the representation of the Green functions on this path are illustrated. We conclude with three numerical examples to demonstrate the proposed scheme.

6.3 Formulation

6.3.1 Problem definition

Consider a layered medium, i.e. a structure of $N-1$ horizontal layers stacked in the z-direction, with relative permitivities $\varepsilon_{r,n}$ and thicknesses d_n. The space below the stack has permittivity $\varepsilon_{r,N}$ and above the stack has permittivity $\varepsilon_{r,0}$. In layer i a two-dimensional dielectric object in the $x-z$ plane is described by the permittivity function $\varepsilon_r(x,z)$. The object is contained within the rectangle $x \in [-W, W]$, $z \in [z_{min}, z_{max}]$, which we call the simulation domain. Figure 6.1 (a) shows our scattering setup for $N=3$ and $i=1$.

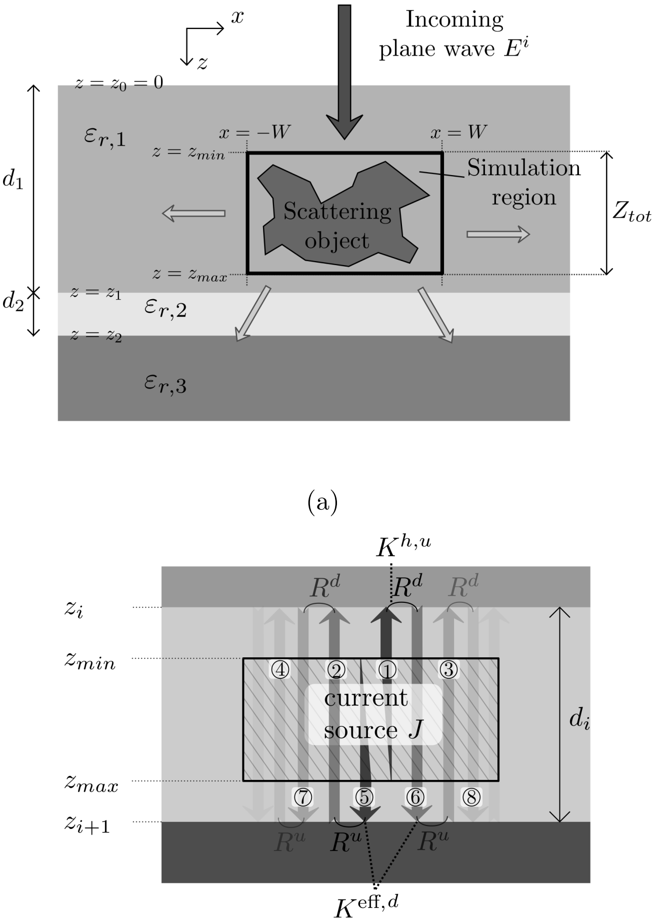
Due to the two-dimensional configuration and the polarization of the incoming field, only the y componenent of the electric field is nonzero, which turns our problem into a scalar problem $\mathbf{E}(x,z) = \hat{\mathbf{y}}E(x,z)$. Putting this into the Maxwell equations [85] with time convention $\exp(j\omega t)$ yields a second order differential equation.

We make a distinction between the incident electric field E^i, which solves the problem in abscence of the scattering object, and the field scattered by the object E^s. The combination of these fields yields the total electric field $E = E^i + E^s$ in the simulation domain.

The contrast function in layer i is defined by

$$\chi(x,z) = \frac{\varepsilon_r(x,z)}{\varepsilon_{r,i}} - 1 \tag{6.1}$$

The contrast current density can be obtained through $J(x,z) = j\omega\varepsilon_0\varepsilon_{r,i}\chi(x,z)E(x,z)$. Similarly $J^i(x,z) = j\omega\varepsilon_0\varepsilon_{r,i}\chi(x,z)E^i(x,z)$ and $J^s(x,z) = j\omega\varepsilon_0\varepsilon_{r,i}\chi(x,z)E^s(x,z)$ are obtained from the incident and scattered part of the electric field.



(a)

(b)

Figure 6.1: (a) The scattering setup. (b) The source of reflections including the definition of the different K waves.

6.3.2 The homogeneous medium Green function

With the Green function the field radiated by a contrast current can be calculated. The Green function will be represented in the spectral domain in the x-direction. The spectral domain is defined through the Fourier transformation

$$\varphi(k_x) = \mathcal{F}_x[\varphi(x)](k_x) = \int_{-\infty}^{\infty} dx\, \varphi(x)e^{-jk_x x}. \tag{6.2}$$

and its inverse $\mathcal{F}^{-1}$. We use the variable k_x for the spectral domain. Whenever a function has x as its argument the spatial version of the function is meant, when k_x is used as an argument its Fourier transform is meant.

We will start working on the problem within one layer, i.e. we assume a homogeneous dielectric constant in the background layer. By adding reflection and transmission coefficients we will turn this into the solution for a multi layered problem in Section 6.3.3. In a homogeneous medium the Green function, $\mathcal{G}^h$, for a contrast current density J can be written as

$$G^h(k_x, z|z') = \frac{e^{-\gamma|z'-z|}}{2\gamma}, \tag{6.3}$$

where $\gamma = \sqrt{k_x^2 - k_b^2}$ and $k_b = \omega\sqrt{\varepsilon_0\varepsilon_{r,i}\mu_0}$ with standard branch cut in the square root. Again i indicates the layer in which the simulation domain is located, and the h subscript will indicate the homogeneous part in this context.

We can calculate the scattered electric field E^h from current source J (not yet taking reflections into account) by

$$\begin{aligned} E^h(k_x, z) = &\int_{z_{min}}^{z} dz' \, \frac{e^{-\gamma(z-z')}}{2j\omega\varepsilon_0\varepsilon_{r,i}\gamma} J(k_x, z') + \\ &\int_{z}^{z_{max}} dz' \, \frac{e^{-\gamma(z'-z)}}{2j\omega\varepsilon_0\varepsilon_{r,i}\gamma} J(k_x, z') \\ = &\, K^{h,d}(k_x, z) + K^{h,u}(k_x, z). \end{aligned} \tag{6.4}$$

Here we recognize that the factor $1/2j\omega\varepsilon_0\varepsilon_{r,i}\gamma$ factor calculates the spectral electric field from the spectral current density at the same altitude and that the factor $e^{-\gamma z}$ propagates this over an altitude displacement z in the z-direction. We will call $e^{-\gamma z}$ the propagation function in medium i. The first term in Eq. (6.4) represents a wave moving down from the source, which we will denote by $K^{h,d}$, and the second term a wave moving up denoted by $K^{h,u}$.

6.3.3 Reflections from layer interfaces

In a multi-layered medium we have to include reflections from the layer interfaces to arrive at the complete scattered field E^s from a current source. With the use of [58, Chapter 2] and [87, Chapter 5], we can calculate the reflection coefficient from the stack of layers above layer i, $R^d(k_x)$ and from the stack below layer i, $R^u(k_x)$[2]

From Figure 6.1(b), we deduce how the scattered field E^s can be constructed from the homogeneous field E^h together with a sum of all reflections. For the upward directed part of E^s, i.e $K^{s,u}$, two contributions can be identified. First, there is the homogeneous part, indicated by ① in Fig. 6.1(b) and denoted by $K^{h,u}(k_x, z)$. Second, there is a set of reflections that propagates upward from the layer interface below at z_{i+1} ②-④. Now we write the sum of all upward-directed reflections ②-④ as a single effective reflection

[2] This notations is different from the notation in Chapter 2, where they were called $\mathcal{T}_{ii}$.

coefficient $R^{\text{eff},u}$, multiplied by a single effective downward directed wave $K^{\text{eff},d}$ ⑤,⑥. The sum of all reflections can then be propagated through the medium using the propagation function $e^{-\gamma(z_{i+1}-z)}$, yielding

$$K^{s,u}(k_x, z) = K^{h,u}(k_x, z) + e^{-\gamma(z_{i+1}-z)} R^{\text{eff},u}(k_x) K^{\text{eff},d}(k_x, z_{i+1}). \tag{6.5}$$

Now the effective downward directed wave, $K^{\text{eff},d}$ is defined as the sum of the homogeneous downward directed wave ⑤, and the reflection of the upward-directed wave ⑥, i.e.

$$K^{\text{eff},d}(k_x, z_i) = K^{h,d}(k_x, z_{i+1}) + R^d(k_x) e^{-\gamma d_i} K^{h,u}(k_x, z_i). \tag{6.6}$$

The reflection generated by the effective downward-directed $K^{\text{eff},d}$ reflects upwards with reflection coefficient R^u. Thereafter it can bounce back and forth several times between the layer interfaces at z_i and z_{i+1}, where it propagates a distance $2d_i$ between each bounce and is reflected with both R^u and R^d. The first bounce is indicated by ⑦ and ⑧. This behavior is summarized in a single effective reflection coefficient $R^{\text{eff},u}$ as

$$R^{\text{eff},u} = R^u(k_x) \sum_{n=0}^{\infty} \left(R^u(k_x) R^d(k_x) e^{-2\gamma d_i} \right)^n .$$

This sum can be calculated using the geometric series [55, Section 2.1]

$$R^{\text{eff},u}(k_x) = \frac{R^u(k_x)}{1 - R^u(k_x) R^d(k_x) e^{-2\gamma d_i}}. \tag{6.7}$$

Similarly, the down-directed wave is defined through

$$\begin{aligned} K^{s,d}(k_x, z) =& K^{h,d}(k_x, z) + e^{-\gamma(z-z_i)} R^{\text{eff},d}(k_x) K^{\text{eff},u}(k_x, z_i) \\ R^{\text{eff},d}(k_x) =& \frac{R^d(k_x)}{1 - R^d(k_x) R^u(k_x) e^{-2\gamma d_i}} \\ K^{\text{eff},u}(k_x, z_i) =& K^{h,u}(k_x, z_i) + R^u(k_x) e^{-\gamma d_i} K^{h,d}(k_x, z_{i+1}). \end{aligned} \tag{6.8}$$

The sum of all upward and downward directed waves equals the scattered field $E^s(k_x, z) = K^{s,u}(k_x, z) + K^{s,d}(k_x, z)$, which includes the homogeneous contribution of the Green function and reflections from the layer interfaces. We are now able to calculate the field including reflections as a function of a current source. When we call $\mathcal{G}[J](k_x, z)$ the integral operator that calculates the scattered field E^s generated by current source J, we can write down the integral equation as

$$\begin{aligned} J^s(x, z) =& j\omega\varepsilon_0\varepsilon_{r,i}\chi(x, z) E^s(x, z) \\ =& j\omega\varepsilon_0\varepsilon_{r,i}\chi(x, z) \mathcal{F}^{-1}_{k_x} \left\{ \mathcal{G}[J^i + J^s](k_x, z) \right\} (x, z), \end{aligned} \tag{6.9}$$

where J^s is the unknown part of the contrast current density. Again, J^i is known and is obtained through $J^i(x, z) = j\omega\varepsilon_0\varepsilon_{r,i}\chi(x, z) E^i(x, z)$. The equation can be ordered with the unknown J^s on the right-hand side as

$$\begin{aligned} & j\omega\varepsilon_0\varepsilon_{r,i}\chi(x, z) \mathcal{F}^{-1}_{k_x} \left\{ \mathcal{G}[J^i](k_x, z) \right\} (x, z) = \\ & \qquad - J^s + j\omega\varepsilon_0\varepsilon_{r,i}\chi(x, z) \mathcal{F}^{-1}_{k_x} \left\{ \mathcal{G}[J^s](k_x, z) \right\} (x, z). \end{aligned} \tag{6.10}$$

Note that this formulation is somewhat different from that in other papers, e.g. [143].

6.4 Discretization

6.4.1 The z direction: Piecewise-linear functions

We use piecewise-linear (PWL) expansion functions as discretization in the z direction. The expansion functions are

$$\Lambda_n(z) = \begin{cases} 1 - \frac{|z-n\Delta-z_{min}|}{\Delta} & \text{if } |z-n\Delta-z_{min}| < \Delta \\ 0 & \text{if } |z-n\Delta-z_{min}| > \Delta \end{cases}, \tag{6.11}$$

with Δ the step size in the z discretization. For testing we use Dirac-delta functions, since these lead to a well-conditioned problem [71]. We use a subscript n to indicate the basis function to which a field corresponds e.g. $E_n(k_x) = E(k_x, z_{min} + n\Delta)$. The maximum value for n is $N_z = (z_{max} - z_{min})/\Delta$. Now we can approximate for example the electric field using these expansion function by

$$E(k_x, z) \approx \sum_{n=0}^{N_z} E_n(k_x)\Lambda_n(z). \tag{6.12}$$

With the use of this discretization in the z direction we can write the integral with G^h in Eq. (6.4) in more detail as

$$K_n^{h,u}(k_x) = \int_{z_{min}}^{z_{min}+n\Delta} dz' \sum_{n'=0}^{n+1} G^h(k_x, n\Delta|z')k_b^2 J_{n'}(k_x)\Lambda_{n'}(z').$$

We use a recursive algorithm [72] to find the result of the integral for each n. We find

$$\begin{aligned} K_{n+1}^{h,u}(k_x) =& K_n^{h,u}(k_x)e^{-\gamma\Delta} \\ & - J_n(k_x)\int_0^\Delta dz' k_b^2 \Lambda_0(z')\frac{e^{-\gamma(\Delta-z')}}{2\gamma} \\ & - J_{n+1}(k_x)\int_0^\Delta dz' k_b^2 \Lambda_1(z')\frac{e^{-\gamma(\Delta-z')}}{2\gamma} \\ =& K_n^{h,u}(k_x)e^{-\gamma\Delta} + J_n(k_x)h_m^u(k_x) + J_{n+1}(k_x)h_e^u(k_x), \end{aligned} \tag{6.13}$$

where we introduced $h_m^u(k_x)$ and $h_e^u(k_x)$ for the result of the integrals over z', here the m and e subscript stand for integrals to the middle or to the end of basis function $\Lambda_n(z)$. In the other direction, $K_{n+1}^{h,d}(k_x)$ is defined as

$$\begin{aligned} K_{n-1}^{h,d}(k_x) =& K_n^{h,d}(k_x)e^{-\gamma\Delta} \\ & - J_n(k_x)\int_{-\Delta}^0 dz' k_b^2 \Lambda_0(z')\frac{e^{-\gamma(z'-\Delta)}}{2\gamma} \\ & - J_{n+1}(k_x)\int_{-\Delta}^0 dz' k_b^2 \Lambda_{-1}(z')\frac{e^{-\gamma(z'-\Delta)}}{2\gamma} \\ =& K_n^{h,d}(k_x)e^{-\gamma\Delta} + J_n(k_x)h_m^d(k_x) + J_{n-1}(k_x)h_e^d(k_x), \end{aligned} \tag{6.14}$$

The discretized scattered field $E_n^s(k_x)$ can be calculated by adding the terms that represent the reflections in Eq. (6.5) to the homogeneous waves $K_n^{h,u/d}(k_x)$.

6.4.2 The x direction: Gabor frames

For the discretization in the x direction we use a Gabor frame in the spatial as well as the spectral domain. The Gabor frame is defined in the exposition [99, Chapter 8] and Chapter 4. We employ the Gaussian window function

$$g(x) = 2^{\frac{1}{4}} e^{\left(-\pi \frac{x^2}{X^2}\right)}, \tag{6.15}$$

where X defines the width of the window function. The Gabor frame is defined as

$$g_{mn}(x) = g(x - m\alpha X) e^{jn\beta Kx}, \tag{6.16}$$

where $K = 2\pi/X$ the spectral window distance and α and β define the oversampling. We use rational oversampling where $\alpha\beta = p/q$ with $p < q$ and $p, q \in \mathbb{N}$. In principle, the Gabor coefficients can be calculated from the dual frame $\eta_{mn}(x)$ by

$$f_{mn} = \int_{-\infty}^{\infty} dx\, f(x) \eta_{mn}(x), \tag{6.17}$$

where the dual frame is found from

$$\eta_{mn}(x) = \eta(x - m\alpha X) e^{jn\beta Kx}. \tag{6.18}$$

When there is oversampling there is a freedom of choice for the dual window η. The dual window η we use to calculate Gabor coefficients is the one obtained by using the Moore-Penrose inverse [99, 100].

The spectral Gabor frame is simply the Fourier transform of the Gabor frame in Eq. (6.16). More details can be found in [118].

6.5 Complex-plane spectral path

6.5.1 Poles, branch cuts and rapid oscillations

From Eq. (6.9) it is clear that Fourier transformations are an essential part of the presented algorithm, since the contrast multiplication should be executed in the spatial domain and the Green function operator can be handled in the spectral domain. It is however difficult to represent the Green function $\mathcal{G}$, for which an exact definition only exists in the spectral domain, because $\mathcal{G}$ contains poles, branch cuts, and rapid oscillations in the spectral domain. It is therefore not trivial to represent the Green function efficiently in terms of Gabor frames. Calculation of the Green operator in the spatial domain, through so-called Sommerfeld integrals, is possible but tedious. We would like to find an efficient representation of J, E and $\mathcal{G}$ in the spectral domain to compute $\mathcal{G}$ working on J. We have several requirements for this representation.

1. We aim for a straightforward representation in which we can contain the complete integral operator $\mathcal{G}$. We do not want to take the poles and branch cuts into account separately.

2. A fast transformation to the spectral-domain representation should be available for J and a fast transformation back to the spatial domain for E.

3. Since a spectral multiplication corresponds to a spatial convolution, we need the representation to hold over the entire spatial domain. Otherwise, subsequent convolutions will induce errors.

4. We need the fields only in the simulation domain. A representation that holds over the entire spatial domain will be inefficient, as it carries more information than we need.

There is clearly some tension between points 3 and 4. A method to solve these issues is to represent functions not on the real spectral axis, but in the spectral complex plane.

The first challenge is that the Green operator $\mathcal{G}$ contains branch cuts in the effective reflection coefficients starting at $k_x^2 = \omega^2\mu_0\varepsilon_0$ and $k_x^2 = \omega^2\mu_0\varepsilon_N$ towards $k_x = \pm j\infty$ as indicated in Fig. 6.2.(a).

Additionally, the effective transmission and reflection coefficients (6.7) can also have poles corresponding to guided waves. In [55, Chapter 2.7] and [60] bounds are given within which the poles are located. For a lossless multi-layered medium it can be shown that the poles must have a distance of at least $k_0 = \omega\sqrt{\max\{\varepsilon_0, \varepsilon_{rb,N}\}\mu_0}$ from $k_x = 0$ and that they lie on the real k_x axis. It was shown that for lossy media the poles also have a minimum distance away from $k_x = 0$ and can be found in solely in the northwest and southeast quadrants of the complex plane.

The last difficulty in the Green function are rapid oscillations. The propagation function $\exp(-\gamma z)$ oscillates rapidly for $-Re(k_b) < k_x < Re(k_b)$ at large z, since $\gamma = \sqrt{k_x^2 - k_b^2}$ has a dominant imaginary part on this range. It is hard to capture these oscillations with a small number of Gabor coefficients.

6.5.2 Functions in the complex plane

In [55, Chapter 2.7] several methods are mentioned for integration paths that avoid the poles and branch cuts to calculate Sommerfeld integrals efficiently. Since the discontinuities in the reflection coefficients are located in the northwest and southeast quadrant of the complex plane, we would like to discretize functions on a path through the southwest and northeast quadrant. We want to represent all functions in the spectral domain on such an integration path to bypass the poles and branch cuts.

The key observation is the complex shift

$$\mathcal{F}[f](k_x \pm jA) = \int_{-\infty}^{\infty} f(x)e^{-j(k_x \pm jA)x}dx = \mathcal{F}[f(x)e^{\mp Ax}](k_x). \tag{6.19}$$

We see that it is possible to generate a shift of a coordinate in the complex spectral domain of $\pm jA$, by multiplying the spatial representation by $e^{\mp Ax}$ and then carrying out a standard real-axis Fourier transform. As mentioned before, the Gabor frame allows for fast and efficient Fourier transformations and multiplications of functions.

To arrive at an integration path passing through the southeast and northwest quadrants we have to split the spectral domain integration path into pieces that we will handle separately. The path we choose can be written as the union of three line segments, parametrized by the real-valued parameter τ, see Figure 6.2.(a), i.e.

$$k_x(\tau) \in \begin{cases} \tau - jA & \text{if } \tau < -A \\ (1+j)\tau & \text{if } -A \leq \tau < A \\ \tau + jA & \text{if } \tau > A. \end{cases} \tag{6.20}$$

This path is also used in [61] and the path also bears some resemblance to the steepest-decent path used in [59].

In principle, we should also integrate over the line segments $[-\infty, -\infty - jA]$ and $[\infty, \infty + jA]$ to close the contour, but asymptotically all functions of interest are zero on these intervals, so we can safely ignore them owing to Jordan's lemma. We chose A small compared to the complete spectral range that we discretize, so there is only a little bit of information contained in the middle part of this representation. Although there is some freedom for the choice of A within which the overall algorithm performs well, we found that the choice $AW \approx 3$, with W the width of the simulation domain, will yield satisfactory results in general. An optimal choice for A will somewhat vary, depending on the shape of the simulation domain and the required accuracy.

We can identify three types of functions that are part of the Green function that we need to represent on this complex path. These three types are the propagation function $\exp(-\gamma z)$, the current to field functions h_m^u, h_m^d, h_e^u and h_e^d (Eq. (6.13)), and the reflection coefficients $R^{\text{eff},u}$ and $R^{\text{eff},d}$ (Eq. (6.7)). As can be seen in Fig. 6.2(b), the propagation function is much better behaved on the complex path and the same holds for $h_{m/e}^{u/d}(k_x)$, since these functions consist of a z-integral over the propagation function. Fig. 6.2(c) shows the effect of going around the poles on the complex path: the poles are smoothened as well. The obstacles mentioned in Section 6.5.1 have disappeared through the use of this complex spectral path.

6.5.3 Transforming to and from the complex spectral path

For the spectral domain representations on the outer parts in Eq. (6.20), i.e. $|\tau| > A$ we use Eq. (6.19). We will adopt the notation $f_L(\tau) = f(\tau - jA)$ and $f_R(\tau) = f(\tau + jA)$ for the representation on the left and on the right part of the complex spectral path, respectively.

To calculate the spatial representation of a function from a spectral representation we employ Eq. (6.19) the other way around. However, here we also have to enforce the end of

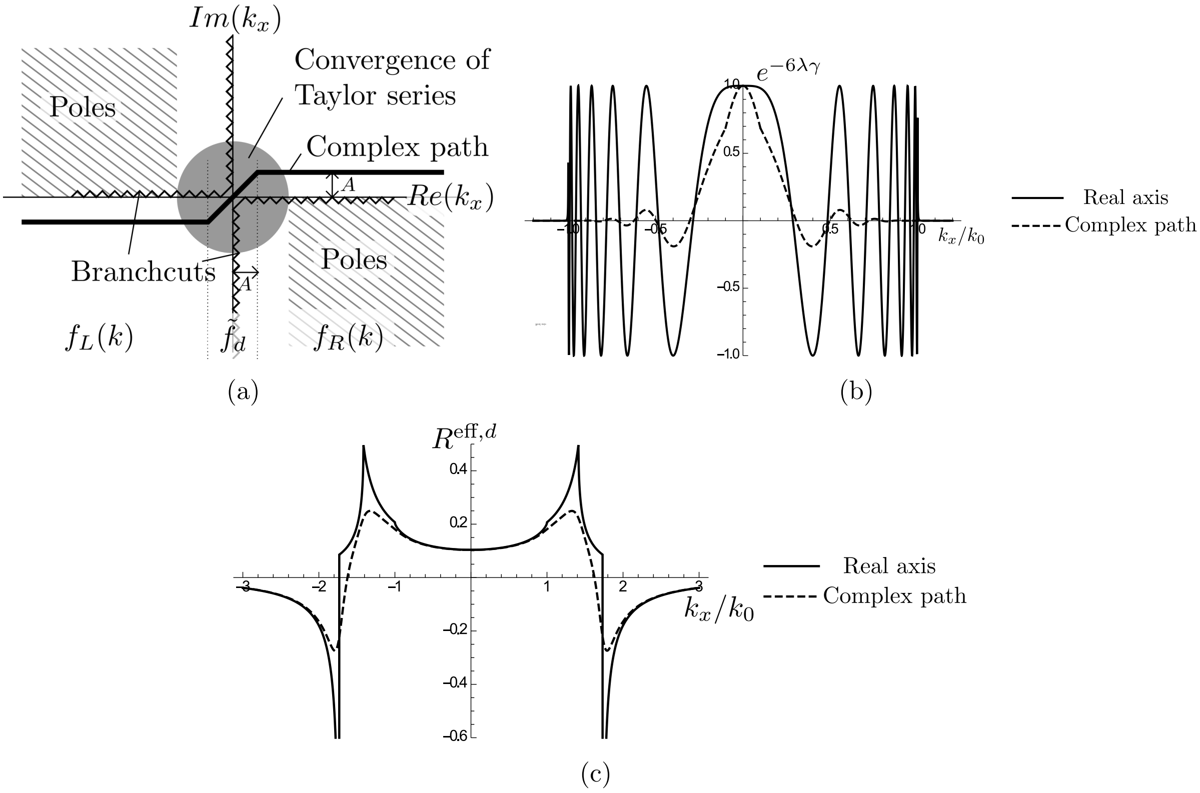


Figure 6.2: (a) The spectral path in the complex plane. (b) Better behavior of the propagation function. (c) Better behavior for the effective reflection coefficient.

the domain at $\tau = \pm A$ by means of cutoff functions c defined by

$$\begin{aligned} c_L(\tau) =& U(A-\tau) \\ c_R(\tau) =& U(\tau - A), \end{aligned} \tag{6.21}$$

with U the Heaviside step function. We can now calculate the contribution of $f_{L/R}$ to the spatial domain by

$$\begin{aligned} f(x) = e^{Ax}\mathcal{F}_\tau^{-1}[c_L(\tau) f_L(\tau)](x) + \\ e^{-Ax}\mathcal{F}_\tau^{-1}[c_R(\tau) f_R(k)](x) + \int_{-A-jA}^{A+jA} dk_x\, f(k_x) e^{jk_x x}, \end{aligned} \tag{6.22}$$

where the first two terms can be readily computed using Gabor-frame based operations and where the integral in the last term is still there since we did not yet describe the discretization of the middle part of the integral.

An important remark on the middle part of the complex integration path is that it does not contain much information. Optimization for speed is not crucial on the middle part as long as the the time spent to calculate its contribution is negligible compared to the time needed for the left and right parts. The applied method only needs to be 'accurate enough'.

When we assume that there are no poles in a range of $\sqrt{2}A$ around 0 (otherwise we can choose a smaller value for A), we can employ a Taylor series to approximate functions in this part of the spectral domain. The Taylor series has the advantage that we can use derivatives around $k_x = 0$ to make a continuation into the complex plane.

To calculate the derivatives of a function $f(k_x)$ represented by Gabor coefficients f_{mn} we define

$$\tilde{f}_d = f^{(d)}(0) = \sum_{m,n} f_{mn} g_{mn}^{(d)}(0), \tag{6.23}$$

where index d runs over the derivatives. Here the derivatives from the spectral Gabor frame can be easily calculated from Eq. (6.16). To calculate the spatial contribution due to the derivatives $\tilde{f}_d$, the Taylor series is used to approximate $f(k_x)$ around $k_x = 0$. From

$$f(k_x) = \sum_d \frac{k_x^d \tilde{f}_d}{d!} \tag{6.24}$$

we can write the integral in Eq. (6.22) as

$$\int_{-A-jA}^{A+jA} dk_x \, f(k_x) e^{jk_x x} = \sum_d \frac{\tilde{f}_d}{d!} i_d(x) \tag{6.25}$$

where

$$i_d(x) = \int_{-A-jA}^{A+jA} dk_x \, (k_x)^d e^{jk_x x}. \tag{6.26}$$

The Gabor coefficients of the $i_d(x)$ functions can be computed during the initialization phase of our algorithm. Additionally, the Taylor series of the product of two functions is needed when multiplying quantities that are represented on the spectral path, such as in Eqs (6.4)-(6.8). This product can be obtained through the general Leibnitz rule:

$$\tilde{(fh)}_d(0) = \sum_m \frac{d!}{(d-m)!m!} \tilde{f}_{d-m} \tilde{h}_d. \tag{6.27}$$

We typically need around 10 terms in the Taylor series for a simulation region of one wavelength in the z direction and three digits precision.

6.6 Approximation of functions

The functions that make up the Green operator $\mathcal{G}$, i.e. $h_{m/e}$, R^{eff} and $e^{-\gamma\Delta}$, need to be approximated accurately on the complex-plane spectral path. Additionally, in the spatial domain χ has to be approximated using Gabor coefficients. We will now give some details on how we obtain approximations for these functions.

6.6.1 General remarks

We use the same methods as Bastiaans employs in [100] to calculate the Gabor coefficients for a function efficiently. To increase the accuracy we use some oversampling, i.e. we calculate Gabor coefficients for a larger range of index n in Eq. (6.17), and then discard the coefficients with large n, which we do not need. This leads to a finer sampling of the function and therefore to a higher accuracy. Typically oversampling by a factor of four yields three digits precision.

The functions $f_L(\tau)$ and $f_R(\tau)$ are defined on the $Im(k) = \mp A$ and $Re(k_x) < -A$ and $Re(k_x) > A$ respectively (see Eq. (6.20)). However, since we use limited number of Gabor coefficients, the approximation of the function can not just stop abruptly at $Re(k_x) = \mp A$. Beyond $\mp A$ the $f_{L/R}$ have to be attenuated smoothly. As we illustrate in Figure 6.3, the approximation on the solid line is continued for some distance along the dashed line to let $f_{L/R}$ be smooth at $Im(k_x) = \mp A$. Continuing the function $f_{L/R}$ on the dashed line with an attenuating factor is not always possible, since γ has branch cuts located at $k_x = \pm k \mp jA$ with $k > 0$, making approximations discontinuous at some point beyond $Re(k_x) > 0$, $Re(k_x) < 0$ respectively. The Gibbs ringing created by such a discontinuity significantly deteriorates the accuracy of the $f_{L/R}$ approximation, since the Gibbs ringing carries over some distance. For this reason it is important that the functions to be approximated are made continuous along the complete attenuating region.

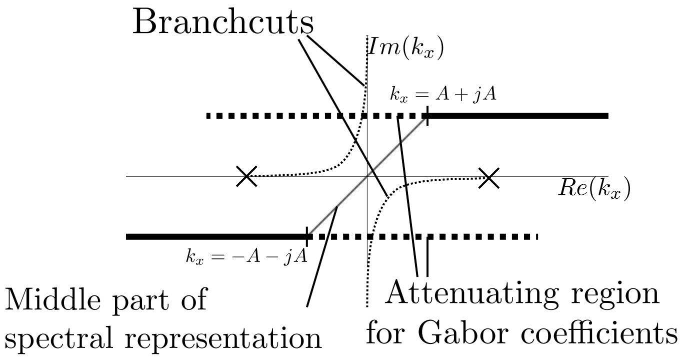
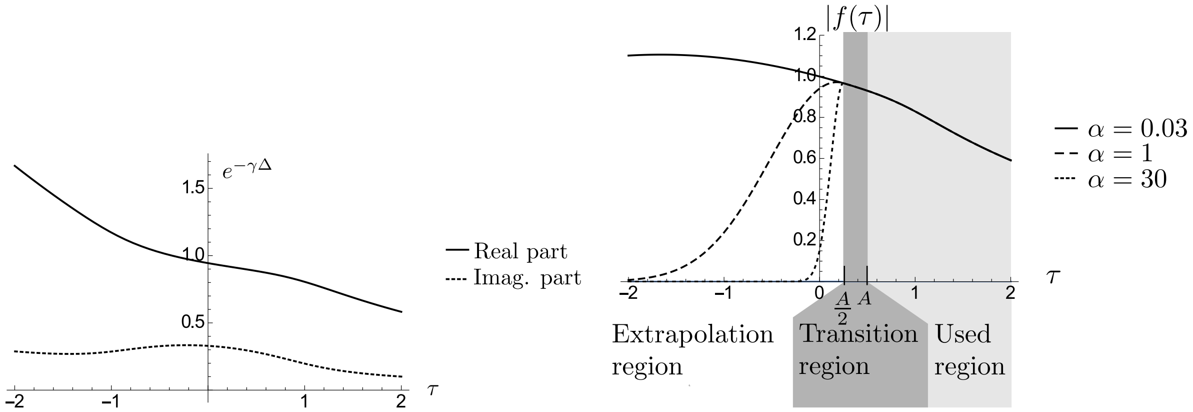


Figure 6.3: Branchcuts can interfere with the attenuating parts of $f_{L/R}(k_x)$

6.6.2 Propagation function

The Δ-distance propagation function $\exp(-\gamma\Delta)$ is multiplied by itself N_z times during the recursion in Eq. (6.13). To guarantee numerical stability, this function has to be equal to or smaller than 1 in modulus everywhere. In the domains of f_L and f_R this function is well behaved, but it exhibits branch cuts at $k_x = \mp jA$. Avoiding the branch cut by continuing on the other Riemann surface lets the function increase beyond one, as is shown in Fig. 6.4.a.



(a) (b)

Figure 6.4: (a) A plot of the propagation function $e^{-\gamma\Delta}$ on the line $k_x = k - jA$. Its absolute value is clearly larger than one for $k < 0$. (b) The continuation using Eq. (6.28) for different values of α. For $\alpha = 1$ and $\alpha = 30$ the absolute value of the propagation function stays below 1. However, for $\alpha = 30$ the function is not smooth enough for good approximation with Gabor coefficients.

We solve this issue by multiplying a linear continuation of $f_{L/R}$ beyond $k_x = \pm A/2 \pm jA$ by a Gaussian. For f_R we calculate the Gabor coefficients from the function

$$f_R^c(\tau) = \begin{cases} e^{-\gamma\Delta} & \text{if } \tau > A/2 \\ (a\tau + b)e^{-\alpha(\tau - A/2)^2} & \text{if } \tau \leq A/2, \end{cases} \tag{6.28}$$

where a and b are fitted so that the function has a continuous derivative at $\tau = A/2$ and α is chosen such that this function will just be smaller than 1 for all k_x with $Im(k_x) = \mp A$, as is shown in Fig. 6.4 (b). When α is large, the transition is very fast and the Gabor frame may need too many coefficients. When α is small, f_R^c increases to values larger than one, potentially destabilizing our algorithm. We use a numerical optimization to find the largest α for which the approximation does not increase beyond 1.

6.6.3 Reflection coefficients

For the reflection coefficients the continuation of the function beyond $\tau = \pm A$ is difficult, because we can expect to encounter poles in that region. However, we do not have the requirement of the function being smaller than 1, so we can use the same technique as applied for the propagation in Eq. (6.28) function but now with $\alpha = 1$.

Another critical point is that an analytic expression for the reflection coefficients does not exist in general. Therefore, we cannot calculate the derivatives for $\tilde{f}_d$ analytically. The way we approximate the derivatives is by fitting a power series through the values of f at different k_m. From this power series we can approximate the derivatives. We define $k_m = -\sqrt{2A}m/M, \ldots, \sqrt{2A}m/M$ for $m \in \{-M, -M+1, \ldots, M-1, M\}$, with $2M + 1$

larger than the total required number of derivatives. Now we can calculate fit values $f_m = f(k_m)$.

If the Taylor series Eq. (6.24) is to hold, then we enforce:

$$f_m = \sum_{d=0}^{2M} \frac{k_m^d \tilde{f}_d}{d!} = \sum_{d=0}^{2M} K_{m,d} \tilde{f}_d, \tag{6.29}$$

which is a matrix equation on the right, with $\mathbf{K}_{m,d} = k_m^d/d!$. This Vandermonde system can be solved by (pseudo-)inverting the matrix, which yields the coefficients $\tilde{f}_d$. Since this Vandermonde system is small, ill conditioning is not a problem.

6.6.4 Cut function and contrast function

The cut functions $c_{L/R}(k)$ in Eq. (6.21) and the contrast function $\chi(x)$ Eq. (6.1) are the only discontinuous functions present in the numerical scheme. This means that we cannot simply use the fast Gabor transformation to calculate their coefficients, because this method requires samples of the function on an equidistant grid and can therefore not sample discontinuities accurately. In principle we would have to calculate the integrals in Eq. (6.17), which is challenging. The easiest way out is to use massive oversampling (by a factor of 1000 or more) with the fast algorithm of [100]. Since this function needs to be calculated only once, during initialization, its computation time is not very critical.

6.7 Summary of the algorithm

We will now summarize the steps that need to be carried out for the complete algorithm. We have ordered these steps in a list to emphasize the chronological order of these operations.

- **Gabor frame** Set up the Gabor frame in both the spatial and spectral domains, i.e. calculate an interpolation list of the dual window function in Eq. (6.18) in both the spectral and spatial domains. Initialize the multiplication operation on the spatial and spectral Gabor frame by calculating the $A_{k;lm}$ in Eq. (26) of [118].

- **Spectral path** Initialize the spectral path by calculating spectral Gabor coefficients of cut-off functions c_L and c_R in Eq. (6.21), spatial Gabor coefficients of the $i_d(x)$ integrals in Eq. (6.26) and spatial Gabor coefficients of the exponential factors $\exp(\pm Ax)$ that are required in Eq. (6.19) and Eq. (6.22).

- **Green function** Discretize all parts the Green function on the complex spectral path using the continuation technique and other directives in Sections 6.6.2 and 6.6.3. The Green function consists of $\exp(-\gamma\Delta)$, h_e^u and h_m^u in Eq. (6.13), h_d^u and h_m^d in Eq. (6.14), $R^{\text{eff},u}$ in Eq. (6.7), $R^{\text{eff},d}$ in Eq. (6.8), $\exp(-\gamma d_i)R^u$ in Eq. (6.6) and $\exp(-\gamma d_i)R^d$ in Eq. (6.8).

- **Initialize problem** Calculate Gabor coefficients corresponding to χ in Eq. (6.1). Calculate the incoming electric field by using e.g. [87, Chapter 5] and use it to calculate the left hand side of Eq. (6.10).

- **Solve problem** Use an iterative solver, e.g. BiCGStab(2), to solve the integral equation in Eq. (6.10) for J^s. The matrix vector product on the right-hand side is computed in the following steps:

  1. Transform the contrast current density to the spectral path, Eq. (6.19) and Eq. (6.23).
  2. Compute the homogeneous waves $K^{h,u}$ and $K^{h,d}$, Eqs (6.13)-(6.14).
  3. Compute $K^{\text{eff},u}$ Eq. (6.6) and $K^{\text{eff},u}$ in Eq. (6.8) to find $E^s = K^{s,u} + K^{s,d}$ through Eqs (6.5)-(6.8).
  4. Transform back to the spatial domain, Eq. (6.22) and Eq. (6.25).
  5. Multiply by the contrast function.
  6. Add the result to J^s in the spatial domain.

- **Postprocess** When J^s is calculated, the contrast current density $J = J^i + J^s$ can be calculated and this can be used to compute various other quantities, such as the scattered field $E^s = G[J]$.

6.8 Numerical results

6.8.1 Accuracy

We have simulated two cases to validate our code. The first case consists of two blocks embedded in a three-layer medium as depicted on the left-hand side of Figure 6.5. The second case consists of eight lines on top of a silicon substrate and is depicted on the right-hand side of Figure 6.5.

The first case was validated using JCMWave software package [134], which uses a finite-element algorithm. The second case was validated using the algorithm of [124], which uses RCWA with PMLs.

For both simulations we used a Gabor frame with window width of $X = 250$ nm and $\alpha = \beta = \sqrt{2/3}$ oversampling. For the first case we used 13 spatial window functions and 25 modulation frequencies, which is a total of 325 unknowns per sample in z. To increase the accuracy we used a a factor 1.5 oversampling in the spectral domain. In the z direction we used 21 samples in z. This results in one unknown per 5 nm in the z direction and one unknown per 8 nm in the x-direction. For the second test case we used nine spatial windows, since the contrast source extends over a wider range in the x-direction. For case two a spectral oversampling of 1.2 was already sufficient. In Figure 6.6 we show the electric field strength around the objects for both cases.

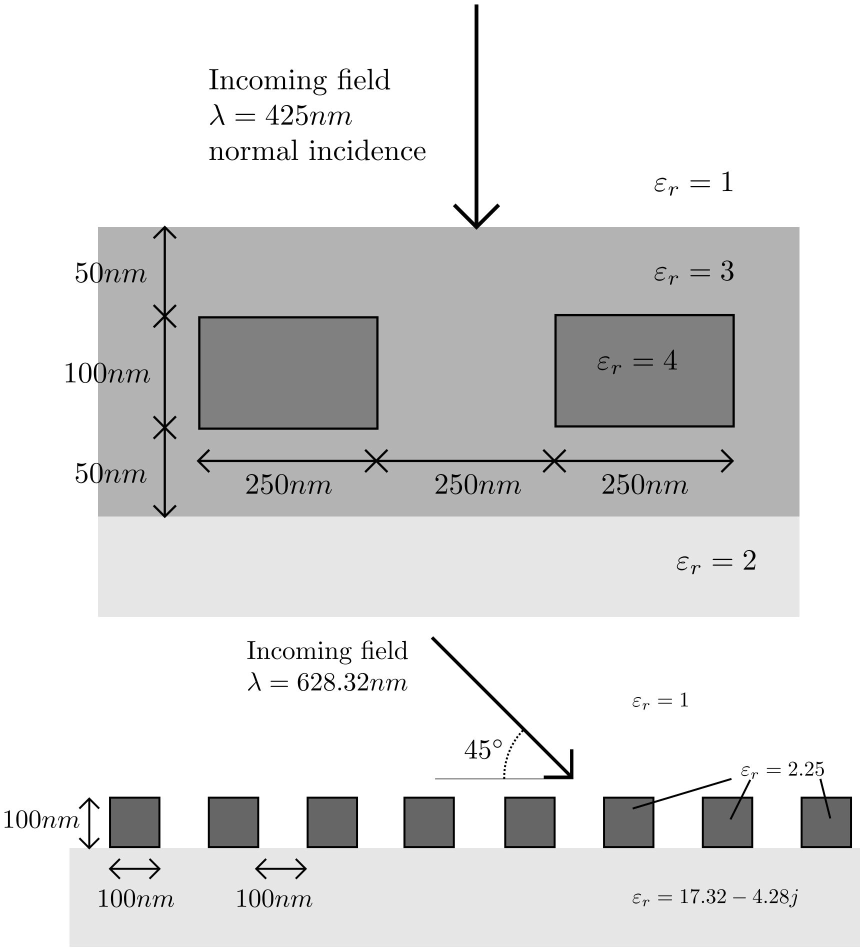


Figure 6.5: Top: the first testcase, bottom: the second testcase

From Figure 6.7, it can be clearly seen that the results from this algorithm coincide with the reference results up to a relative difference of around 10^{-3}. Only around the edges of the blocks a somewhat larger error is observed, because the analytic Gabor frame cannot exactly represent the jump in the second derivative of the electric field.

We choose to plot the error for both cases at the edge of the blocks, since the scattered field generally behaves worse around discontinuities. However, we observed that the error remains more or less constant across the whole simulation domain. The difference between the reference solution and our simulation results can be tightened by increasing the sampling.

6.8.2 Computational efficiency

To study the performance of the proposed algorithm, we first consider the convergence of the iterative solver BiCGstab(2) [97, 98], since fast convergence is critical to computational efficiency. We have used the scattering setup of the first testcase in Figure 6.5, but with

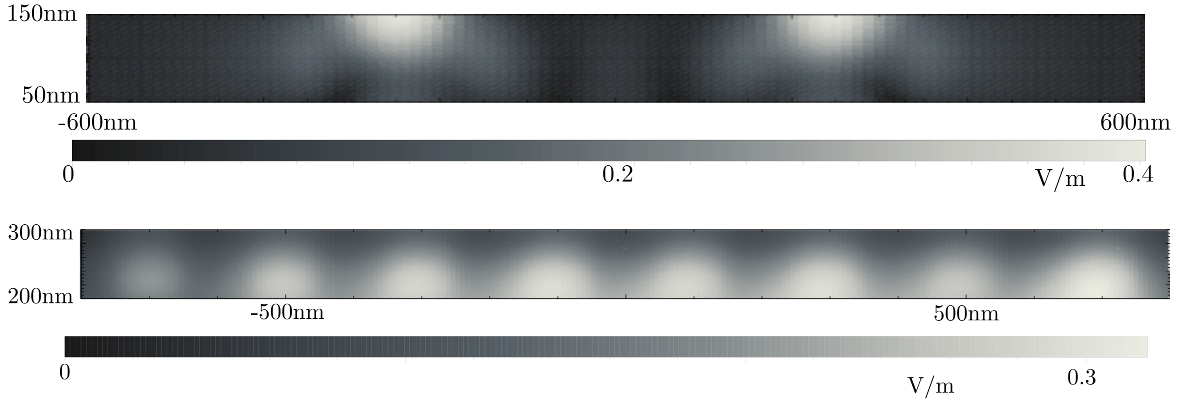


Figure 6.6: The absolute value of the scattered field E^s. Top: the first testcase. Bottom: the second testcase.

the dielectric constant of the objects increased to $\varepsilon_r = 30$. We show results with a higher contrast χ since a high contrast usually requires more iterations, and with a low contrast the convergence is too fast for an insightful plot. Figure 6.8 shows that the residual error converges rapidly and that the convergence is not very sensitive to the discretization in the x direction.

Another important aspect of this type of algorithm is the computation time. In [144], several methods have been tested for scattering problems in multilayered media. According to this article, the most competitive methods seem to be RCWA (or Fourier Modal Method) and FEM. Other methods mentioned here are FDTD implementations, which all scored poorly in accuracy, a hybrid method utilizing both FEM and RCWA, and a volume integral method (VIM) that was accurate but slower than some of the competition. Therefore, we focus on FEM and RCWA e.g. in [134, 145, 146].

To compare the present method to RCWA and FEM, we have first plotted timing results of the other methods in Figure 6.9. On the vertical axis we put the relative computation time divided by the number of discretization points, N_x, so a horizontal line corresponds to an $O(N_x)$ algorithm, with N_x proportional to the number of unknowns used in the x-direction. Some of the results are generated for periodic scatterers, which is in principle a different class of solver. However, the scaling of the computational efficiency is the same as their a-periodic counterparts that use PMLs or supercell techniques. In this graph we clearly see that RCWA is computationally more intense than $O(N_x \log N_x)$, where N_x represents the number of harmonic functions in the x-direction. RCWA scales as $O(N_x^2)$ or $O(N_x^3)$, depending on the implementation. For FEM, there are two variables, the number of mesh elements N_t and the polynomial refinement N_p. Since with FEM the discretization is in two directions we compare its performance by putting N_x equal to $N_x = \sqrt{N_t}$ or $N_x = \sqrt{N_p}$ on the horizontal axis.

We would like to emphasize that in [33] a FEM-algorithm has been published that scales linearly as $O(N_x)$ for a 3D case. Although this is somewhat superior to the $O(N_x \log N_x)$ scaling in the present method, FEM has the downside of having to discretize the entire multilayer domain and added PMLs, whereas the present method only discretizes at po-

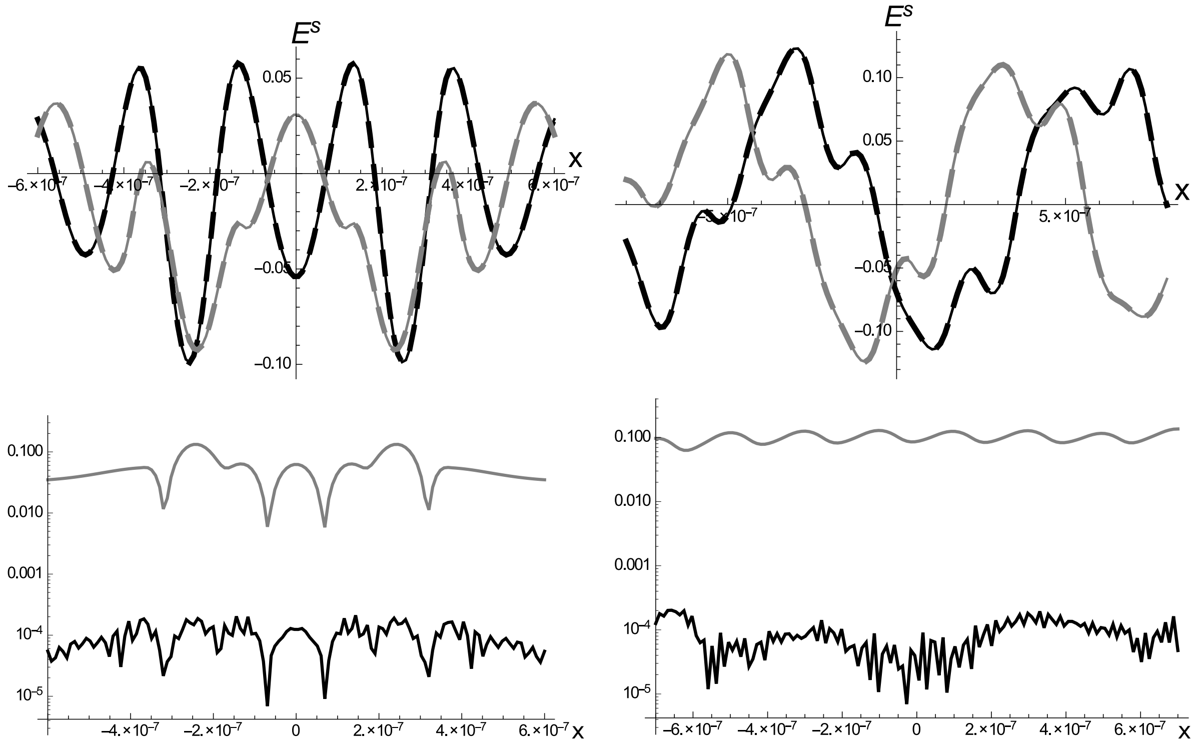


Figure 6.7: Left: The electric field for the first case, for $z = z_{min}$ at the top of the blocks as defined in Fig. 6.1; Right the second case at the interface between the layers; z_{max}. Top: The electric field (real part, black; imaginary part, gray) plotted at a cross section as depicted in Figure 6.5. The reference data (thin) and the simulation results (thick, dashed) are plotted through each other. Bottom: The absolute value of the electric field (gray) and the difference between the reference solution and the solution obtained by Gabor coefficients (black).

sitions where objects are located. There are cases where this advantage will outweigh the $O(\log N_x)$ penalty of the present method, for example in the second testcase. For that testcase, FEM needs a region of at least 1220 nm in the z-direction to be discretized, while the present algorithm only uses 70 nm of discretized space in the z-direction.

In Figure 6.10 we have plotted the computation time of the present method against the timing of a RCWA implementation [124] for the scattering case on the right-hand side of Figure 6.5. We used 20 discretization steps in the z direction for the present method and 20 slices for RCWA to mimic slanted-boundaries behavior. In the case of RCWA the wavenumber of the highest harmonic function was calculated through $2\pi/HW$, H being the number of harmonics, from 1 tot 512 and $W = 1900$ nm the period of the simulation domain. For the present method, we increased the spectral range by increasing the number of the n-index basis function in Eq. (6.16). Clearly, for RCWA the difference in computation time between TE and TM polarization is small, which justifies that we compare our method against results for TM polarization in Figure 6.9. Although RCWA

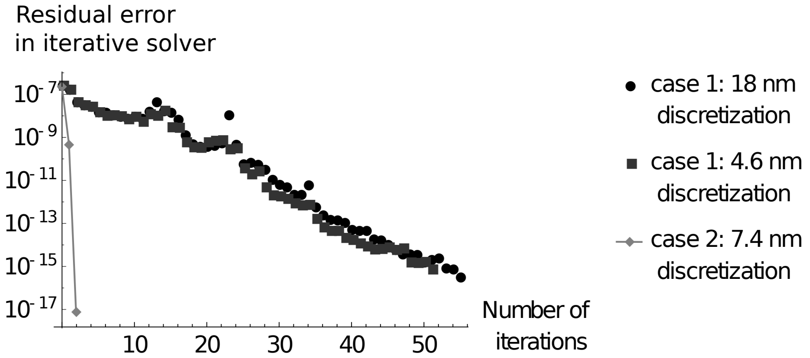


Figure 6.8: The convergence of the BiCGStab(2) algorithm for the first testcase with an increased contrast and for the unchanged second testcase of Figure 6.5.

is noticeably faster for a small number of harmonics, this spectral range is too low to accurately capture the details of the problem. When a large spectral range is required, either for higher accuracy and/or for a larger simulation domain, the present method is clearly beneficial.

6.8.3 Application to a grating coupler

As a final example we show results for a grating coupler that was inspired by [23]. The application to this type of problem is challenging, since a grating coupler is a device that couples an incident field to a guided wave and therefore it shows the accuracy of the presented method for guided waves. The dielectric waveguide consists of a thin high-contrast layer deposited on a thick low-contrast layer. Waves are coupled into the high contrast layer through a set of grooves, as illustrated in Figure 6.11. The dimensions of this grating coupler were not further optimized for this particular wavelength to the extent of what was done in [23].

To characterize this setup, a Gabor frame with window width $X = 1550$nm and $\alpha = \beta = \sqrt{2/3}$ truncated to 17 window functions with 241 modulation frequencies was used in the x-direction. In the z-direction 21 basis functions were used with a width of 3.5 nm to span the height of the grooves. The Taylor series in the spectral integration path was truncated at 33 terms. The required computation time to solve the complete problem was 560 seconds. Again, this result was validated against results generated with the FEM-algorithm JCMWave [134], which yielded a relative error of 4.6×10^{-5} between the present algorithm and the FEM reference. Figure 6.12 clearly shows that a guided wave is induced in the high-contrast layer that travels to the left.

The convergence of the accuracy against different discretization parameters is shown in Figure 6.13. For each of the figures all simulated parameters are kept at the values mentioned above, except for the one in which a parameter sweep is performed. In Figure 6.13(a), the convergence is shown against the number of modulation frequencies in the Gabor frame. The number of modulation frequencies ranges from 11 to 241, which corresponds to a sampling frequency of 120 nm down to 5.2 nm. In Figure 6.13(b), the

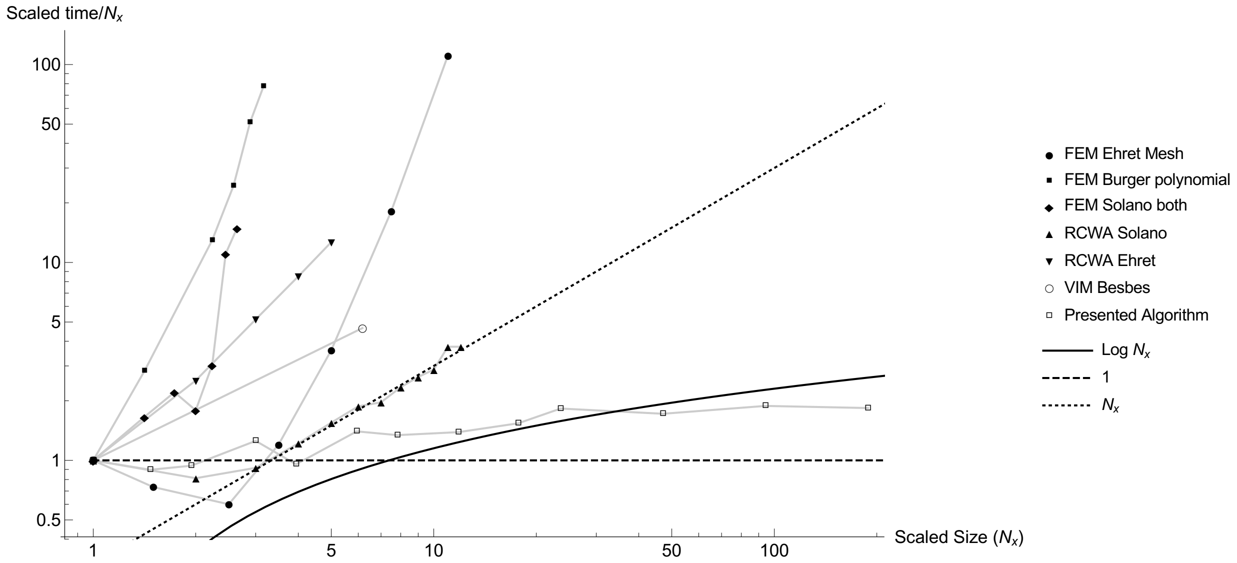


Figure 6.9: The scaling of computation time with the number of unknowns in the x-direction (N_x). The data has been scaled to coincide for the smallest simulation size. The origins of the data are Ehret: [145] on a periodic problem, Burger: [134] on an a-periodic problem, Solano: [146] on a periodic problem, Besbes: [144] on an a-periodic problem.

convergence is shown against the number of PWL functions that discretize the z-direction in the range from 2 to 21. This corresponds to a width Δ of the PWL function, Eq. (6.11), from 70 nm to 3.5 nm. In Figure 6.13(c), the convergence is shown against the number of terms in the Taylor series Eq. (6.25). Since the calculation time is not significantly affected by the number of terms, the safest strategy is to use a rather large number of terms in the Taylor series. Clearly, a truncation at 8 terms is already sufficient. For a large number of terms in the Taylor series, the accuracy is limited via the PWL basis and the Gabor frame.

6.9 Conclusion

We have presented an algorithm capable of calculating the scattering from 2D dielectric objects embedded in a layered medium, using a fully a-periodic approach under illumination of a TE-polarized wave based on a domain integral equation. The use of a Gabor frame makes it possible to efficiently approximate functions in both the spatial and the spectral domain simultaneously, which we exploit by carrying out the contrast multiplication in the spatial domain and the Green operator in the spectral domain.

In the spectral domain we use a representation on a path in the complex spectral plane instead of on the real axis. An advantage of this complex spectral path is that the Green operator is smoother and therefore easier to approximate. The particular choice for this path allows for a fast FFT-based transformation between the spatial representation and the complex-path spectral representation. The computation time of the resulting algorithm scales as $O(N_x \log N_x)$.

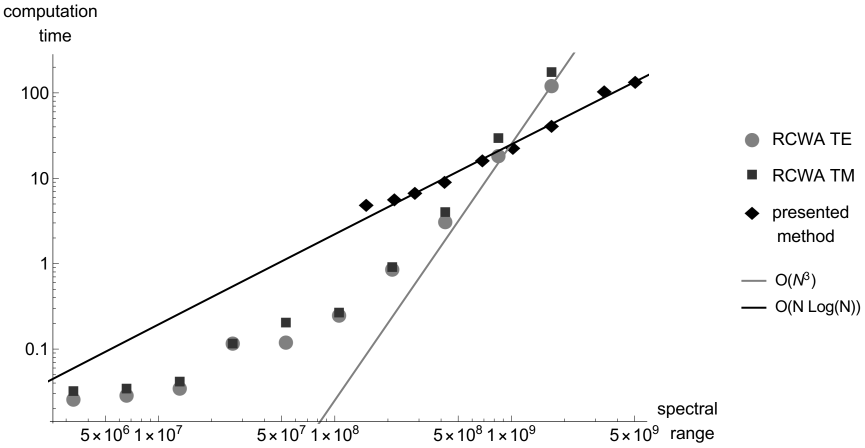


Figure 6.10: The computation times for the RCWA implementation in [124] compared to the present method. The cross-over point of both trend lines corresponds to a spatial detail of 7 nm, still coarser than the detail level of 5 nm in the z-direction. The present algorithm was testend on a single core of an intel i7-4600u processor.

Acknowledgements

The authors would like to thank dr. Maxim Pisarenco for providing us with reference data using the algorithm described in [124].

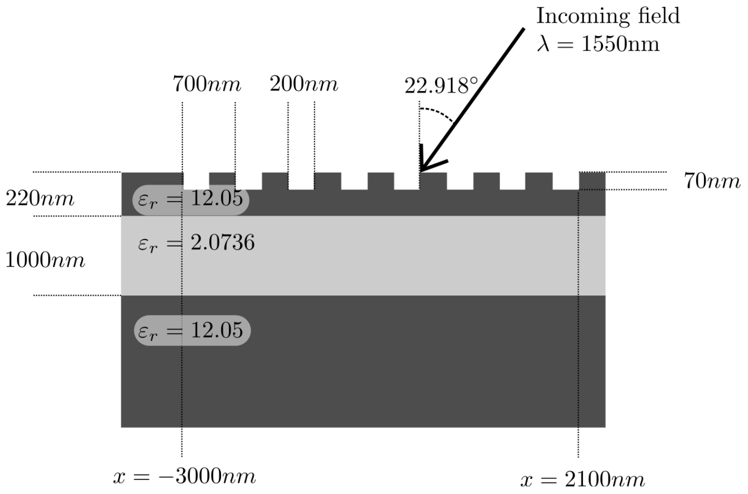


Figure 6.11: A gratingcoupler for TE mode waves.

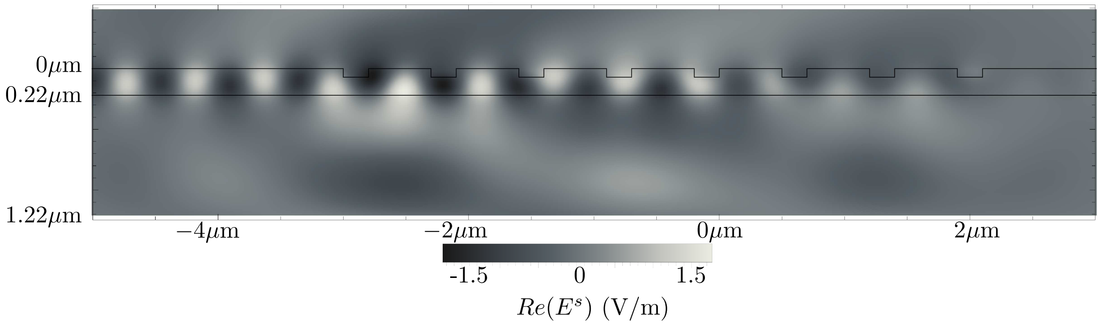


Figure 6.12: The real part of the scattered electric field in the gratingcoupler of Figure 6.11.

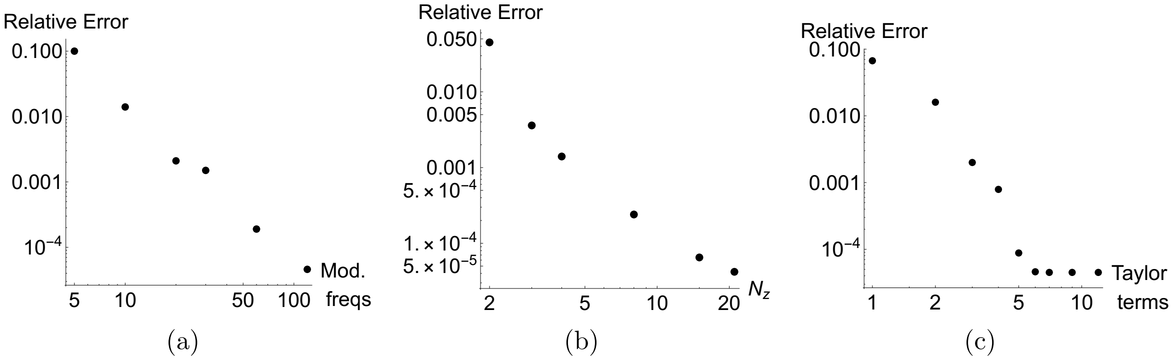


Figure 6.13: Convergence and computation time of the present algorithm to (a) an increasing number of modulation frequencies in the Gabor frame in the x-direction, (b) an increasing number of PWL basis functions in the z-direction, (c) an increasing number of Taylor coefficients.

Chapter 7

The 2D TM scattering problem for finite dielectric objects in a dielectric stratified medium employing Gabor frames in a domain integral equation[1]

7.1 Introduction

The simulation of electromagnetic scattering from finitely sized dielectric objects in a multilayered dielectric medium has several important applications. Among these applications are metrology for integrated-circuit production [148], metamaterials [138], and elements on nanophotonic chips [23]. Fast and accurate numerical methods are very important in these fields of research.

In a preceding article [135] (Chapter 6), we proposed a method to calculate the scattering from a two-dimensional dielectric object illuminated by a wave with transverse electric (TE) polarization in a layered medium. We used a domain integral equation to solve the scattering problem. There are two key ingredients to this method. The first is the use of a Gabor frame as a discretization, which ensures a fast and exact Fourier transformation. The second key ingredient is the use of a specially chosen path through the complex plane in the spectral domain on which we discretize the fields. On this path we are able to circumvent the poles and branchcuts that are present in the Green function.

In this article we show that the same ingredients can also be used for solving 2D scattering problems with Transverse Magnetic (TM) polarization. The challenge with TM polarization is that the electric field is discontinuous wherever the contrast function is discontinuous. After Lalanne and Granet discovered a method to accurately calculate the TM-polarized scattering from an object [83, 82], Li put this into a rigorous framework [81], which resulted in the so-called Li factorization rules. The key point of these articles is that when two functions with discontinuities at the same spatial position are approximated

[1] This chapter was first published as the article [147].

by a Fourier series, the product of the Fourier series does not converge well. Lalanne [83] solves the issue by replacing the discontinuous contrast operator by the inverse of a truncated inverse contrast operator. Granet [82] avoids the multiplication of functions with discontinuities at the same positions altogether by a reformulation. The way Granet handles spatial discontinuites can also be applied in the differential method formulation, where a generalization to more arbitrary shapes in three dimensions exists as the normal-vector-field formulation [84, 89]. This class of methods to handle spatial discontinuites is not unique to each spectral method, they are applicable to many different spectral methods such as the Rigorous Coupled-Wave Analysis (RCWA) also known as the Fourier Modal Method [74, 73], the periodic Volume Integral Method (pVIM) [92, 71] and the Differential Method [149].

We show that slow convergence of multiplications of functions with discontinuities at the same positions is also an issue when functions are represented by Gabor coefficients. However, following [82] we replace the electric field by an auxiliary field that is continuous. Multiplication of the discontinuous contrast function with this continuous auxiliary field yields a well-converging solution similar to the periodic case in [82]. We consider two validation cases to demonstrate that this spatial-spectral approach yields accurate results.

7.2 Formulation

7.2.1 Problem description

Consider a two-dimensional dielectric object of finite size, described in the x-z plane by its relative permittivity function $\varepsilon_r(x, z)$. This dielectric object is embedded in one layer of a multilayered medium defined by $N - 1$ layers with dielectric constants $\varepsilon_{rb,n}$ in the region between z_n and z_{n+1} and thickness $d_n = z_{n+1} - z_n$. This is illustrated in Figure 7.1. Above the top layer there is vacuum $\varepsilon_{rb,0} = 1$ and below the lowest layer there is a halfspace with relative permitivity $\varepsilon_{rb,N}$. We assume that the dielectric object is completely embedded in layer i. We define the contrast function $\chi(x, z)$ by

$$\chi(x, z) = \frac{\varepsilon_r(x, z)}{\varepsilon_{rb,i}} - 1, \tag{7.1}$$

which is nonzero only inside the object. The simulation domain with bounds $-W \leq x \leq W$ and $z_i \leq z_{min} \leq z \leq z_{max} \leq z_{i+1}$ contains the dielectric object completely.

We define the incoming field $\mathbf{E}^i$ as the field in the multilayer medium in absence of the scatterer. This field has transverse magnetic (TM) polarization, i.e. its magnetic field $\mathbf{H}^i$ is directed in the transverse y direction, so the $\mathbf{E}^i$ field lies in the x-z plane. Since scattering will keep $\mathbf{H}$ pointing in the y direction, $E_y = 0$ everywhere, which turns this problem into a two-dimensional one. When we define the total electric field $\mathbf{E}$ as the solution to this scattering problem, the scattered field $\mathbf{E}^s$ can be found from $\mathbf{E}^i = \mathbf{E} - \mathbf{E}^s$.

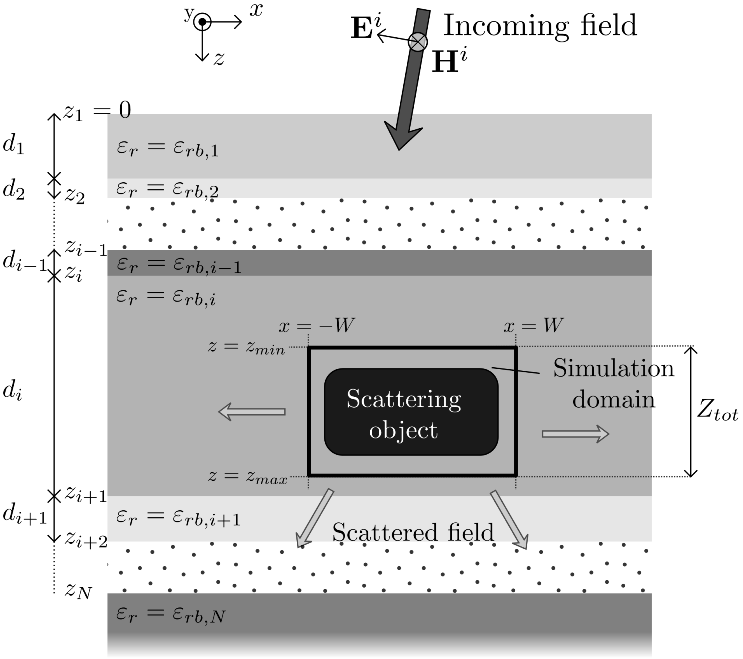


Figure 7.1: Scattering setup. A TM polarized field is incident on a dielectric object located in layer i of a multilayered background medium.

7.2.2 The integral equation in the spatial domain

For $e^{j\omega t}$ time convention, the integral equation can be written as the combination [2]

$$\begin{aligned} \mathbf{E}^i(x,z) =& \mathbf{E}(x,z) - \mathbf{E}^s(x,z) = \mathbf{E}(x,z) - \\ & \int_{-W}^{W} dx' \int_{z_{min}}^{z_{max}} dz' \, \frac{k_0^2}{j\omega\varepsilon_0\varepsilon_{rb,i}} \, \mathcal{G}(x,z|x',z') \cdot \mathbf{J}(x',z') \\ \mathbf{J}(x,z) =& j\omega\varepsilon_0\varepsilon_{rb,i}\chi(x,z)\mathbf{E}(x,z), \end{aligned} \tag{7.2}$$

where the $\mathcal{G}$ denotes the rank-two Green-function tensor in x and z, $\mathbf{J} = (J_x, J_z)$ defines the contrast current density and $k_0^2 = \omega^2\varepsilon_0\mu_0$ defines the squared wavenumber in vacuum. With the first of these equations we can compute the scattered field from the contrast current density. The integrals of the integral equation are in the form of an integration with the Green tensor. The second equation will be called the field-material interaction.

[2] Note that this formulation is different from the one presented in Chapters 5 and 6. Previously a current-based formulation of the integral equation was employed that evades multiplications with two functions that are nonzero at the edge of the simulation domain. Such a multiplication leads to artifacts originating from the truncation of the Gabor frame, visible in the solution as strong oscillations at the edge of the domain. In this formulation such multiplications are needed in the transformation to the spatial domain, where $\exp(\pm Ax)$ is multiplied with $E_L/R(x)$ in Eq. (6.22), but the problem can also be avoided by discretizing $\exp(\pm Ax)$ for a larger range of M in Eq. (4.3). In this way the artifacts induced by the multiplication are located outside of the simulation domain.

In the x direction, the calculation of the scattered field can be handled most efficiently in the spectral domain, since there we can exploit the x-directed translation symmetry in the layered background medium. In the z direction, perpendicular to the layer interfaces, it is most convenient to work in the spatial domain, since there is no translation symmetry. For the field-material interaction we work in the spatial domain in both directions.

In the next sections we will first discuss the Green function operator (Section 7.2.3). Then we describe how we discretize (7.2) (Section 7.2.4) and afterwards we will explain how we can accurately compute the field material interaction (Section 7.2.5).

7.2.3 The Green operator in the spectral domain

We use the Fourier transformation defined by

$$f(k_x) = \mathcal{F}_x[f(x)](k_x) = \int_{-\infty}^{\infty} dx\ f(x)e^{-jk_x x}. \tag{7.3}$$

In the spectral domain, we write functions with k_x as an argument and in the spatial domain with argument x. The Fourier transform of a function will be meant when the argument has changed from x to k_x and vice versa.

The Green operator can be written as a sum of two parts. The first part, $\mathcal{G}^h$, represents the radiation into a homogeneous space with background dielectric constant $\varepsilon_{rb,i}$. The second part represents the reflections at the layer interfaces. The scattered field due to the homogeneous part of the Green function can be written as

$$\mathbf{E}^h(k_x, z) = \frac{k_0^2}{j\omega\varepsilon_0\varepsilon_{rb,i}} \int_{z_{min}}^{z_{max}} dz'\ \mathcal{G}^h(k_x, z|z') \cdot \mathbf{J}(k_x, z), \tag{7.4}$$

where the homogeneous Green function is given by

$$\mathcal{G}^h(k_x, z|z') = -\begin{pmatrix} k_0^2\varepsilon_{rb,i} - k_x^2 & jk_x\partial_z \\ jk_x\partial_z & k_0^2\varepsilon_{rb,i} + \partial_z^2 \end{pmatrix} \frac{e^{-\gamma|z-z'|}}{2\gamma k_0^2}, \tag{7.5}$$

with $\gamma^2 = \varepsilon_{rb,i}k_0^2 - k_x^2$. Note how we can identify a propagating part $e^{-\gamma|z-z'|}$ in $\mathcal{G}^h$, that governs how the electric field propagates and/or decays over a distance $|z - z'|$ in the z direction.

The second part of the Green operator adds the reflections, originating from the layer interfaces, to the scattered electric field, i.e.

$$\begin{aligned} &\frac{k_0^2}{j\omega\varepsilon_0\varepsilon_{rb,i}}(\mathcal{G} \circ \mathbf{J})(k_x, z) = \mathbf{E}^s(k_x, z) = \mathbf{E}^h(k_x, z) \\ &+ \left(\mathcal{R}^{u,u}(k_x) \cdot \mathbf{E}^h(k_x, z_{min}) + \mathcal{R}^{u,d}(k_x) \cdot \mathbf{E}^h(k_x, z_{max})\right) e^{-\gamma(z-z_{min})} \\ &+ \left(\mathcal{R}^{d,d}(k_x) \cdot \mathbf{E}^h(k_x, z_{max}) + \mathcal{R}^{d,u}(k_x) \cdot \mathbf{E}^h(k_x, z_{min})\right) e^{-\gamma(z_{max}-z)}. \end{aligned} \tag{7.6}$$

Here the $\mathcal{R}^{\alpha,\beta}$ denote the effective reflection coefficients, see [87, Chapter 5], from the layers below and above layer i including the offsets $z_{min} - z_i$ and $z_{max} - z_{i+1}$. Here $\beta = u/d$ (up/down) denotes the z propagation direction of the wave which generates the reflection and α denotes the direction in which the reflected wave itself propagates [135].

7.2.4 Discretization and spectral path

For discretization in the x-direction we employ the Gabor frame as defined in [99] Chapter 8, with Gaussian window function

$$g(x) = 2^{\frac{1}{4}} e^{\left(-\pi \frac{x^2}{X^2}\right)}, \tag{7.7}$$

where X defines the width of the window function. For better convergence, rational oversampling by a factor $1/\alpha\beta$ is employed, with the Gabor frame defined by

$$g_{mn}(x) = g(x - m\alpha X) e^{jn\beta K x}, \tag{7.8}$$

where $K = 2\pi/X$, the spectral step. To calculate Gabor coefficients, we use the dual frame found from the Moore-Penrose inverse [99, 100]. More details on the use of Gabor frames as a discretization for integral equations can be found in [118, 135].

Following the approach of [72, 150, 71], we use piecewise-linear expansion functions in the z direction

$$\Lambda_n(z) = \begin{cases} 1 - \frac{|z-n\Delta-z_{min}|}{\Delta} & \text{if} \quad |z - n\Delta - z_{min}| < \Delta \\ 0 & \text{if} \quad |z - n\Delta - z_{min}| > \Delta \end{cases}. \tag{7.9}$$

For the test functions we use Dirac delta functions at $z = n\Delta + z_{min}$. In [135] (Chapter 6) it is explained how the z'-integral in (7.4) can be computed efficiently.

In the x direction we use the Gabor frame as a basis and its dual to test, as explained in [118] (Chapter 5). There it is also explained that in the spectral domain we do not represent functions on the real axis, but instead on the path, $\tau \in \mathbb{R}$,

$$k_x(\tau) \in \begin{cases} \tau - jA & \text{if } \tau < -A \\ (1+j)\tau & \text{if } -A \leq \tau < A \\ \tau + jA & \text{if } \tau > A \end{cases}. \tag{7.10}$$

For A we choose a fixed value such that $AW \approx 3$. When a function $f(x)$ is transformed to the spectral domain, it is split up into $f_L(k_x)$, $f_M(k_x)$ and $f_R(k_x)$, each corresponding to one of the subsequent cases in (7.10). For $f_L(k_x)$ and $f_R(k_x)$ we use Gabor frames to represent these functions and for the middle part $f_M(k_x)$ we use a Taylor series. Since A is small compared to the total spectral range in which information is contained, the middle part contains little information and the Taylor series can be truncated after a few terms.

7.2.5 The field-material interaction

The main difficulty encountered in the TM scattering problem compared to the TE scattering problem is that the electric field has discontinuities wherever the contrast function has discontinuities. The electric field for TE scattering is continuous, so there we do not encounter this problem.

For RCWA it was pointed out in [83, 82, 81] that the convergence of a spatial-domain multiplication of two functions with a discontinuity at the same position is poor. In a spectral basis, such as in RCWA, this spatial multiplication is represented in the spectral domain by a convolution. When both functions have a spatial discontinuity, their spectral convergence is poor and the convergence of their convolution cannot be guaranteed. Wherever the contrast function is discontinuous, the electric field also has a discontinuous component, which leads to poor convergence in the field-material interaction in (7.2).

Although we use the Gabor frame instead of a Fourier series as a discretization, the same convergence problem comes into play. A function $f(x)$ represented by a set of Gabor coefficients f_{mn} can be written as

$$\begin{aligned} f(x) &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f_{mn} g(x - m\alpha X) e^{jn\beta Kx} \\ &= \sum_{m=-\infty}^{\infty} \mathring{f}_m(x) g(x - m\alpha X), \end{aligned} \tag{7.11}$$

with

$$\mathring{f}_m(x) = \sum_{n=-\infty}^{\infty} f_{mn} e^{jn\beta Kx}. \tag{7.12}$$

Now $\mathring{f}_m(x)$ is the resulting periodic function of the Fourier series in n, so a Gabor-frame representation can be seen as a collection in m of Fourier series in n. If $f(x)$ is discontinuous, then also (some of) the $\mathring{f}_m(x)$ are discontinuous. For a spatial multiplication, products with $\mathring{f}_m(x)$ are required and therefore again poor convergence is obtained with the Gabor frame when both functions have discontinuities at the same locations. In Figure 7.2 we illustrate this effect for a Heaviside step function. We use the tilde here to denote a truncated Gabor approximation of a function. Since the Heaviside step function $H(x)$ equals its square: $H(x) = H^2(x)$, no noticeable difference should be visible between the Gabor approximated $\tilde{H}$ and the Gabor approximated square $\widetilde{\tilde{H} * \tilde{H}}$. Obviously, there is a significant difference visible in Figure 7.2, hence the multiplication of discontinuous functions represented by Gabor coefficients is not accurate. An important difference is that the location of the step has shifted to the right. When this convolution would be applied to the field-material interaction, this would lead to a significantly smaller contrast current density and therefore to inaccurate results. For a good approximation these functions should overlap, since the same discretization is used on both. Although this example is different from the example used in [81], it is obvious that a significant error is made in the multiplication of discontinuous functions.

A reformulation of the problem is possible such that only one function is discontinuous [82, 92, 88, 71]. Let us consider a rectangular scatterer that is aligned with the layer interfaces. The electric-field component normal to a material interface is discontinuous and the electric field parallel to the interface is continuous. However, the electric flux density $\mathbf{D} = \varepsilon_r \varepsilon_0 \mathbf{E}$ normal to a material interface is continuous, whereas the electric flux density parallel to the interface is discontinuous [85], Section 1.5. According to the Li

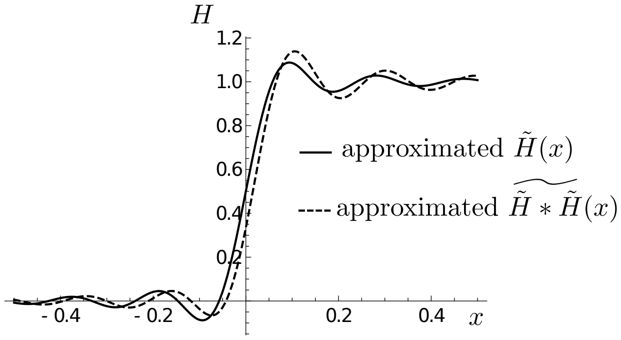


Figure 7.2: Step functions approximated by Gabor coefficients truncated to $m \in -6, \ldots, 6$ and $n \in -6, \ldots, 6$ in a Gabor frame with $X = 1$ and $\alpha = \beta = \sqrt{2/3}$. Solid line: a direct approximation of the step function. Dashed line: An approximation of the square of the Gabor-represented step function, computed by a truncated convolution.

rules [81], we should select the continuous components. Let us assume the scattering from a rectangular object aligned with the coordinates. The discontinuity at the top and the bottom of the rectangle can be dealt with in the spatial z discretization. However, at the left and right side of the rectangle, $E_x(x, z)$ is discontinuous along the x coordinate and therefore the Li rules are violated, so poor convergence can be expected for these interfaces. Now $D_x(x, z)$ and $E_z(x, z)$ are continuous at the sides of the of the rectangle

To address the problem of convergence we define the field $\mathbf{F}(x, z) = \hat{\mathbf{x}} D_x(x, z)/\varepsilon_0 \varepsilon_{rb,i} + \hat{\mathbf{z}} E_z(x, z)$. We can calculate the electric field from $\mathbf{F}$ by

$$\mathbf{E} = \mathcal{L}_\chi \cdot \mathbf{F} = \begin{pmatrix} \frac{1}{1+\chi} & 0 \\ 0 & 1 \end{pmatrix} \cdot \mathbf{F} \tag{7.13}$$

and

$$\mathbf{J} = \mathcal{M}_\chi \cdot \mathbf{F} = j\omega \varepsilon_0 \varepsilon_{rb,i} \begin{pmatrix} 1 - \frac{1}{1+\chi} & 0 \\ 0 & \chi \end{pmatrix} \cdot \mathbf{F}. \tag{7.14}$$

Following the notation in [92], we rewrite (7.2) in a single equation as

$$\mathbf{E}^i = \mathcal{L}_\chi \cdot \mathbf{F} - k_0^2 \mathcal{G} \circ (\mathcal{M}_\chi \cdot \mathbf{F}). \tag{7.15}$$

In the next section this formulation will be shown to converge much better when we use Gabor frames in the x direction compared to the case where the Li rules have been ignored, i.e. when we choose

$$\mathcal{L}_\chi = \mathrm{Id} \tag{7.16}$$

$$\mathcal{M}_\chi = \chi \, \mathrm{Id}, \tag{7.17}$$

with Id the 2×2 identity matrix. We note that more general objects can in principle be treated by using normal-vector fields [84, 151, 89, 90].

7.3 Results

7.3.1 Accuracy

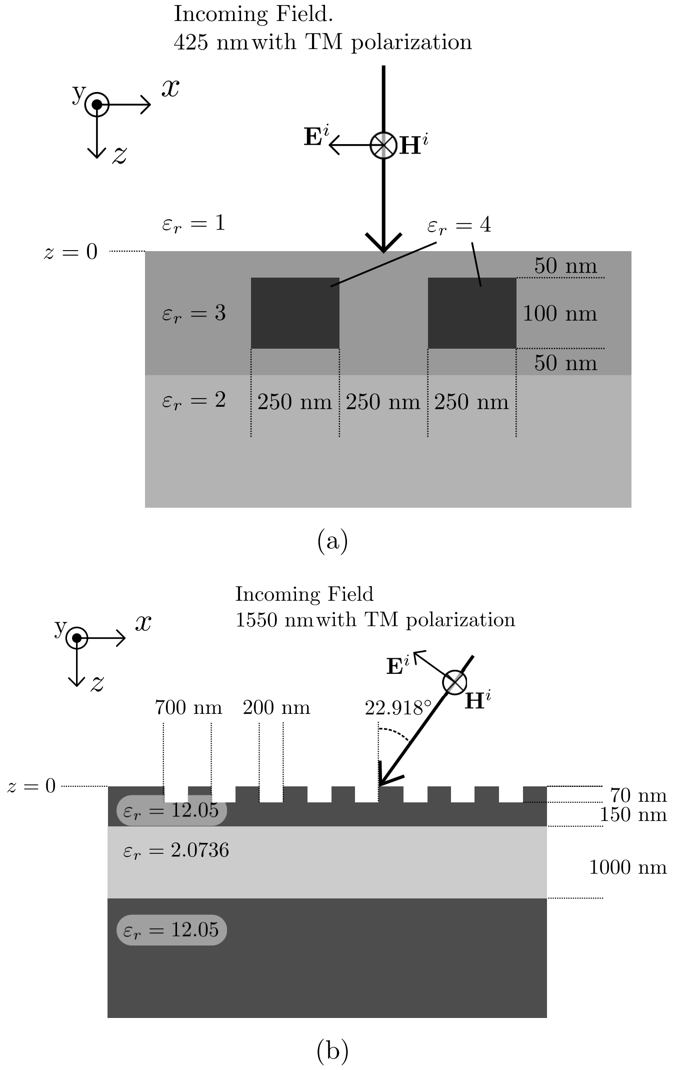
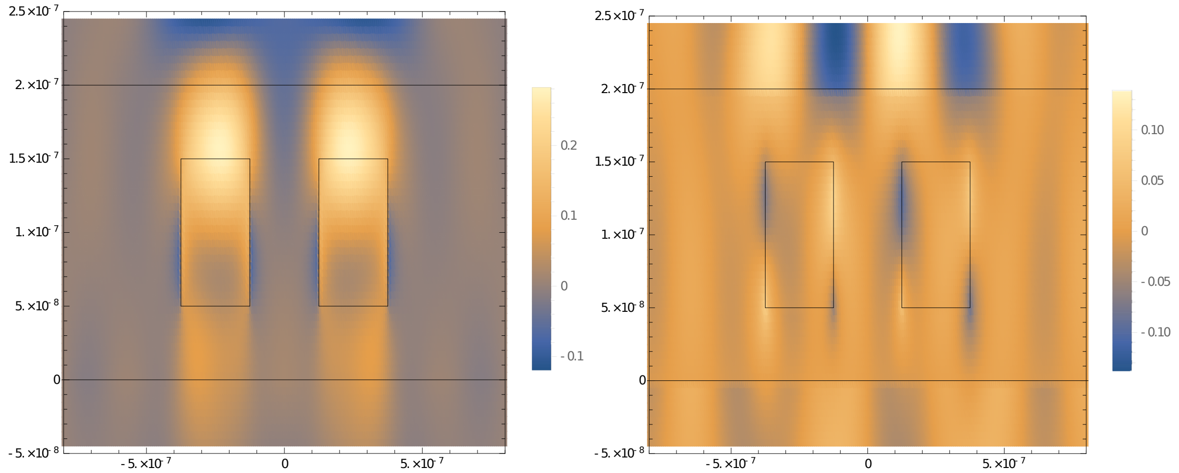


Figure 7.3: (a) The first usecase consists of two blocks in a layered medium. (b) A grating coupler consisting of grooves in a thin high-contrast medium on top of a thick low-contrast layer.

We have validated the above outlined algorithm against the JCMWave software package [134] for two different usecases. We aimed for a relative accuracy of 10^{-3}, since engineering parameters like the material properties are often determined with less or similar precision for most practical applications. The simulation parameters were chosen with this criterion in mind and optimized for speed. The first usecase, Figure 7.3(a), consists of two blocks



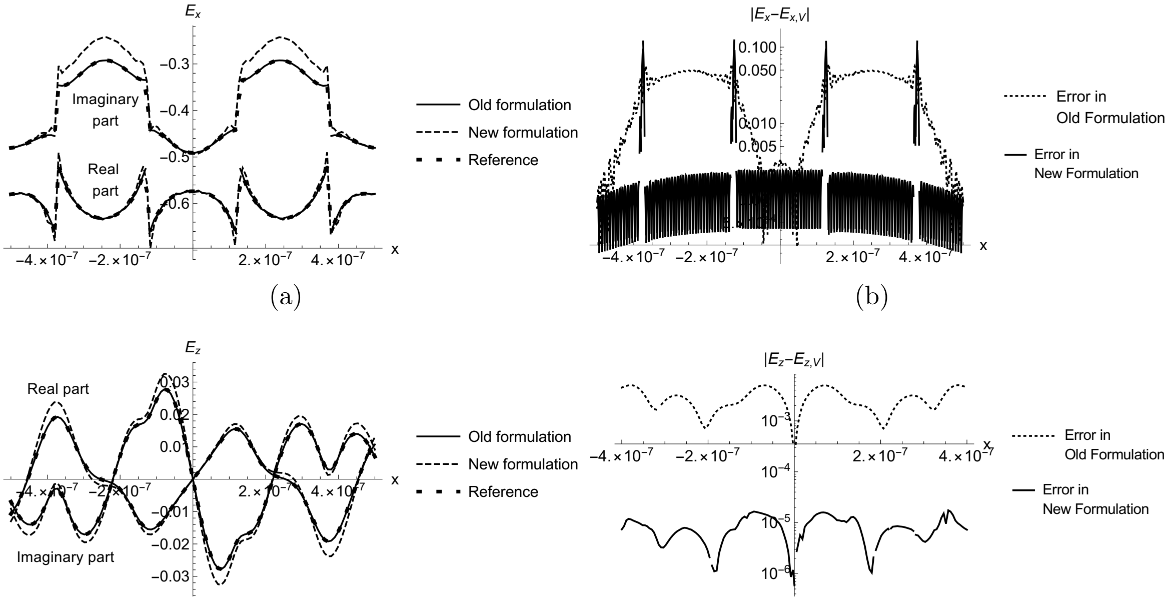
(a) (b)

Figure 7.4: The real part of (a) the scattered electric field in the x-direction and (b) the electric field in the z-direction.

in a layered medium. This is a relatively low-contrast case, since the difference in relative permitivity between the background, $\varepsilon_{rb,1} = 3$, and the blocks, with $\varepsilon_r = 4$, is small.

Figure 7.4 presents the real part of the scattered electric field $\mathbf{E}^s(x, z)$ for the geometry in Fig. 7.3 (a) excited by a normally incident plane wave of unit amplitude. The first figure represents the x-directed component of the electric field and the second figure shows the z-directed component. For this simulation we used one piecewise-linear basis function ((7.9)) per 2.5 nm in the z direction. In the x direction a Gabor frame was chosen with $X = 250$ nm in (7.7), $\alpha = \beta = \sqrt{3/2}$ in (7.8), and index $m \in \{-5, \ldots, 5\}$ and index $n \in \{-6, \ldots, 6\}$ in (7.8), totalling 143 Gabor coefficients in the x direction, equaling one coefficient per 15.7 nm on a simulation domain at some distance around the object. We chose the discretization in both x- and z-directions such that it contributed approximately the same error to the end result. Clearly, the Gabor coefficients in the x direction are more efficient in accurately discretizing the problem than the piecewise-linear functions in the z direction in the sense that a coarser discretization can be applied. We used 40% extra Gabor coefficients in the spectral domain for a finer sampling of the auxiliary field in (7.4).

Figure 7.5 shows a comparison with respect to JCMWave for the case in Figure 7.3 (a), in Figure (a) and (b) through the middle of the blocks at $z = 100$ nm, and in (c) and (d) at $z = 10$ nm, just below the upper interface. Results with the auxiliary field formulation ((7.13) and (7.14)) and without the auxiliary field formulation ((7.16) and (7.17)) are shown, so they can be compared. In Figure 7.5 (a) we show the electric field $E_{x,V}(x, 100\text{nm})$ from the JCMWave validation, $F_x(x, 100\text{nm})/(1 + \chi(x, 100\text{nm}))$ from the presented algorithm with auxiliary field formulation, and $E_x(x, 100\text{nm})$. It is clear that the accuracy found by using the auxiliary field formulation is much better, although we observe some Gibbs ringing in the auxiliary field formulation as well in Figure 7.5(b). The discontinuity of the dielectric object induces the Gibbs phenomenon on the solution. Since this



(a)

(b)

(c)

(d)

Figure 7.5: The electric field for the case in Figure 7.3 (a). In (a),(b) it is E_x at $z = 100$ nm and in (c),(d) it is E_z at $z = 10$ nm. (a),(c) show field strength. With old formulation we mean results obtained without the auxiliary field formulation ((7.16) and (7.17)), and with the new formulation the described algorithm with auxiliary field formulation $\mathbf{F}$ is meant ((7.13) and (7.14)). (b),(d) show the difference between simulation and reference.

Gibbs error has a very high frequency, it does not radiate very far away from the blocks. For example, in scattering calculations this error does not contribute to the long-distance scattering. On the blocks, the Gibbs phenomenon dominates the error, but at a distance the Gibbs ringing is attenuated, so only the error that really radiates dominates there. In Figures 7.5 (c) and (d) we have plotted the electric field 10 nm below the upper layer. Here the Gibbs phenomenon does not play a role anymore and the results obtained using the auxiliary field formulation ((7.13) and (7.14)) have a relative accuracy better than 10^{-3}. However, without the auxiliary-field formulation, the error is at least two orders of magnitude larger.

The second usecase, Figure 7.3(b), was inspired by [23], where several setups for grating couplers with TE polarization were introduced. We have chosen the grating coupler geometry and angle of incidence such that it couples TM waves efficiently into the same multilayer medium. However, the geometry was not optimized for optimal coupling to the same degree as in the original article.

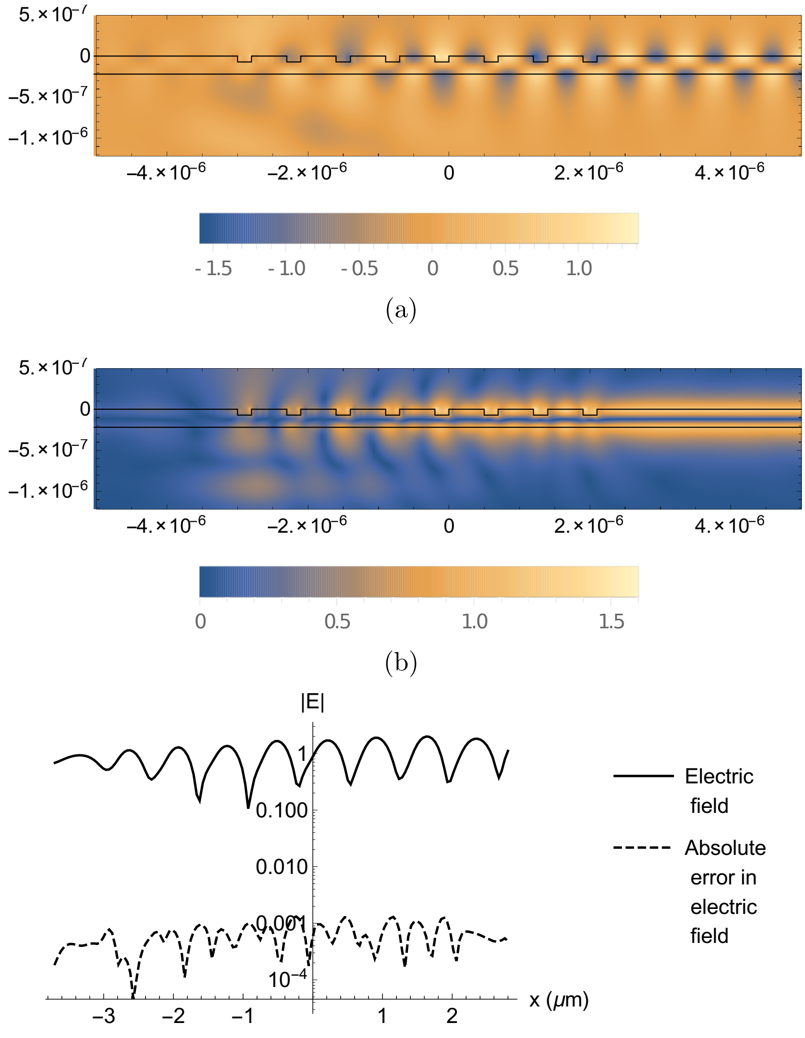
The electric field of the second test case is presented in Figure 7.6 for excitation by a plane wave of unit amplitude. It can be clearly seen that the incoming waves couple to a right-travelling wave trapped within the 220 nm high-contrast layer. Since the simulation domain can be limited to the grooves in the multi-layer medium, the simulation domain was chosen from $z = 0$ tot $z = 70$ nm in 15 piecewise-linear expansion functions, which equals one basis function per 4.6 nm. In the x direction a Gabor frame was chosen with $X = 1550$ nm in (7.7), $\alpha = \beta = \sqrt{3/2}$ in (7.8), and index $m \in \{-7, \ldots, 7\}$ and index $n \in \{-40, \ldots, 40\}$ in (7.8), totalling 1215 Gabor coefficients in the x direction, which equals one coefficient per 15.6 nm on a simulation domain around the object.

These results were also validated using JCMWave. In Figure 7.6 (c), the difference between JCMWave and results obtained with the present algorithm are shown. The results obtained with the auxiliary field formulation ((7.13) and (7.14)) agree well up to a level of 10^{-3}, however, the iterative solver did not converge to even 1 digit precision in 300 iterations[3]for the formulation without auxiliary field ((7.16) and (7.17)), whereas the auxiliary field formulation converged in fewer than 25 iterations with BiCGStab(2) [39]. We calculated the error from the field strengths at the lower side of the high-contrast layer at $z = 220$nm to reduce the Gibbs ringing. From this we can conclude that the amplitude of the wave coupled into the layer agrees with the JCMWave results for the auxiliary field formulation.

7.3.2 Computation time

To see how the computation time of our algorithm scales when the discretization is refined, we have refined the discretization both in the x direction and the z direction, while keeping the discretization in the other direction constant. Figure 7.7 shows that the computation time scales as $O(N_z)$ in the z-direction, with N_z the number of piecewise-linear

[3] This indicates that the formulation with (7.16) and (7.17) yields an approximation error in the field-material interaction that is is so large that condition number of the resulting matrix equation is significantly altered.



(a)

(b)

(c)

Figure 7.6: The x-directed scattered field $E_x^s(x, z)$ for an incoming field of unit amplitude; (a) the real part, (b) the absolute value, (c) top line: the $|E|$ field at $z = 220$nm, bottom line: the absolute difference with the JCMWave results.

basis function in the z direction, starting from the reference at $N_z = 21$ piecewise-linear functions. The same figure also shows an $O(N_x \log N_x)$ dependence with N_x corresponding to the range of n in the number of included Gabor frame functions ((7.8)), starting from the reference $N_x = 143$ frame functions.

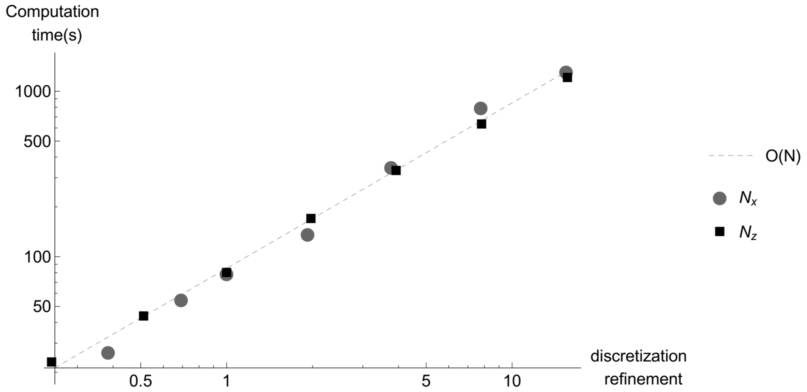


Figure 7.7: Scaling of the computation time for the case in Figure 7.3 (a), with on the x axis the factor with which the numbers of unknowns N_x and N_z were increased compared to the results in Figure 7.4 and Figure 7.5, where $N_x = 143$ unknowns and $N_z = 21$ unknowns.

7.4 Conclusion

We have succesfully reformulated the two-dimensional TM scattering problem for finitely sized dielectric scatterers in a dielectric layered medium with a volume integral equation in a mixed spatial and spectral basis in terms of a continuous auxiliary field **F** ((7.13) and (7.14)), which leads to a satisfactory convergence. A formulation without such a continuous field ((7.16) and (7.17)), which violates the Li rules, shows much poorer accuracy in one test case and in the other test case convergence of the iterative solver was not reached.

We showed numerical evidence that the computation time scales as $O(N_x N_z \log N_x)$ with respect to refinements in the discretization.

For two cases we have shown that the proposed algorithm, that employs a discretization on a path through the complex spectral plane, combined with a Gabor frame, can be used for TM polarization.

This algorithm is capable of characterizing both the scattering from dielectric objects and the coupling of waves into a dielectric layer via a grating coupler.

Chapter 8

A 3D spatial spectral integral equation method for electromagnetic scattering from finite objects in a layered medium[1]

8.1 Abstract

The generalization of a two-dimensional spatial spectral volume integral equation to a three-dimensional version for electromagnetic scattering from dielectric objects in a stratified dielectric medium is explained. In the spectral domain, the Green function, contrast current density, and scattered electric field are represented on a complex integration manifold that evades the poles and branch cuts that are present in the Green function. In the spatial domain, the field-material interactions are reformulated by employing a normal-vector field approach, which obeys the Li factorization rules. Numerical evidence is shown that the computation time of this method scales as $O(N \log N)$ with the number of unknowns. The accuracy of the method for three representative examples is compared to a finite-element method reference.

8.2 Introduction

Efficient solvers for electromagnetic scattering in stratified media are important in e.g. metrology [153, Chapter 18], metamaterials [19, 138], and integrated optics [25]. Especially for three-dimensional structures, where the number of unknowns is often very large, there is a demand for fast solvers, for which the computed complexity scales well for large numbers of unknowns. A good strategy to find a potentially efficient algorithm is to exploit symmetries. For stratified media such a symmetry is the translation symmetry in the

[1] This chapter was submitted as an article for the journal Optical and Quantum Electronics [152].

layered background medium. This symmetry can be exploited via the use of the Green function in a volume integral formulation.

In a stratified dielectric medium, an analytic expression exists for the Green function in the electromagnetic case, as a function of one spatial coordinate in the direction of stratification and two spectral coordinates in the two directions perpendicular to the stratification [53]. It is advantageous to use the stratified-medium Green function, since it incorporates the response of the multilayer medium analytically. Therefore, little computation time or memory is used for computing the scattered electromagnetic field throughout the entire layered stack, since the electric field on a domain slightly larger than the scattering object suffices. It is possible, using Sommerfeld [54] or Fourier integrals, to transform the Green function completely to the spatial domain and then use it in an integral-equation method [55, Chapter 8], [56, Chapter 5], [57, Chapter 4],[58, Chapter 2]. However, these Sommerfeld integrals are often tedious to compute, because of poles and branch cuts present in the Green function that can be located on or close to the integration path. Since the Green function has to be re-calculated for every modification in the multilayer medium, caching the Green function in a library is only advantageous when the exact same multilayered medium is used many times.

It is also possible to use the Green function directly in the spectral domain, where it is known analytically. For a periodically repeating object, the Green function decomposes into a discrete set of modes as derived in for example [92, 71]. Poles and oscillations along branch cuts in the Green function [55, 56] can be avoided on such a discrete set of modes since the modes and locations of the poles will most likely not coincide. However, for a finite scatterer the spectral domain is continuous and now the contribution of the poles and oscillations along the branch cuts are hard to discretize [118, 135]. Deformations of the Sommerfeld integration path to a complex-plane path [60, 61, 59, 62] can help to evade these poles and branch cuts. In [135] an algorithm for two-dimensional electromagnetic scattering with TE polarization in a multilayered medium is presented, where both contrast-current density and scattered field are represented on a path in the complex plane of the spectral domain. It is this path that allows for the use of Gohberg's and Koltracht's [72] fast, flexible and recursive Green-function convolution in the stratification direction.

The first challenge in the three-dimensional case is that, instead of one, now two directions perpendicular to the stratification direction need to be handled. The complex integration path is turned into a complex integration manifold and since the transformation from the spatial domain to the complex integration manifold is part of the core of the algorithm, transformations back and forth need to be computationally efficient. We show an integration plane consisting of nine regions of three distinct types and show transformations to and from the spatial domain that can be computed in $O(N \log N)$ operations, where N is the number of spectral unknowns.

The second challenge is that the discontinuity of both the permittivity and the electric field at material interfaces leads to poor convergence in spectral formulations [80]. This effect was also observed for a Gabor-frame based solver for TM-polarized scattering [147]. For periodic scattering problems with a discrete spectral expansion a reformulation of the field-material interactions corrects this poor rate of convergence [82, 83], which is explained in more detail in [81] introducing the so-called Li rules. In [147] it is shown that the same mechanism can also be used for a continuous spectral expansion and the algorithm of [135] is extended to efficiently deal with the discontinuous field-material interaction in a way that does abide by these Li rules. Here, we propose a generalization of this method to three dimensions. Inspired by [91], we show that a normal-vector field formulation [84] can be used for three-dimensional scattering to replace the field-material interaction.

We start by a brief formulation of the volume integral equation in Section 8.3. Subsequently, we give a more detailed explanation of the discretization, with emphasis on the complex-plane spectral domain representation in Sections 8.4 and 8.5, followed by a short summary of the normal-vector field framework in Section 8.6. The applicability of the present algorithm is highlighted by three numerical examples, with numerical evidence that the computation time scales as $O(N \log N)$ with the number of unknowns and comparison against a finite-element reference calculation in Section 8.7.

8.3 The volume integral equation

Consider a stratified dielectric medium where homogeneous layers with different relative permittivities are stacked in the z-direction. Layer n is located between z_n and z_{n+1} and has relative permittivity $\varepsilon_{rb,n}$. Index $n = 0$ coincides with the top half space, $z < 0$, and index $n = N_L$ with the half-space $z > z_{N_L+1}$ below all layers, an example of which is also illustrated in Figure 8.1. In layer i a three-dimensional dielectric object is contained within the simulation domain $\mathcal{D} = [-W_x, W_x] \times [-W_y, W_y] \times [z_{min}, z_{max}]$, with $z_i \leq z_{min}$ and $z_{max} \leq z_{i+1}$. This dielectric object is characterized by a relative permittivity function $\varepsilon_r(\mathbf{x})$, with $\mathbf{x} = (x, y, z)$, or more conveniently by the contrast function

$$\chi(\mathbf{x}) = \frac{\varepsilon_r(\mathbf{x})}{\varepsilon_{rb,i}} - 1, \tag{8.1}$$

which is nonzero only inside the object.

An incident electromagnetic field originates from the upper half-space at arbitrary angle and with arbitrary polarization. The electric field in presence of the multilayered background medium $\varepsilon_{rb,n}$, but in absence of the dielectric object can be readily calculated [55, 87] and is denoted as $\mathbf{E}^i(\mathbf{x})$. The dielectric object generates a scattered field $\mathbf{E}^s(\mathbf{x})$ that together with the incident field $\mathbf{E}^i(\mathbf{x})$ adds up to the total electric field $\mathbf{E}(\mathbf{x})$, i.e.

$$\mathbf{E}(\mathbf{x}) = \mathbf{E}^i(\mathbf{x}) + \mathbf{E}^s(\mathbf{x}). \tag{8.2}$$

The scattered field $\mathbf{E}^s(\mathbf{x})$ can be calculated via the multilayer Green tensor $\mathcal{G}$ through

$$\mathbf{E}^s(\mathbf{x}) = \int_{\mathcal{D}} d\mathbf{x}' \, \mathcal{G}(\mathbf{x}|\mathbf{x}') \cdot \mathbf{J}(\mathbf{x}'), \tag{8.3}$$

where the contrast current density $\mathbf{J}(\mathbf{x})$ is given by the field-material interaction

$$\mathbf{J}(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb,i}\chi(\mathbf{x})\mathbf{E}(\mathbf{x}), \tag{8.4}$$

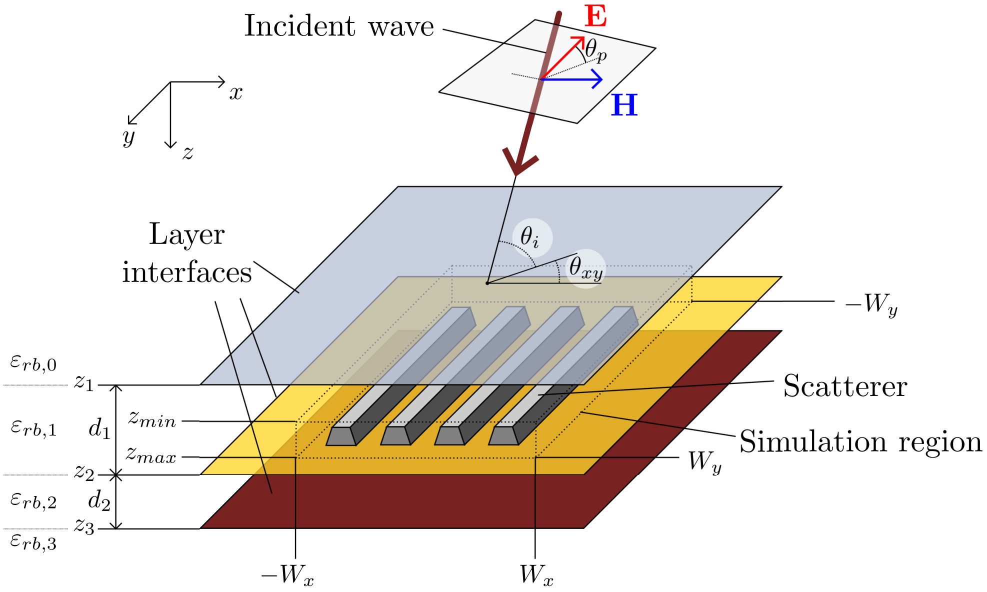


Figure 8.1: An illustration of a possible scattering setup

which is again nonzero only in the scattering object. Combining Eqs. (8.2), (8.3) and Eq. (8.4) yields the integral equation that we propose to solve

$$\mathbf{E}^i(\mathbf{x}) = \mathbf{E}(\mathbf{x}) - \int_{\mathcal{D}} d\mathbf{x}'\, \mathcal{G}(\mathbf{x}|\mathbf{x}') \cdot [j\omega\varepsilon_0\varepsilon_{rb,i}\chi(\mathbf{x}')\mathbf{E}(\mathbf{x}')]\,. \tag{8.5}$$

However, for an efficient numerical scheme several refinements have to be made.

8.4 The spectral domain representation

8.4.1 The Green function

Computing the three-dimensional integral in Eq. (8.3) involves, when implemented naively, an $O(N^2)$ matrix-vector product, with $N = N_x N_y N_z$ the total number of unknowns. This matrix-vector product can be employed in an iterative solver. Analogous to [118, 135, 147], we represent the Green function, the contrast current density $\mathbf{J}$, and scattered field $\mathbf{E}^s$ in the spectral domain in the transverse xy-plane. We denote coordinates in the transverse plane as $\mathbf{x}_T = (x, y)$, and in the spectral domain as $\mathbf{k}_T = (k_x, k_y)$. We use a Fourier transformation defined as

$$f(k_\xi) = \mathcal{F}_\xi\left[f(\xi)\right](k_\xi) = \int_{-\infty}^{\infty} d\xi\, f(\xi)e^{-jk_\xi\xi}, \tag{8.6}$$

where we distinguish functions in the spectral domain by arguments containing k_x, k_y and $\mathbf{k}_T$ and in the spatial domain by the arguments x and y and $\mathbf{x}_T$.

In the spectral domain, a spatial convolution can be executed with $O(N_x N_y)$ complexity, with N_α the number of unknowns used in direction α. The transverse convolution in Eq. (8.5) can be carried out efficiently. The remaining integration in the z-direction can be calculated in $O(N_z)$ time via the recursive algorithm proposed by Gohberg and Koltracht [72].

The multilayer Green tensor in Eq. (8.3), can be separated in a homogeneous-medium part yielding $\mathbf{E}^{s,h}$ and reflected waves moving up, $\mathbf{K}^u$, and down, $\mathbf{K}^d$. The homogeneous-medium part of the scattered field is given by

$$\mathbf{E}^{s,h}(\mathbf{k}_T, z) = \int_{z_{min}}^{z_{max}} dz' \, \mathcal{G}^h(\mathbf{k}_T, z|z') \cdot \mathbf{J}(\mathbf{k}_T, z'), \tag{8.7}$$

where the homogeneous-medium Green tensor is given in Cartesian components (x, y, z), respectively, as

$$\mathcal{G}^h(k_x, k_y, z|z') = \begin{pmatrix} \varepsilon_{rb,i}k_0^2 - k_x^2 & -k_x k_y & -k_x\gamma \\ -k_x k_y & \varepsilon_{rb,i}k_0^2 - k_y^2 & -k_y\gamma \\ -k_x\gamma & -k_y\gamma & \gamma^2 - 2\gamma\delta(z - z') \end{pmatrix} \frac{e^{-\gamma|z-z'|}}{2\gamma}. \tag{8.8}$$

Here, k_0 is the wave number $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$ and γ is defined as $\gamma = \sqrt{\mathbf{k}_T^2 - \varepsilon_{rb,i}k_0^2}$, where $\mathbf{k}_T = (k_x, k_y, 0)$ so $\mathbf{k}_T^2 = k_x^2 + k_y^2$. Note that the factor $\exp(-\gamma|z - z'|)$ propagates (and/or attenuates) the electric field over a distance $|z - z'|$, and will therefore be referred to as the propagation function.

Now the scattered field $\mathbf{E}^s$ can be found by adding reflected waves $\mathbf{K}^{u/d}$ from the layer interfaces to the homogeneous scattered field $\mathbf{E}^{s,h}$, where u and d refer to waves moving up or down respectively. Consequently, we have

$$\begin{aligned} &\mathbf{E}^s(\mathbf{k}_T, z) = \mathbf{E}^{s,h}(\mathbf{k}_T, z) + \\ &\quad \left(\mathcal{R}^{u,u}(\mathbf{k}_T, z)\mathbf{E}^{s,h}(\mathbf{k}_T, z_{min}) + \mathcal{R}^{u,d}(\mathbf{k}_T, z)\mathbf{E}^{s,h}(\mathbf{k}_T, z_{max})\right) e^{-\gamma(z - z_{min})} + \\ &\quad \left(\mathcal{R}^{d,u}(\mathbf{k}_T, z)\mathbf{E}^{s,h}(\mathbf{k}_T, z_{min}) + \mathcal{R}^{d,d}(\mathbf{k}_T, z)\mathbf{E}^{s,h}(\mathbf{k}_T, z_{max})\right) e^{-\gamma(z_{max} - z)} \end{aligned} \tag{8.9}$$

with $\mathcal{R}^{\alpha,\beta}(\mathbf{k}_T, z)$ the three-dimensional effective reflection coefficient that contains both h and e polarization, which can be calculated from the effective reflection coefficients for h polarization $r_h^{\alpha,\beta}(\mathbf{k}_T)$ [135], and for e polarization $r_e^{\alpha,\beta}(\mathbf{k}_T)$ [147] as

$$\mathcal{R}^{\alpha,\beta} = \begin{pmatrix} \frac{k_x^2 r_e^{\alpha,\beta}(\mathbf{k}_T) - k_y^2 r_h^{\alpha,\beta}(\mathbf{k}_T)}{\mathbf{k}_T^2} & \frac{k_x k_y (r_e^{\alpha,\beta}(\mathbf{k}_T) - r_h^{\alpha,\beta}(\mathbf{k}_T))}{\mathbf{k}_T^2} & 0 \\ \frac{k_y k_x (r_e^{\alpha,\beta}(\mathbf{k}_T) - r_h^{\alpha,\beta}(\mathbf{k}_T))}{\mathbf{k}_T^2} & \frac{k_y^2 r_e^{\alpha,\beta}(\mathbf{k}_T) - k_x^2 r_h^{\alpha,\beta}(\mathbf{k}_T)}{\mathbf{k}_T^2} & 0 \\ 0 & 0 & r_e^{\alpha,\beta}(\mathbf{k}_T) \end{pmatrix}. \tag{8.10}$$

This matrix projects the e and h polarized parts of the electric field onto effective transmission coefficients $r_e^{\alpha,\beta}$ and $r_h^{\alpha,\beta}$, respectively. The definition of these effective reflection coefficients is given in [135, 147], which is based on the expositions about multilayer media in [57, Chapter 4], [58, Chapter 2], [87].

Since the field-material interaction in Eq. (8.4) is calculated in the spatial domain and the Green-function operation in Eq. (8.9) in the spectral domain, we need a fast and efficient means of transforming the current density $\mathbf{J}(\mathbf{x}_T, z)$ to the spectral domain and the scattered field $\mathbf{E}^s(\mathbf{k}_T, z)$ back to the spatial domain. We propose to use a two-dimensional Gabor frame in the transverse plane, since a Gabor frame is efficient to represent the operation of Fourier transformation. It can be represented analytically by a mere transposition of the coefficient matrix in $O(N_x N_y)$ operations [118].

8.4.2 The Gabor frame

We use a Gabor frame with Gaussian window function

$$g(x, y) = 2^{\frac{1}{2}} \exp\left(-\pi \frac{x^2}{X^2} - \pi \frac{y^2}{Y^2}\right), \tag{8.11}$$

with width X in the x-direction and Y in the y-direction. This defines the oversampled two-dimensional Gabor frame as

$$g_{\mathbf{mn}}(x) = g(x - m_x \alpha X, y - m_y \alpha Y) e^{jn_x \beta K_x x + jn_y \beta K_y y}, \tag{8.12}$$

with two-dimensional indices $\mathbf{m} = (m_x, m_y)$ and $\mathbf{n} = (n_x, n_y)$. Here, the spectral spacing is $K_x = 2\pi/X$ and $K_y = 2\pi/Y$ and rational oversampling $\alpha_x \beta_x < 1$ and $1 > \alpha_y \beta_y \in \mathbb{Q}$. The number of coefficients in $\mathbf{m}$ and $\mathbf{n}$ is allowed to differ for both directions. Gabor coefficients can be calculated as

$$f_{\mathbf{mn}} = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x, y) \eta_{\mathbf{mn}}(x, y), \tag{8.13}$$

with dual frame

$$\eta_{\mathbf{mn}}(x, y) = \eta(x - m_x \alpha X) \eta(y - m_y \alpha Y) e^{jn_x \beta K_x x + jn_y \beta K_y y}. \tag{8.14}$$

There is freedom of choice for the dual window function $\eta(x)$, but we choose the dual frame function calculated via the Moore Penrose inverse [99, 100], since it is widely used and exhibits a convenient exponential decay in both the spatial and spectral domain.

We use the Fourier transform of Eq. (8.12) to discretize functions in the spectral domain. This has the advantage that the operation of Fourier transformation reduces to merely a tranposition of coefficients. Details on operations such as Fourier transformation and multiplication of Gabor-represented functions can be found in [118] for one dimension and the generalization to two dimensions is straightforward. The one-dimensional operations can be done for both the x and y direction consecutively.

8.5 A complex-plane deformation of the integration manifold

In the z-direction, the integration with the Green tensor in Eq. (8.7) is discretized completely in the spatial domain. Since it was shown that a piecewise-linear discretization in the z-direction is effective [118, 135, 147], we propose to use it here as well. In the z-direction, the basis functions are then defined as

$$\Lambda_\ell(z) = \begin{cases} 1 - \frac{|z-\ell\Delta_z - z_{min}|}{\Delta_z} & \text{if} \quad |z - \ell\Delta_z - z_{min}| < \Delta_z \\ 0 & \text{if} \quad |z - \ell\Delta_z - z_{min}| > \Delta_z \end{cases}, \tag{8.15}$$

with Δ_z the discretization step in the z-direction.

For the discretization in the xy plane, a method similar to the two-dimensional cases in [135, 147] is proposed. The Green function contains poles due to the effective reflection coefficients and many oscillations along the branch cuts may occur. Both these poles and oscillations cannot be represented efficiently in a Gabor frame representation. In the two-dimensional case, these problems can be circumvented by representing the Green-function in the transverse direction in Eq. (8.9) on a path in the spectral complex plane. For three-dimensional problems, this path can be generalized to a two-dimensional integration manifold in the transverse $\mathbf{k}_T$ coordinates on which the transformation back to the spatial domain takes place. In the k_x-direction, the complex spectral path is defined by the function $\tau_x(k_x)$, with $k_x \in \mathbb{R}$ and $\tau_x \in \mathbb{C}$ as

$$\tau_x(k_x) \in \begin{cases} k_x - jA_x & \text{if } k_x < -A_x \\ (1+j)k_x & \text{if } -A_x \leq k_x < A_x \\ k_x + jA_x & \text{if } k_x > A_x. \end{cases} \tag{8.16}$$

and a similar definition for $\tau_y(k_y)$ with A_y defining the imaginary displacement along the k_y-direction. Here, A_x and A_y are constants that can be chosen individually. Numerical experiments show that a choice such that A_xW_x and A_yW_y are in the range $2 \ldots 5$, yields optimal accuracy, with W_x and W_y as in Figure 8.1. With the coordinate change from $\mathbf{k}_T$ to τ_T, Eqs (8.7) and (8.9) contain smooth functions and these can be used in combination with the Gabor-frame discretization.

This complex spectral manifold divides the complex $\mathbf{k}_T$ domain into nine regions as depicted in Figure 8.2. All functions in the spectral domain will be represented on this τ_T manifold. With the aid of Jordan's lemma, the Fourier transformation to the spatial domain can be carried out over the τ_T manifold. Closing the contour at $\mathbf{k}_T \to \infty$ is not needed, since the representation using Gabor frames converges to zero rapidly outside the simulation domain.

8.5.1 Discretization in regions of Type 1

Most information is contained in regions of Type 1, since A_x and A_y are relatively small compared to the complete spectral range to be discretized. The contrast current density is transformed to the complex spectral integration manifold via

$$\mathbf{J}(k_x + jA_x, k_y + jA_y, z) = \mathcal{F}_{\mathbf{x}_T}[\mathbf{J}(\mathbf{x}_T, z)e^{-xA_x - yA_y}](\mathbf{k}_T), \tag{8.17}$$

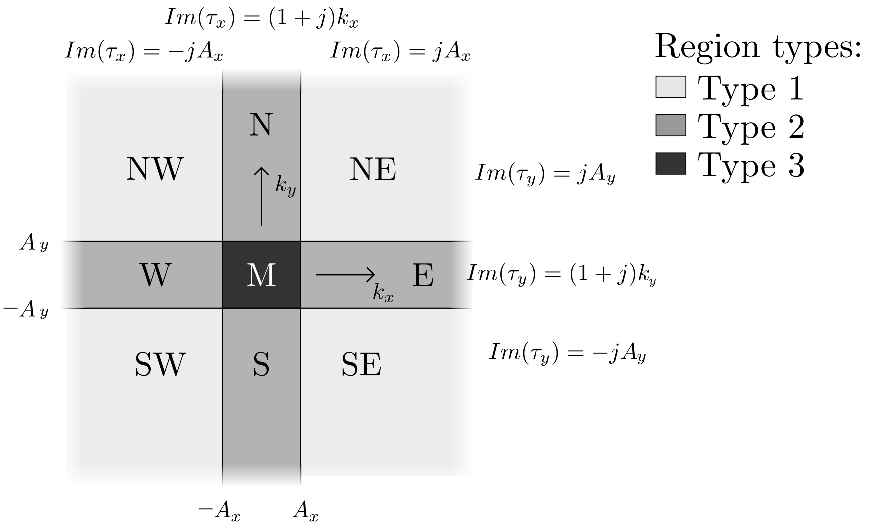


Figure 8.2: The complex-plane integration domain in the spectral domain consisting of nine regions, of three types.

for the northeast (NE) quadrant, i.e. $k_x \geq A_x \wedge k_y \geq A_y$, and similarly for the other regions of Type 1. Analogously, the transformation of the scattered field back to the spatial domain is obtained as

$$\mathbf{E}^{s,\mathrm{NE}}(\mathbf{x}_T, z) = e^{-xA_x - yA_y}\mathcal{F}^{-1}_{\mathbf{k}_T}[c_{\mathrm{NE}}(\mathbf{k}_T)\mathbf{E}^{s,\mathrm{NE}}(k_x + jA_x, k_y + jA_y, z)](\mathbf{x}_T), \tag{8.18}$$

with the cut-off function $c_{\mathrm{NE}}(\mathbf{k}_T)$ equalling 1 on the NE region and zero elsewhere. The Fourier transformation can be performed in $O(N_xN_y)$ operations and the operation of multiplication in $O(N_xN_y \log N_xN_y)$ operations, for functions represented by N_xN_y Gabor coefficients. Therefore, the total of these operations allows for an $O(N_xN_y \log N_xN_y)$ computational complexity.

All this means that the scattered electric field $\mathbf{E}^s$ is represented by a five-dimensional array of coefficients $\mathbf{E}^{s,\mathrm{NE}}_{\mathbf{m},\mathbf{n},l}$, with m_x, n_x and m_y, n_y corresponding to the Gabor frame on the coordinates, $k_x + jA_x$ and $k_y + jA_y$ respectively. The ℓ index corresponds to a piecewise-linear (PWL) representation in the z-direction. The scattered electric field in region NE is then approximated as

$$\mathbf{E}^{s,\mathrm{NE}}(k_x + jA_x, k_y + jA_y, z) \approx \sum_{\mathbf{m},\mathbf{n}} \sum_{\ell=1}^{N_z} g_{\mathbf{m},\mathbf{n}}(k_x, k_y)\Lambda_\ell(z)\mathbf{E}^s_{\mathbf{m},\mathbf{n},\ell}. \tag{8.19}$$

The Green function consists of several parts, some of which are depending on the complex propagation constant $\gamma(\mathbf{k}_T) = \sqrt{-\varepsilon_{rb,i}k_0^2 + \mathbf{k}_T^2}$. On the real k_xk_y-plane $\gamma(\mathbf{k}_T)$ touches, but does not cross, two branch cuts at $\mathbf{k}_T = (0, 0)$ in the case of lossless media. For lossy media the branchcuts are located at some distance from the origin. For both cases, the τ path passes just in between these two branchcuts. However, when a Type-1 region such as the NE-region is continued to the complete $(k_x + jA_x, k_y + jA_y)$ plane, a branch cut is crossed just outside the NE region, as illustrated in Figure 8.3. The branch cut is located on a straight line through $\tau_T = (0 + jA_x, 0 + jA_y)$ and the direction of the line depends on the choice of A_x and A_y. The continuous nature of a Gabor-frame representation does not allow for an abrupt stop of the discretization domain at the borders of a Type-1 region. Therefore, such a Gabor-frame representation of the Green function exhibits significant Gibbs ringing from the branch cut that spreads into the Type-1 regions. For a two-dimensional case, this is described in [135], where a linear continuation of the Green function is proposed that suppresses strong Gibbs ringing.

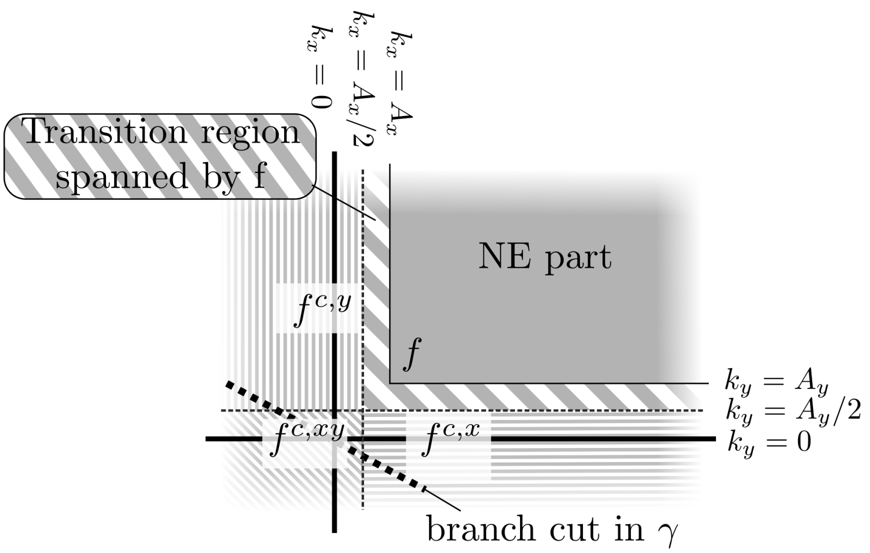


Figure 8.3: A function f represented in the NE regions on $\tau(k_x, k_y) = (k_x + jA_x, k_y + jA_y)$ that depends on $\gamma(\tau(\mathbf{k}_T))$ contains a branchcut on a straight line through the $k_x < 0 \vee k_y < 0$ region. On the regions indicated with solid and striped grey the original function f is discretized and on regions indicated by fine lines the continuations $f^{c,x}$, $f^{c,xy}$ and $f^{c,y}$ are discretized, the discontinuity of the branch cut is therefore avoided by the continuous functions.

In three dimensions, this issue can also be resolved by making a first-order continuation of the functions to eliminate the branch cut. Since the branch cut can be located close to the $k_x = 0$ or $k_y = 0$ axes, and the function values are needed at $k_x > A_x$ and $k_y > A_y$, we start the continuation of the functions in the middle at $k_x^c = A_x/2$ and $k_y^c = A_y/2$. Then the Gibbs phenomenon from the discontinuous second derivative will be at a short distance from $k_x = A_x$ and $k_y = A_y$, where Region NE begins. For the continuation of a function $f(k_x, k_y)$ along the k_y-axis we choose

$$f^{c,x}(\mathbf{k}_T) = \left[f(\frac{A_x}{2}, k_y) + (k_x - \frac{A_x}{2})\partial_{k_x} f(\frac{A_x}{2}, k_y) \right] e^{-\alpha(k_x - A_x/2)^2}, \tag{8.20}$$

for $k_x < A_x/2$ and $k_y > A_y/2$. Similarly $f^{c,y}(k_x, k_y)$ can be constructed for $k_x > A_x/2$ and $k_y < A_y/2$, which is illustrated in Figure 8.3. The Gaussian factor is added to make the continuation decay to zero slowly. The third part, $k_x < A_x/2$, $k_y < A_y/2$ is a continuation of $f^{c,y}$ into

$$f^{c,xy}(\mathbf{k}_T) = \left[f^{c,y}(\frac{A_x}{2}, k_y) + (k_x - \frac{A_x}{2})\partial_{k_x} f^{c,y}(\frac{A_x}{2}, k_y) \right] e^{-\alpha(k_y - A_y/2)^2}. \tag{8.21}$$

Note that this expression equals the expression obtained from continuing $f^{c,x}$ onto this domain. The derivative of the function f is calculated using a forward finite-difference method, with a difference of $10^{-4}A_x$ or $10^{-4}A_y$ for the x and y direction, respectively. For most functions, $\alpha = \min(X^2, Y^2)$ is a good choice. However, for one part of the Green function, notably the propagation function $e^{-\gamma\Delta_z}$, care has to be taken that its absolute value does not exceed one in the continuation. By increasing the value of α, this condition can always be satisfied. More details can be found in [135].

A general remark about the importance of this continuation is appropriate. In principle, the Gibbs phenomenon in a Gabor frame representation is not of much significance, unless two functions with discontinuities at the same position are multiplied. The Li rules [81] state that when two functions with spatial discontinuities at the same position are multiplied to form a convolution in the spectral domain, the convergence of this convolution is poor. The Li rules also apply to Gabor frames [147] and since the spatial and spectral domain are both represented by a Gabor frame, a spatial version of the Li rules is also applicable to the Gabor frame. These spatial Li rules state that when two spectral functions with discontinuities are multiplied in a Gabor-frame representation, a poor convergence is observed. Now when the NE region of the electric field with its branch cut is multiplied by the cut-off function $c_{\mathrm{NE}}(\mathbf{k}_T)$ in Eq. (8.18), which is discontinuous at $k_x = A_x$ and at $k_y = A_y$, functions are multiplied for which the locations of discontinuities almost touch each other. This leads to near-violation of the spatial Li rules. Since the discontinuities are not exactly at the same location, a high sampling would in principle solve this issue. However, this would require an excessive sample density that is avoided by the continuation of the Green function parts proposed in Eqs (8.20) and (8.21).

8.5.2 Discretization in regions of Type 2

First we will approximate the contrast current density in k_x around $k_x = 0$ with a Taylor expansion that is found through a Vandermonde matrix. This Taylor expansion is then applied to find corresponding values of the contrast current density on the line $Im(\tau_x) = Re(\tau_x)$, on which a PWL basis is used as a discretization. This PWL basis consists of $2N_s + 1$ basis functions, sampled at $p(1+j)A/N_s$, with $p \in \{-N_s, \ldots, N_s\}$. Afterwards, we give a means to directly Fourier transform from the discretized N region to spatial-domain Gabor coefficients. We will only consider the northern (N) region of the complex spectral integration manifold since the E, W, and S region follow by analogy.

For the calculation of the current density in the N region, function values of $\mathbf{J}^{\mathrm{N}}$ are available at the lines $\tau_x = k_x \pm jA_x$, which were calculated via the Gabor representation in the NE and NW regions. The analyticity of $\mathbf{J}$ allows to produce a Taylor expansion of $\mathbf{J}$ around $k_x = 0$ from values at the lines $Im(\tau_x) = \pm A_x$. Afterwards, this Taylor expansion is used to calculate values of the contrast current density at the line $Re(\tau_x) = Im(\tau_x)$, where they are needed for discretization in the N region, as is shown in Figure 8.4. Close to $k_x = 0$, $\mathbf{J}^{\mathrm{N}}$ can be approximated as

$$\mathbf{J}^{\mathrm{N}}(k_x, k_y, z) \approx \sum_{n=0}^{4N_v+1} \frac{k_x^n}{n!} \mathbf{a}_n(k_y, z), \tag{8.22}$$

where $\mathbf{a}_n(k_y, z) = \partial_{k_x}^n \mathbf{J}^{\mathrm{N}}(k_x, k_y, z)$, and $4N_v + 2$ is the total number of terms in this Taylor expansion. Values for $\mathbf{J}^{\mathrm{N}}(k_x \pm jA_x, k_y)$ can be obtained from the results for the NE and NW regions, by using a fast Gabor transformation [118, 100] that yields values at $k_x = n\Delta_{k_x} \pm jA_x$ for $n \in \mathbb{Z}$, with Δ_{k_x} the spectral sample spacing corresponding to the Gabor frame. Values for $\mathbf{a}_n$ can be found by solving a small Vandermonde system [38, Chapter 2.8]. By constructing the vector $\underline{k}$ of k_x values as $\underline{k} = (-N_v\Delta_{k_x} - jA_x, \ldots, N_v\Delta_{k_x} - jA_x, -N_v\Delta_{k_x} + jA_x, \ldots, N_v\Delta_{k_x} + jA_x)^T$, this Vandermonde system can be written as a matrix equation $\underline{\underline{K}} \cdot \underline{\mathbf{a}} = \underline{\mathbf{j}}(k_y, z) = \mathbf{J}^{\mathrm{N}}(\underline{k}, k_y, z)$. The element $\underline{\underline{K}}_{mn}$ of the m'th row and n'th column of matrix $\underline{\underline{K}}$ is given by $\underline{\underline{K}}_{mn} = (\underline{k}_m)^n$, the n'th power of element m in $\underline{k}$. We solve this system by using the inverse of $\underline{\underline{K}}$, i.e.

$$\underline{\mathbf{a}}(k_y, z) = \underline{\underline{K}}^{-1} \cdot \underline{\mathbf{j}}(k_y, z). \tag{8.23}$$

Now that it is possible to express the Taylor coefficients $\mathbf{a}_n$ in terms of the $2N_\nu + 1$ samples on the NW-region, i.e. $\mathbf{J}(k_x - jA_x, k_y + jA_y, z)$, and the $2N_\nu + 1$ samples in the NE-region, i.e. $\mathbf{J}(k_x + jA_x, k_y + jA_y, z)$, they can be used to evaluate the Taylor expansion in Eq. (8.22) on the N-region, where $Im(\tau_x) = Re(\tau_x)$. We will write this as a matrix-vector product using the matrix $\underline{\underline{T}}$. The matrix $\underline{\underline{T}}$ transforms from a Taylor series to an equidistant sampling on the line $[-A_x - jA_x, A_x + jA_x]$. The elements are $\underline{\underline{T}}_{pm} = ((1+j)pA_x/N_s)^m$, where $p \in \{-N_s, \ldots, N_s\}$ and where $m \in \{0, \ldots, 4N_\nu + 1\}$, i.e.

$$\mathbf{J}_p^{\mathrm{N}}(k_y, z) = [\underline{\underline{T}} \cdot \underline{\mathbf{a}}(k_y, z)]_p = [\underline{\underline{T}} \cdot \underline{\underline{K}}^{-1} \cdot \underline{\mathbf{j}}(k_y, z)]_p. \tag{8.24}$$

We use the array of numbers $\mathbf{J}_{p,m_y,n_y,\ell}^{\mathrm{N}}$ to represent the current in the N region of the complex integration domain. Index $p \in \{-N_s, N_s\}$ points to the set of piecewise-linear basis functions that are used in the k_x-direction on the line $\tau_x((1+j)pA/N_s)$. In the k_y-direction we use a Gabor frame, denoted here by indices m_y and n_y. This means that the y dependence in Eq. (8.24) is replaced by this set of Gabor indices. Again, a set of N_z PWL functions is used in the z-direction denoted by the index ℓ.

Having dealt with the transformation to the N region, we will now deal with the transformation from the N region back to its spatial-domain counterpart. After multiplication of the contrast current density $\mathbf{J}_{p,m_y,n_y,\ell}^{N}$ with the Green function (see Section 8.4.1), the contribution of the North part of the scattered electric field yields $\mathbf{E}_{p,m_y,n_y,\ell}^{s,N}$. From this array we can make an approximation on the N region of the scattered electric field

$$\mathbf{E}^{s,\mathrm{N}}(k_x + jk_x, k_y + jA_y, z) \approx \sum_{p=-N_s}^{N_s} \Lambda_{s_x,p}(k_x) \sum_{n_y,m_y} g_{m_y,n_y}(k_y) \sum_{\ell=1}^{N_z} \Lambda_{z,\ell}(z) \mathbf{E}_{p,m_y,n_y,\ell}^{s,N}. \tag{8.25}$$

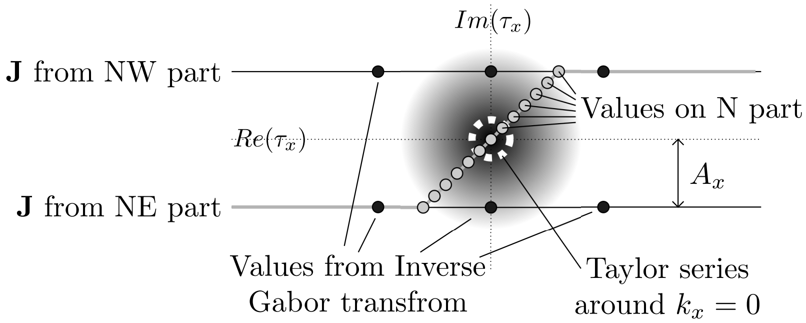


Figure 8.4: Illustration of the expansion for $N_\nu = 1$ of the six known values (dark grey circles) from the NE and NW regions to values for the N region (light-gray circles).

Here $\Lambda_{s_x,p}$ are piecewise-linear (PWL) basis functions

$$\Lambda_{s_x,p}(k) = \begin{cases} 1 - \frac{|k-p\Delta_{k_x}|}{\Delta_{k_x}} & \text{if} \quad |k - p\Delta_{k_x}| < \Delta_{k_x} \\ 0 & \text{if} \quad |k - p\Delta_{k_x}| > \Delta_{k_x} \end{cases}, \tag{8.26}$$

with width $\Delta_{k_x} = A_x/N_s$ in the k_x-direction. To transform the N region of the scattered electric field $\mathbf{E}^{s,\mathrm{N}}_{p,m_y,n_y,\ell}$ back to the spatial domain where it is discretized in the Gabor frame with coefficients $\mathbf{E}^{s,N}_{\mathbf{m},\mathbf{n},l}$, we use the Fourier transforms of the PWL functions in Eqs. (8.26) and (8.25), i.e.

$$I^{\mathrm{N}}_p(x) = \int_{-A_x}^{A_x} dk_x\, \tau'(k_x)\, \Lambda_p(k_x) e^{j\tau_x(k_x)x}. \tag{8.27}$$

Since the x direction is discretized by using Gabor coefficients in the spatial domain, the x-dependence of this function $I^{\mathrm{N}}_p(x)$ must be transformed into Gabor coefficients $I^{\mathrm{N}}_{p,m_x,n_x}$. These Gabor coefficients are calculated during initialization of the algorithm via e.g. Eq. (8.13) or a fast Gabor transformation. Now the contribution of the N region to the scattered electric field in the spatial domain is given by

$$\mathbf{E}^s_{\mathbf{m},\mathbf{n},l} = \cdots + \sum_{p=-N_s}^{N_s} \mathbf{E}^{s,N}_{p,m_y,n_y,\ell} I^N_{p,m_x,n_x} \tag{8.28}$$

where the dots indicate the contributions from the other eight regions to the scattered field.

Similar to regions of Type 1, some parts of the Green function are discretized by using a continuation such as in Eq. (8.20), to avoid a branch cut. For example, for the N region the continuation is only needed in the y-direction, since a Gabor frame is employed in this direction only and a PWL discretization does not suffer from Gibbs ringing. The construction for a one-dimensional continuation is described in more detail in [135].

8.5.3 Discretization in the region of Type 3

For the middle (M) region, a two-dimensional version of the construction for the N region is used. Since the generalization is fairly straightforward, we will not write it down explicitly.

The only difference here is that we use a total number of $2N_m + 1$ PWL functions per direction. We use a different number of PWL functions in this region since, depending on the simulation parameters, the accuracy can depend significantly on the choice of N_m. Since the middle part contains information of waves with small $\mathbf{k}_T$, it contains information about waves traveling almost parallel to the z-direction. Especially for scatterers that are larger in the z-direction, a larger N_m is required.

An important remark on the use of Vandermonde matrices is that they are generally ill-conditioned when a uniform sampling is used, such as is the case in the NE and NW regions. In principle, this could lead to a poor conditioning of the $\underline{\underline{K}}$ matrix and therefore to an unstable inversion when the matrix is increased in size. However, the amount of information on the interval $\mathbf{k}_T \in [-A_x, A_x] \times [-A_y, A_y]$ is so small that large matrices are not needed.

There are two reasons that a relatively large number of PWL basis functions (typically $N_m > 10$ and $N_s > 10$) is needed in regions of Type 2 and 3. The first is that a PWL basis is relatively inefficient compared to a Gabor frame. For the second reason we have to look at both the spatial and the spectral domain. Since the contrast current density $\mathbf{J}$ is confined to a finite region only, its Fourier transform is fairly smooth. However, the scattered electric field $\mathbf{E}^s$ is not confined to the simulation domain, and therefore its Fourier transform is much less smooth. On the Type-1 and Type-2 regions this lack of smoothness is compensated by a representation in terms of complex spectral coordinates τ, where the Green function is much smoother. However, the Type-3 region is not shifted as far into the complex plane as the Type-1 and Type-2 regions, and therefore the Green function is less smooth in this region. Since the Green function is implemented recursively, for intermediate results, i.e. the scattered field in between z_{min} and z_{max}, this lack of smoothness should be represented accurately. Afterwards, when the transformation to the spatial domain is performed, this roughness on the M region corresponds to contributions outside the simulation domain. However, ignoring the roughness is not an option since it leads to accumulating errors in the recursive handling of the Green function. This is especially important when $z_{max} - z_{min}$ is large compared to the wavelength.

8.5.4 Correspondence between simulation parameters and accuracy

Since there are many simulation parameters, it is not trivial to find a combination of values for these parameters that produces both a good accuracy and a short computation time. This list is intended to clarify which simulation parameters influence which part of the algorithm. This list is intended as a general guideline for optimal results.

1. Start with a Gabor frame with $X = Y = \lambda$, the wavelength of the light source, and $\alpha = \beta = \sqrt{2/3}$.

2. Choose $m_x > 3 + W_x/\alpha X$ and similarly $m_y > 3 + W_y/\alpha Y$. Choose $n_x = n_y > 5 \max_{x \in \mathcal{D}}(1 + \chi(x))$, which guarantees at least 11 unknowns per wavelength per direction. Test whether a function (e.g. a Gaussian with width X), can be represented

with the required accuracy over the entire simulation domain $\mathcal{D}$. When the accuracy is too low everywhere, increase $\mathbf{n}$, when the accuracy is too low at the boundary of $\mathcal{D}$ only, increase $\mathbf{m}$.

3. Start at $A_x = 3/W_x$ and $A_y = 3/W_y$, $N_v = 1$, $N_s = 10$ and $N_m = 10$. Test whether a set of spatially and spectrally localized functions (e.g. modulated Gaussians that are shifted along the entire simulation domain) can be transformed to the complex spectral integration manifold and back again with the required accuracy. Note that the exponential function in Eq. (8.17) reduces the accuracy of the Gabor frame. Therefore, a simultaneous decrease of A_x and increase of $\mathbf{m}$ improves the accuracy in the transformation between the spatial and spectral domain. Especially when a high accuracy is needed, N_ν may have to be increased when the error in the N, E, S, W and M regions is too large. Also an increase in N_s and N_m can be considered when the error in the PWL interpolation is found to be too large.

4. The Green function (Eq. (8.8)) contains a factor γ^{-1}, that has a strongly peaked behavior around $|\mathbf{k}_T| = \sqrt{\varepsilon_{rb,i}}k_0$. Test whether the function γ^{-1} is represented accurately enough by the Gabor frame with the current parameters. Otherwise A_x can be increased (which decreases the accuracy in the previous step) or $\mathbf{m}$ can be increased (which increases the computation time, but does not affect the accuracy in the previous step).

5. Especially for large $z_{max} - z_{min}$, a lot of information is stored in the M region, which contains information about waves traveling in a narrow cone around the z axis. The main culprit here is the function $e^{-\gamma(z_{max}-z_{min})}$ in Eq. (8.8). Choose N_m such that this function can be well approximated in the M region.

6. Another simulation can be run with lower $\mathbf{n}$ values. When both results agree well, convergence in the spectral range $\mathbf{n}$ has been reached, otherwise $\mathbf{n}$ should be increased for higher accuracy.

Many of the steps in this list can be automated. It is not necessary to manually carry out this procedure for every simulation.

8.6 Efficient field-material interaction

Formulating the field-material interaction as proposed in Eq. (8.4) yields poor convergence since it violates the Li rules [81]. We propose to use a normal-vector field approach[2][84, 91]. In [147] it is shown that when the Li rules are satisfied, good convergence is reached in a continuous spectral discretization in a formulation similar to the RCWA formulation by Granet [82]. We follow the same approach as [84, 70, 71, 91, 90] in constructing normal-vector fields and the following is intended as a short summary of that method.

[2] A more detailed description and an example can be found in Section 3.1

When the permittivity is discontinuous at a material interface, it is observed that the electric field **E** normal to the surface is discontinuous, but the electric flux density **D** normal to such a surface is continuous. Therefore, in the field-material interaction in Eq. (8.4), both χ and the normal component of **E** are discontinuous and multiplication of two discontinuous functions represented by Gabor coefficients shows poor convergence [147], since it violates the Li rules [81]. An auxiliary field **F** is introduced that is composed of **D** in the direction normal to every surface of discontinuous χ and **E** parallel to each of those surfaces. Since this fixes the choice of **F** only at the boundaries of dielectric objects, there is much freedom in choosing it away from the interfaces. Normal-vector fields [84, 154, 155, 91] can be a good tool to systematically construct an auxiliary field **F**.

Since we use Gabor coefficients only in the transverse plane, we apply the normal-vector field formulation only in the transverse plane. For objects with interfaces that are not aligned with the z or transverse plane, a staircasing approximation is needed. When $\mathbf{N}_T(\mathbf{x})$ is a vector field of unit amplitude that is directed normal to the transverse part of all discontinuous surfaces in χ and when $\alpha(\mathbf{x})$ is a scalar function that equals one at these discontinuities, these functions can be used to construct the desired auxiliary field **F** as

$$\mathbf{F}(\mathbf{x}) = \mathbf{E}(\mathbf{x}) + \mathbf{N}_T(\mathbf{x}) \left[\left(\frac{\alpha(\mathbf{x})}{\varepsilon_0 \varepsilon_{rb,i}} \mathbf{D}(\mathbf{x}) - \mathbf{E}(\mathbf{x}) \right) \cdot \mathbf{N}_T(\mathbf{x}) \right]. \tag{8.29}$$

The field-material interaction in Eq. (8.4) can be re-written as

$$\mathbf{J}(\mathbf{x}_T, z) = \, [\chi C_\varepsilon](\mathbf{x}_T, z)\mathbf{F}(\mathbf{x}_T, z), \tag{8.30}$$

and the electric field can be recovered from

$$\mathbf{E}(\mathbf{x}_T, z) = [C_\varepsilon](\mathbf{x}_T, z)\mathbf{F}(\mathbf{x}_T, z), \tag{8.31}$$

where the Cartesian component i of the electric field due to the Cartesian component j of auxiliary field F is calculated by employing the operator C_ε defined as

$$[C_\varepsilon(\mathbf{x})]_{ij} = \delta_{ij} + \mathbf{N}_{T,i}(\mathbf{x})\mathbf{N}_{T,j}(\mathbf{x}) \left[\frac{1}{\alpha(\mathbf{x})(1+\chi(\mathbf{x}))} - 1 \right] \tag{8.32}$$

with δ_{ij} denoting the Kronecker delta and similarly

$$[\,\chi C_\varepsilon(\mathbf{x})]_{ij} = j\omega\varepsilon_0\varepsilon_{rb,i}\chi(\mathbf{x})[C_\varepsilon(\mathbf{x})]_{ij}. \tag{8.33}$$

More details and examples of this construction can be found in [91].

8.7 Numerical results

We have chosen three test cases to validate the present algorithm. As a reference to validate our results we use the commercial FEM code JCMWave [134]. Our goal is to achieve an accuracy of 10^{-3}, which is sufficient in our applications, e.g. due to measurement noise or fabrication tolerances. To achieve high accuracy in the validation, we use

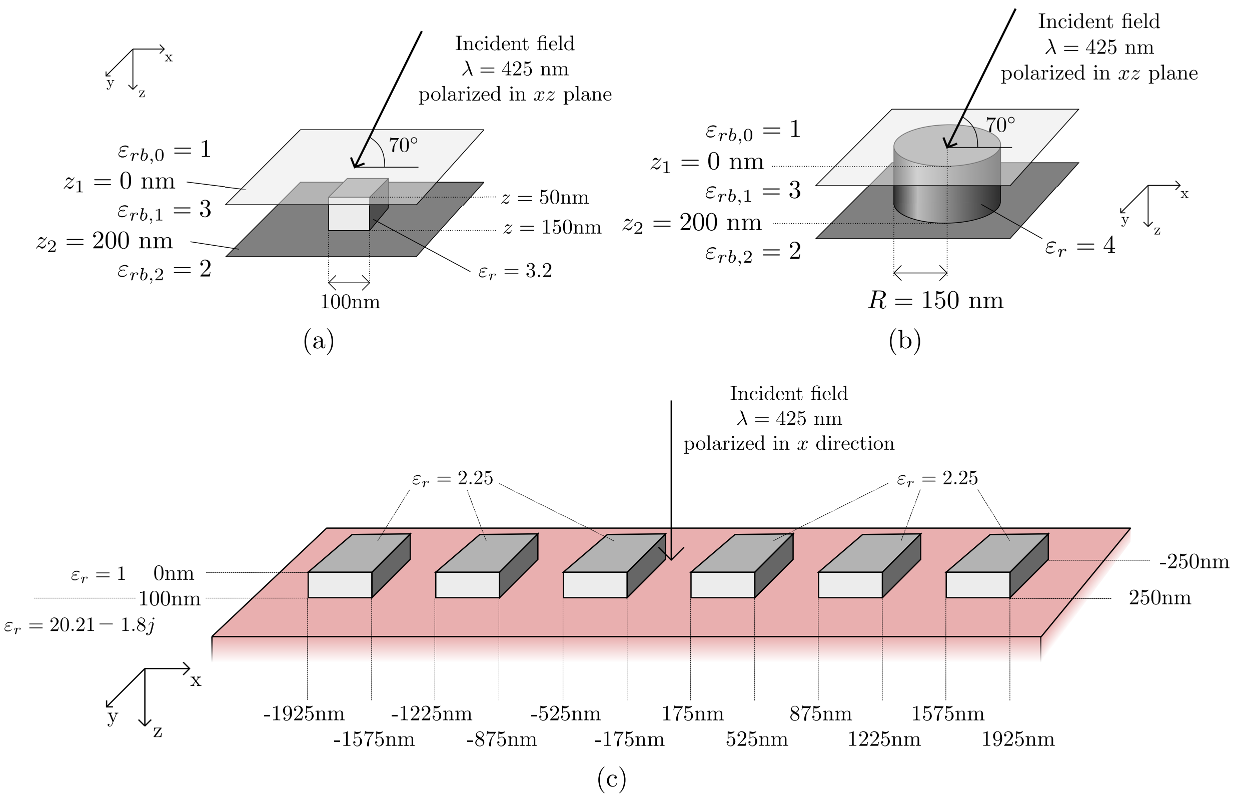
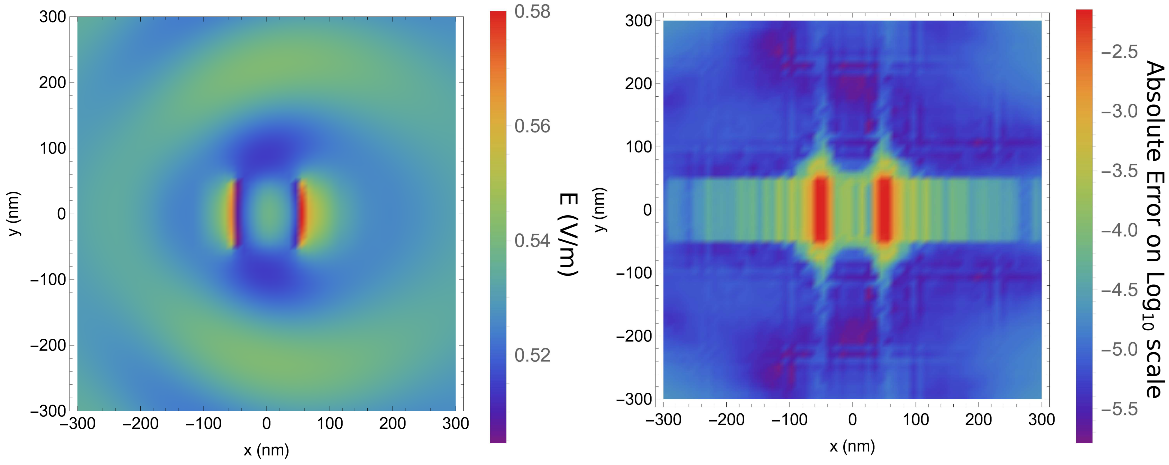


Figure 8.5: (a) The scattering setup for a small, low-contrast 100 nm cube embedded in a multilayered medium. (b) A cylinder, embedded in the same multilayered medium. (c) A finite grating consisting of six repeating blocks located on top of a substrate.

a relatively small, low-contrast scatterer in the first test case. A small dielectric cube is embedded in a dielectric medium as shown in Figure 8.5(a), together with the remaining details of the setup. The incident wave is characterized by Cartesian wavenumber $\mathbf{k} = (-k_0 \sin(70°), 0, k_0 \cos(70°))$, with the electric field polarized in the xz-plane and with unit amplitude.

We choose a Gabor frame with $X = Y = 80$ nm, $\alpha = \beta = \sqrt{2/3}$. For the highest accuracy we use a Gabor frame, Eq. (8.12), restricted to $m_x, m_y \in \{-7, \ldots, 7\}$ and $n_x, n_y \in \{-10, \ldots, 10\}$, which amounts to one basis function per 3.1 nm. In the z-direction we use a step size of 2.5 nm. For the sampling of the regions of Type 2 and 3 we use $N_\nu = 2$, $N_s = N_m = 15$.

With these simulation parameters, the simulation domain in the xy-plane extends over a larger region than the scatterer itself, as is visible in Figure 8.6(a). In this figure, the norm of $\mathbf{E}$ is shown on the plane $z = 100$ nm. In Figure 8.6(b) the absolute difference between results from the present algorithm and the JCMWave reference are shown. Over large regions of the simulation domain the absolute difference is smaller than 10^{-5}. From $y = -50$ nm to $y = 50$ nm and close to the edges of the cube agreement between both simulations is not as good as on the rest of the simulation domain. This is caused by Gibbs-ringing at the discontinuities in the electric field, especially in the x-component.



(a)

(b)

Figure 8.6: The electric field at the $z = 100$ nm plane for the scattering case in Fig. 8.5(a). In Figure (a) $|\mathbf{E}|$ is plotted for an incident plane wave with unit amplitude and in Figure (b) this is compared to the results obtained using JCMWave.

Since this Gibbs ringing has a very small spatial period, it does not radiate into the far field. Because the far field is the most interesting for the application, we use the far field as a reference for the accuracy of the method. As can be observed in Figure 8.7, the error in the far field is much lower than in the near field, because the abscence of the Gibbs ringing. The average relative difference with an $\mathcal{L}^2$-norm in the far field data equals $4 \cdot 10^{-5}$. Clearly, the far-field results agree much better than the near-field results. The small size of the scatterer and its low contrast results in a far field pattern that does not vary much with the angle. This example is therefore somewhat uninteresting, however, it has the advantage that the FEM reference could achieve a high accuracy in a multilayered scattering problem.

In Figure 8.8, we show how both the accuracy and the computation time scale with the number of unknowns used in the calculation. The horizontal axis in Figure 8.8(a) contains the sample density, which was lowered by decreasing the range of the $\mathbf{n}$-index in Eq. (8.12), where $n_x, n_y \in \{-r, \ldots, r\}$ from $r = 10$ down to $r = 1$. This corresponds to a sample density $1/\Delta_x = 1/\Delta_y = (2r + 1)/\sqrt{\alpha/\beta}X$. The other simulation parameters were kept constant throughout this sweep. The results suggest that the computation time scales as $O(1/\Delta_x\Delta_y) = O(N_xN_y)$. Since FFTs are used, we expect an $O(N_xN_y \log(N_xN_y))$ behaviour, but the logarithms are apparently negligible compared to other parts of the algorithm at this simulation size. Figure 8.8(b) shows a clear $O(N_z)$ behaviour, which is expected from Gohberg and Koltracht's recursion [72].

The second example for which we provide computational data consists of a circular dielectric cylinder embedded in a multilayered medium as is described in Figure 8.5(b). In

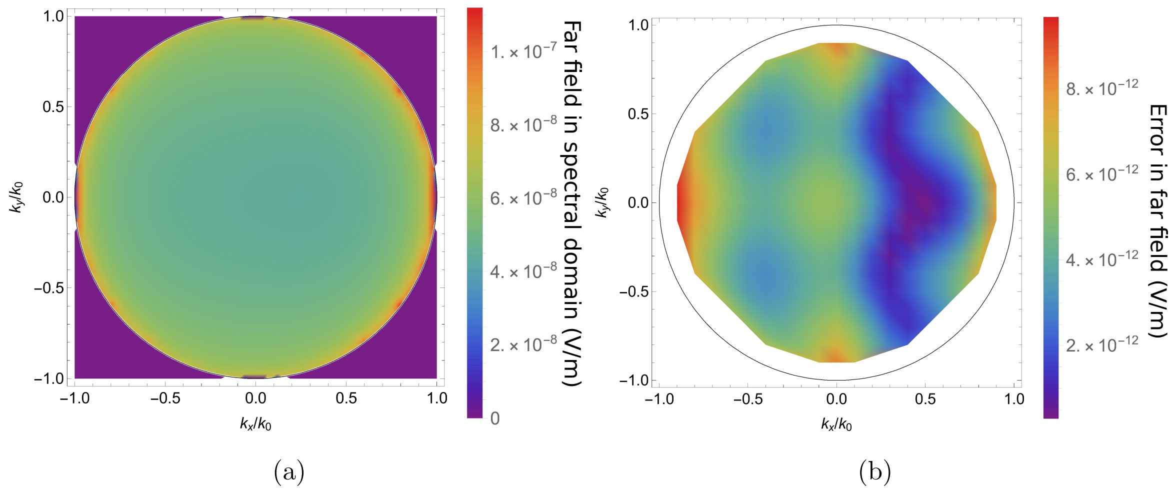
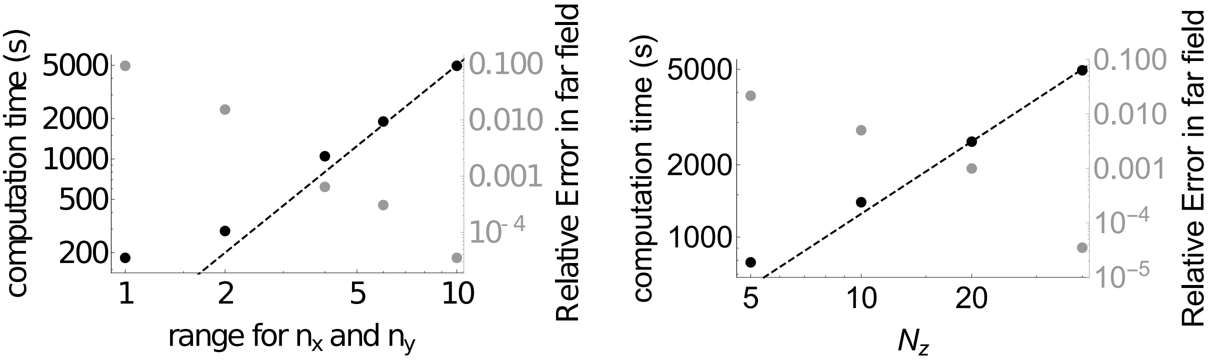


Figure 8.7: The far field for the case in Fig. 8.5(a) as a function of the transverse wavenumber $\mathbf{k}_T/k_0$, scattered back into the half-space $z < 0$. In (a) the magnitude $|\mathbf{E}^s|$ of the scattered electric field is shown. In (b) the difference between a JCMWave validation run and the present algorithm is shown. An average relative error of $4 \cdot 10^{-5}$ was observed. Since an interpolation of the reference data is used that is not accurate at the edge of the radiation circle, the far field data is truncated for large $\mathbf{k}_T$.

Figure 8.9, the electric field is shown at $z = 10$ nm for $X = Y = 100$ nm, $\alpha = \beta = \sqrt{2/3}$ and $m_x, m_y \in \{-4, \ldots, 4\}$ and $n_x, n_y \in \{-7, \ldots, 7\}$, which amounts to one basis function per 5.1 nm. In the z-direction, 41 basis functions are used with step size $\Delta_z = 2.5$ nm. For the sampling of the regions of Type 2 and 3 we use $N_\nu = 2$, $N_s = N_m = 15$. From Figure 8.9(b) it is clear that the difference between the result from the present algorithm and the JCMWave reference is somewhat larger than for the previous case. However, this is due to a less accurate reference that was calculated with a lower order p-refinement.

In Figure 8.10, the absolute value of the far field reflected into the upper half-space is plotted. The relative error for the simulation in the far field is $2.8 \cdot 10^{-3}$, which is significantly larger than for the cubic scatterer, because of the reference, with lower accuracy.

We have calculated the far field with a finer discretization, to show the convergence of the algorithm. We emphasize that the present algorithm was not developed for small computational domains. However, for a small computational domain a more accurate validation result was feasible than for a very large computational domain. The reason that the present algorithm is relatively slow for small simulation domains is that there exists a lower limit on the number of unknowns in the x and y-direction, since the Gabor frame (as we choose it) is inaccurate over a range of at least three window widths αX (see Eq. (8.12)) distance from the point at which the Gabor frame ends. Therefore, at least seven windows are needed for an accuracy of 10^{-3} in the middle of the computational domain, both for the spatial index m_x and for the spectral index n_x in Eq. (8.12). Consequently, at least seven values for index m_x and seven values for index n_x are needed, which amounts to a total of 49 coefficients per direction at minimum. Since we use a Gabor frame in two dimensions,

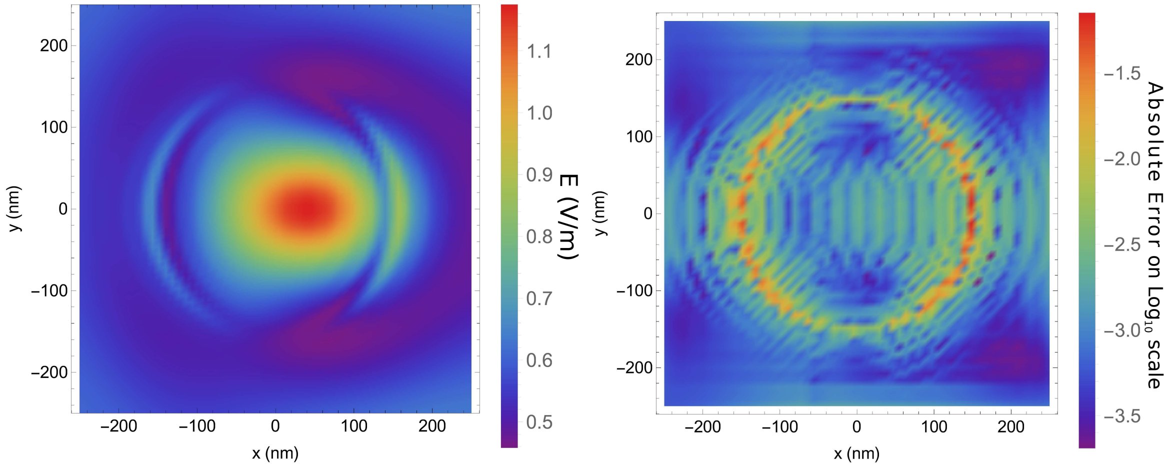


(a)

Figure 8.8: In Figure (a) both the computation time and the relative error in the far field, computed as the average of Figure 8.7(b) for a range of n_x and n_y for the Gabor frame. In (b) the same is shown, but now for different sampling Δ_z in the z-direction.

a minimum of 2401 unknowns is needed for a minimum simulation domain. Note that we assumed the use of a Gabor frame with $\alpha = \beta = \sqrt{2/3}$ and the Moore-Penrose inverse to calculate the dual window, where the accuracy increases exponentially with the distance to the truncated coefficients for this choice [156] for sufficiently smooth functions. This effect only applies to small simulation domains. For large simulation domains, the number of unknowns at the edge of the simulation domain is negligible.

The third and final example for which we computed the scattered electric field consists of six dielectric blocks of $350 \times 500 \times 100$ nm deposited on a slightly lossy dielectric substrate as shown in Figure 8.5(c). In Figure 8.11, the electric field is shown at $z = 10$ nm for $X = Y = 500$ nm, $\alpha = \beta = \sqrt{2/3}$ and $m_x \in \{-7, \ldots, 7\}$, $m_y \in \{-4, \ldots, 4\}$ and $n_x, n_y \in \{-7, \ldots, 7\}$, which amounts to one basis function per 27 nm. In the z-direction, 21 basis functions are used with stepsize $\Delta_z = 5$ nm. For the sampling of the regions of Type 2 and 3 we use $N_\nu = 2$, $N_s = N_m = 20$. Since the scatterer is larger than in the previous examples, it was efficient to choose larger window widths X and Y, which results a coarser sampling. From Figure 8.11(b) it is clear that this coarser sampling generates a somewhat more pronounced Gibbs ringing from the edges. However, in the far field, which is shown in Figure 8.12, the average relative difference with the JCMWave reference calculation is similar to that for the cylinder case, i.e. 2.5×10^{-3}, where the estimated relative accuracy of the reference calculation was of the order of 2×10^{-3}. Even though the scatterer extends much wider in the xy plane, the number of unknowns in the xy-plane was increased by only a factor of 5/3, while the accuracy in the far field remained similar. This clearly shows that the present method performs better for scatterers larger than a wavelength in size.



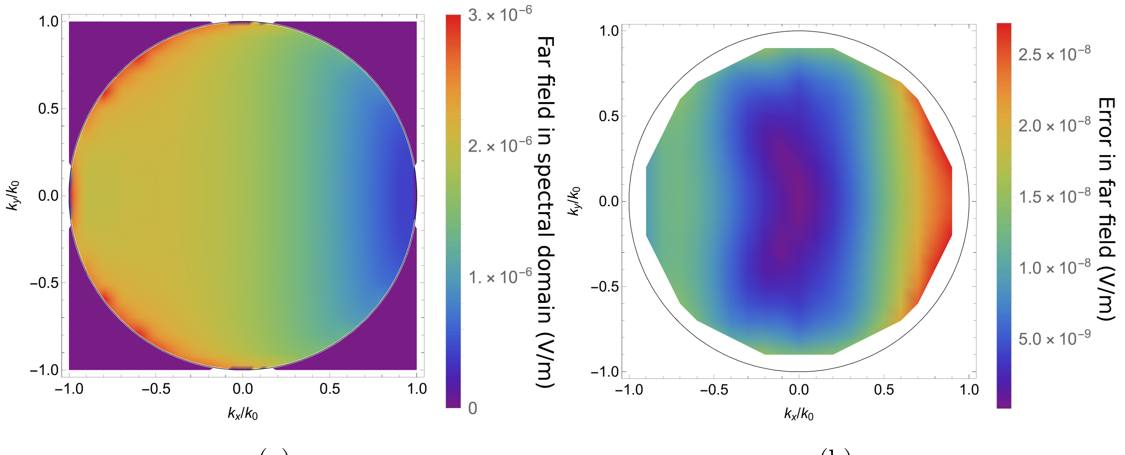
(a)

Figure 8.9: The electric field in the plane $z = 10$ nm plane for the scattering case in Fig. 8.5(b). In Figure (a) $|\mathbf{E}|$ is plotted for an incident plane wave with unit amplitude and in Figure (b) this is compared with the results obtained from JCMWave.

8.8 Conclusion

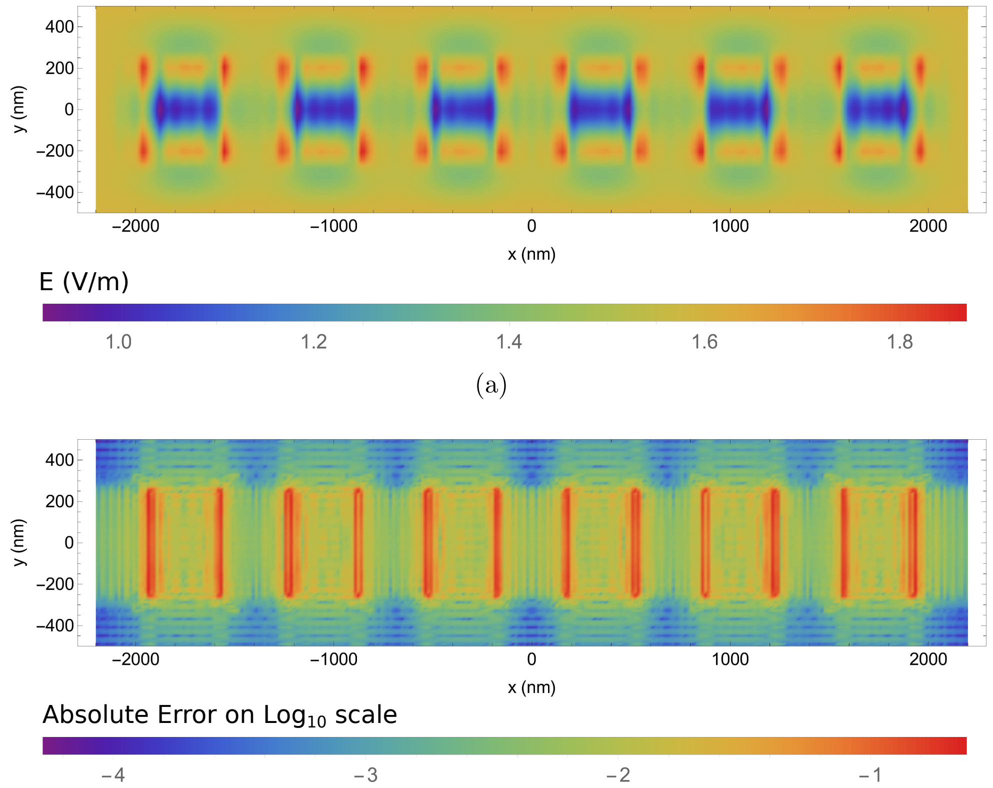
A volume integral equation for 3D scattering from finite dielectric objects embedded in a dielectric layered medium was presented in the mixed spatial and spectral domain and an algorithm based on Gabor frames was presented for the discretization. The algorithm employs a mixed spatial-spectral formulation and Gabor frames for the discretization. A representation of the Green function, contrast current density, and scattered electric field on a complex integration manifold is employed in the spectral domain. A normal-vector field formulation in the transverse spatial domain is employed to improve the convergence in the field-material interaction.

The accuracy of the present algorithm was compared to a FEM algorithm. The results of both algorithms in the far field agree with each other up to a relative error of 4×10^{-5} in one small numerical example. In the other two examples an agreement up to 2.8×10^{-3} and 2.5×10^{-3} were observed, because the FEM algorithm did not fully converge with the computational resources at hand. Numerical evidence was presented that the computational complexity of the present algorithm scales as $O(N \log N)$ with the number of unknowns.



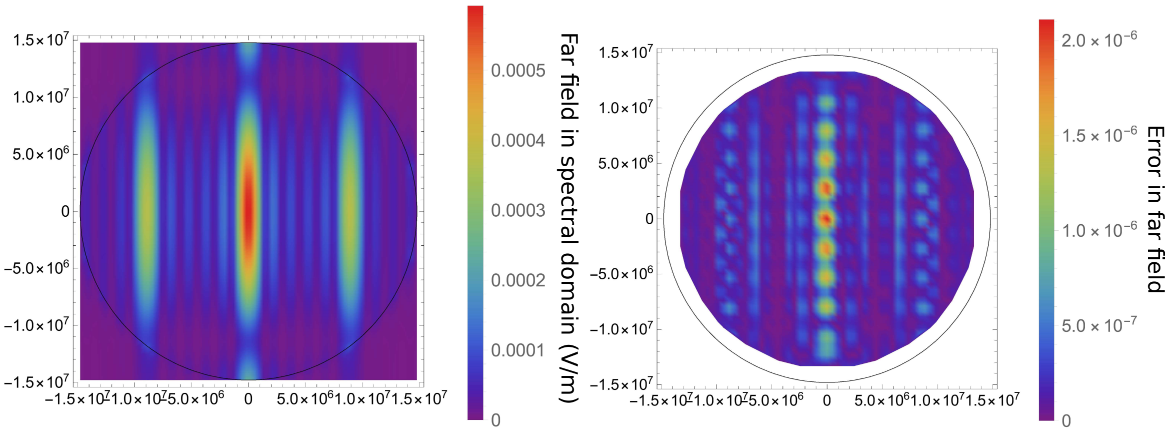
(a) (b)

Figure 8.10: The far field for the case in Fig. 8.5(b) as a function of the transverse wavenumber $\mathbf{k}_T/k_0$, scattered back into the half-space $z < 0$. In (a) the magnitude $|\mathbf{E}^s|$ of the scattered electric field is shown. In (b) the difference between a JCMWave validation run and the present algorithm is shown. An average relative error of $2.8 \cdot 10^{-3}$ was observed.



(a)

Figure 8.11: The electric field in the plane $z = 10$ nm plane for the scattering setup in Fig. 8.5(c). In Figure (a) $|\mathbf{E}|$ is plotted for a normally incident plane wave with unit amplitude and in Figure (b) this is compared with the results obtained using JCMWave.



(a)

(b)

Figure 8.12: The far field for the case in Fig. 8.5(c) as a function of the transverse wavenumber $\mathbf{k}_T/k_0$, scattered back into the half-space $z < 0$. In (a) the magnitude $|\mathbf{E}^s|$ of the scattered electric field is shown. In (b) the difference between a JCMWave validation run and the present algorithm is shown. An average relative error of $2.5 \cdot 10^{-3}$ was observed.

Chapter 9

Scaling to large simulation domains

In this chapter we will demonstrate that the algorithm of Chapters 6 to 8 can be applied to simulation domains of large size, without loss of accuracy, or excessive computational costs. First simulation domains of large transverse extent are covered in Section 9.1. Subsequently, in Section 9.2 a singular value decomposition is introduced for the M region of the complex-plane spectral domain integration manifold of Chapter 8. This singular value decomposition is particularly useful for simulation domains with large longitudinal extent. The discretization error originating from the other parts of the complex-plane spectral domain integration manifolds is controllable when the sampling rate in the spatial domain is kept constant, which is shown in Section 9.3. We end this chapter by showing numerical results for a scatterer of large transverse extent.

9.1 Scaling to large domain in the transverse direction [1]

9.1.1 Identifying potential problems

For a wide applicability of the algorithm based on a complex spectral path deformation of the integral equation, it is important to show that it can be applied to large simulation domains. Since the discretization and complex spectral path is the same in the x and y direction for both two and three-dimensional problems, we will focus on 2D problems for notational efficiency.

The scaling to larger simulation domains is not trivial, since the exponentials of the form $\exp(\pm Ax)$ in Eq.(6.22) may lead to instabilities in the algorithm for large $x \in [-W, W]$. Since we can make a choice for the simulation parameters where $AW = c$ is a constant, independent of the simulation width W, the exponential factors will then be limited to $\exp(c)$. However, this means that $A \propto 1/W$ and consequently for larger simulation domains the deviation from the real k_x axis on the complex-plane path is smaller. This means that

[1] The content of this section was presented at the Progress In Electromagnetic Research Symposium (PIERS) 2017 [157]

for larger simulation domains functions are evaluated closer to the branch cuts and poles in the Green function. On the other hand, a larger W also means that the spectral domain is sampled finer, since the spectral sampling density is directly proportional to the size of the simulation domain W. The remaining part of this section is devoted to identifying the parts of the Green function that could become problematic. For each of these parts we will demonstrate numerically that the finer sampling compensates for the smaller distance to the poles and branch cuts.

In the 3D Green function for a homogeneous medium, Eq. (2.38), two potential problems can be identified in the xz component. The other components do not show any additional poles or branch cuts, so we will ignore them. These potential problems are displayed in boxes in

$$G^h_{x,z}(k_x, k_x, z|z') = (\varepsilon_{rb,i}k_0^2 - k_x^2)\boxed{e^{-\gamma|z-z'|}}\boxed{\frac{1}{2\gamma}}. \tag{9.1}$$

All of these are related to the branch cuts in γ. In a multilayered medium, reflection coefficients come into play. In principle, the reflection coefficient between two interfaces is a continuous function, but in the calculation of the effective reflection coefficients, Eq. (2.42), a factor of

$$\frac{1}{1 - r_e^u(\mathbf{k}_T)r_e^d(\mathbf{k}_T)e^{-2\gamma d_i}} \tag{9.2}$$

is present, where poles can be encountered, when $|r_e^u| = |r_e^d| = 1$ which happens when $\mathbf{k}_T$ corresponds to an angle larger than the critical angle and when $|\mathbf{k}_T| < \varepsilon_{rb,i}k_0$, i.e. such that γ is purely imaginary.

9.1.2 Branch cuts in γ and $1/\gamma$

To show that the branch cuts in γ are not problematic, we will plot γ for several values of A with $k_0 = 1$ and $\varepsilon_{rb,i} = 1$, such that $\gamma(k_x) = \sqrt{k_x^2 - 1}$. Decreasing A corresponds to larger simulation domains and therefore the sampling density increases in the spectral domain. Therefore, we use a scaling on the k_x axis around the branch point at $k_x = 1$ given by $k_x(\kappa) = 1 + A\kappa + jA$. This scaling is chosen such that the sampling density in κ is constant with respect to changes in A. This means that when all lines converge, a constant sampling density on the κ-axis corresponds to a constant discretization error when the simulation domain is increased.

Since $\gamma(kx(\kappa))|_{\kappa=0} = \sqrt{1-(1-jA)^2} \approx \sqrt{2jA}$ we also use a scaling on the y-axis by a factor of $1/\sqrt{A}$. From the results of this procedure, as presented in Figure 9.1, it is clear the results converge to a single line for different values of A.

The same procedure is followed for the $1/\gamma$ function. Since $1/\gamma(kx(\kappa))|_{\kappa=0} = 1/\sqrt{1-(1-jA)^2} \approx 1/\sqrt{2jA}$ the variable on the y axis is now scaled by a factor of $\sqrt{A}$. In Figure 9.2 a good convergence is observed as well.

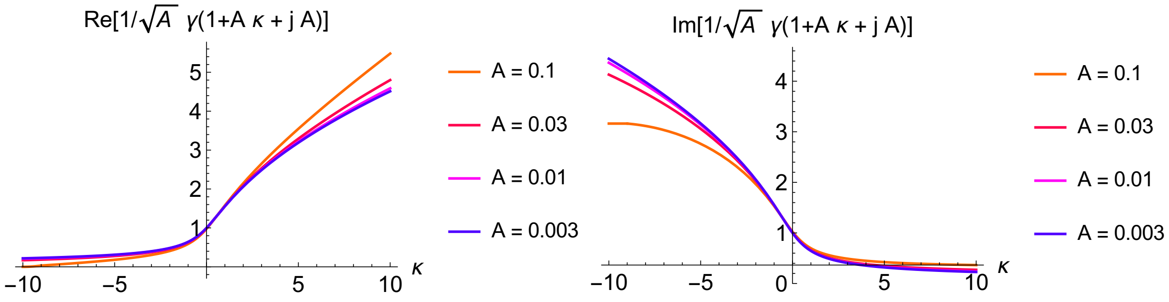


Figure 9.1: Plot showing the shape of the γ function against coordinates that are scaled to correspond to a fixed sampling density for large simulation domains. Left shows the real part and right shows the imaginary part of this function.

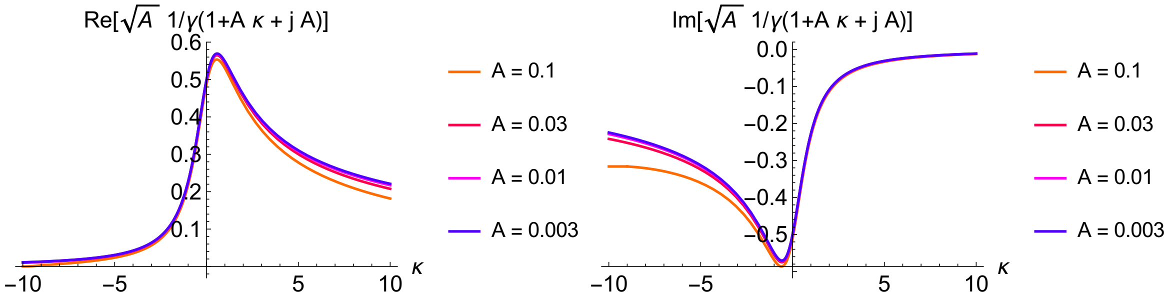


Figure 9.2: Plot showing the shape of the $1/\gamma$ function against coordinates that are scaled to correspond to a fixed sampling density for large simulation domains. Left shows the real part and right shows the imaginary part of this function.

9.1.3 Oscillations in the propagation function $\exp(-\gamma|z-z'|)$

It is not directly the presence of branch cuts in the propagation function that gives rise to problems, rather the presence of oscillations along the branch cut. The number of oscillations is related to the propagation distance $|z-z'|$, which is not directly related to A. However, a large A has the effect of damping the oscillations with the highest frequency, since the oscillations are present on the real k_x axis and dissipate quickly with increasing distance..

This behavior is clearly visible in Figure 9.3, where a plot has been made for $g(k_x) = \exp(-z\gamma(k_x))$, with $z = 20$. For large A, more oscillations are visible, because a larger part of the k_x range is plotted since $k_x = 1 + A\kappa + jA$. For smaller A, the number of oscillations decreases, although the amplitude increases. Convergence to a limiting function is therefore not observed, but clearly a smaller A yields a propagation function that is easier to discretize for a fixed z, since the number of oscillations decreases. On the other hand we can conclude that for a simulation domain that is small in the x direction (Large A) and large in the z direction, the method breaks down, since that yields a large number of undamped oscillations. Therefore a simulation width W should not be chosen too small. In our numerical examples we observed that a simulation of about two wavelengths is usually enough for a good accuracy. A much smaller simulation domain is not realistic, since the middle part of the complex spectral path in Eq. (6.20) becomes too large for large A. This is mainly because the information density in the Green function requires several samples per wavenumber $\sqrt{\varepsilon_{rb,i}}k_0$, and when the simulation domain W is too small, this is not satisfied.

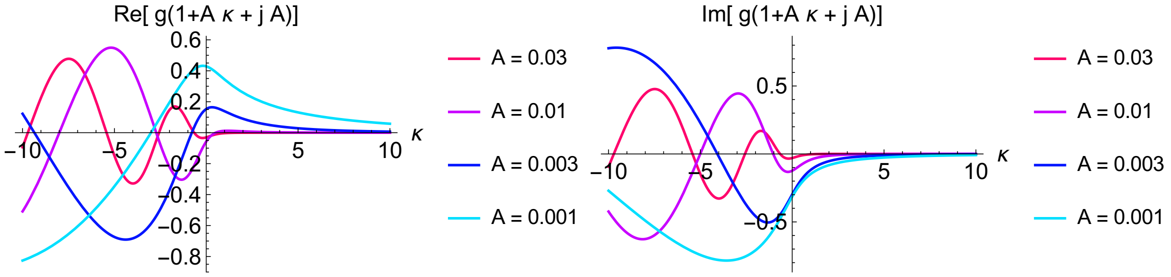


Figure 9.3: Plot showing the shape of the propagation function $g(x) = \exp(-20\gamma)$ against coordinates that are scaled to correspond to a fixed sampling density for large simulation domains. Left shows the real part and right shows the imaginary part of this function.

9.1.4 Poles in the effective reflection coefficients

A pole, such as found in the effective reflection coefficients (e.g. Eq. (2.42)) can be modeled by the function $p(k) = 1/(k^2 - 1)$. For the poles we use a scaling factor of A on the y axis. The same procedure as in Section 9.1.3 can be carried out for this function and in Figure 9.4 we see that all graphs coincide, which means that also poles are not problematic.

To conclude, none of the potential causes mentioned above gives rise to problems for scaling to large simulation domains. Therefore, we conclude that scaling to larger simulation domains will not harm the accuracy of the discretization on the complex-plane path deformation under the conditions that AW is constant and that the sampling on the x-axis remains constant, which means that the sampling rate on the k_x axis is proportional to W.

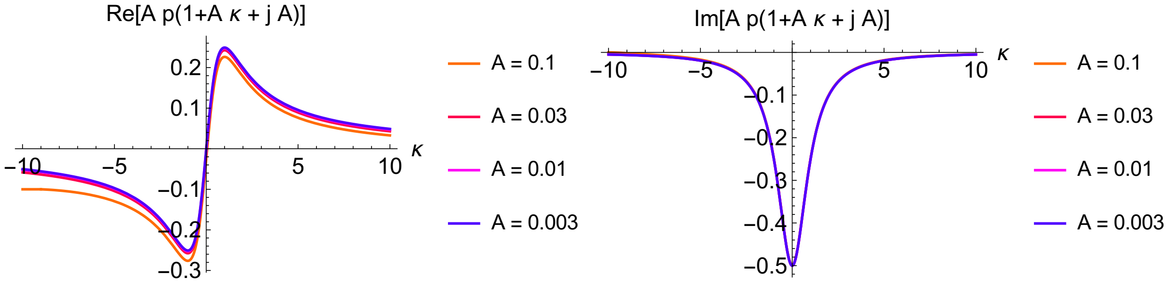


Figure 9.4: Plot showing the shape of the propagation function $p(k) = 1/(k - 1)$ against coordinates that are scaled to correspond to a fixed sampling density for large simulation domains. Left shows the real part and right shows the imaginary part of this function.

9.2 A singular value decomposition for Region M [2]

9.2.1 Abstract

With a 3D spatial spectral integral-equation method for EM scattering from finite objects, a significant part of the computation time is spent on a middle region around the origin of the spectral domain. Especially when the scatterer extends to more than a wavelength in the stratification direction, a fine discretization for this region is required, which consumes much computation time in the transformation to the spatial domain. Numerical evidence is shown that the information in the middle region of the spectral domain is almost linearly dependent. Therefore, a truncated singular-value decomposition is proposed to make the computation time largely independent of the discretization of this middle region. For a practical example the increased computational efficiency and the approximation error of the singular-value decomposition are shown.

9.2.2 Introduction

Previously a 2D and 3D spatial spectral integral equation methods for electromagnetic scattering from dielectric objects in multilayered media were proposed [118, 159, 152] (Chapters 5-8). This approach is based on handling the field-material interaction in the spatial domain and the Green function in the spectral domain. Fourier transformations are needed to transform the contrast current density and the electric field between the spatial and spectral domain. The scattered electric field is computed from the contrast current density by a recursive set of multiplications by several parts of the Green function.

The Green function contains branch cuts and poles. To avoid these singularities the contrast current density and scattered electric field are therefore represented on a manifold that is deformed slightly into the complex plane. This deformation decomposes each of the two transverse spectral directions into three parts, two parts containing most information and a small part connecting them. In total, this yields nine different regions. Although the connecting part is small, still a significant amount of computation time is spent on transforming information represented in this part to the spatial domain. Since the connecting part contains information about waves traveling close to the stratification direction, a fine discretization is needed especially for objects with a large extent in this direction. Each basis function in this connecting part is Fourier transformed individually, although the contained information is largely redundant. We propose a singular-value decomposition to remove the redundancy and speed up these computations.

9.2.3 Spectral path

We use the formulation for electromagnetic scattering as explained in [135, 152] (Chapters 2, 6 and 8. In these articles, the branch cuts and poles in the Green function are evaded

[2] This section has been accepted for the International Conference on Electromagnetics in Advanced Applications ICEAA 2017 Proceedings [158]

by representing the contrast current density, the electric field and the Green function in the spectral domain on a path

$$\tau_\alpha(k_\alpha) \in \begin{cases} k_\alpha - jA & \text{if } k_\alpha < -A \\ (1+j)k_\alpha & \text{if } -A \leq k_\alpha < A \\ k_\alpha + jA & \text{if } k_\alpha > A, \end{cases} \tag{9.3}$$

with $\alpha \in \{x, y\}$. The $k_x k_y$ plane is then divided into nine regions as shown in Figure 9.5, where we will focus on the middle region indicated by M. In the z direction, a piecewise-linear (PWL) expansion in the spatial domain is employed.

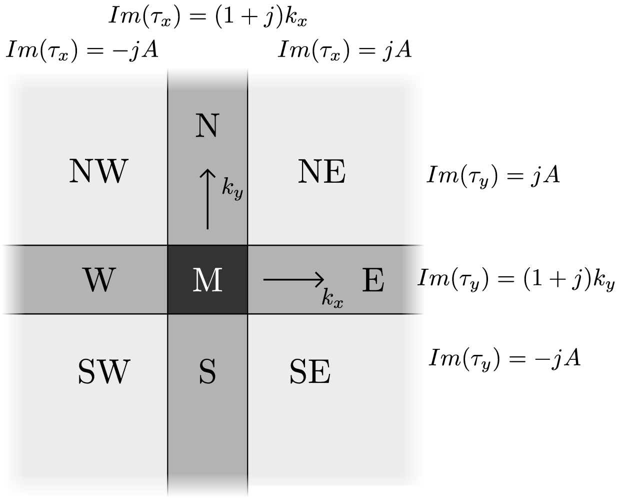


Figure 9.5: Subdivision of the $k_x - k_y$ plane with piecewise-constant and piecewise-linear imaginary shifts.

9.2.4 Discretization of the middle region

2D: one-dimensional Taylor series

Gabor frames are used as a discretization for the contrast current density on the regions $k_\alpha < -A$ and $k_\alpha > A$. In the figure M stands for middle and the other capital letters abbreviate the wind directions. For the two-dimensional algorithms in [159, 135, 147] (Chapters 6 and 7), where the y direction is absent, a Taylor-series approximation with $N_a + 1$ terms is employed on the connecting part $-A < k_x < A$. A function $f(k_x)$ is approximated on the spectral path of Eq. (9.3) for $k_x \in [-A, A]$ by

$$f(\tau_x(k_x)) \approx \sum_{n=0}^{N_a} \frac{f^{(n)}(0)}{n!} \tau_x(k_x)^n. \tag{9.4}$$

So, instead of function values on the path τ_x, the derivatives at $k_x = 0$ are kept as a representation. In the spectral domain, we multiply the contrast current density and the

Green function. The result $m(k_x)$ of multiplying two functions, say $g(k_x)$ and $h(k_x)$, where the derivatives of these functions are known, can then be calculated by the Leibnitz rule as

$$m^{(n)}(0) = \sum_{\ell=0}^{n} \frac{n!}{(\ell-n)!\ell!} g^{(\ell-n)}(0) h^{(n)}(0). \tag{9.5}$$

Since N_a derivatives are needed, the number of operations for such a multiplication scales as $O(N_a^2)$ for the number of terms in the Taylor series. For problems without features extending more than a wavelength in the z-direction, around ten terms are required in the Taylor series. Even though this Taylor series is computationally not very efficient, it does not significantly contribute to the overall computation time, owing to the small number of terms in the truncated Taylor series.

3D: Piecewise-linear functions

For a 3D algorithm, functions are represented on complex paths τ_x and τ_y, in the k_x and k_y direction, respectively. Although a two-dimensional Taylor series would be able to represent the Green function and contrast current density on the region $(k_x, k_y) \in [-A, A]^2$, the quadratic computational complexity in the multiplication in Eq. (9.5) makes it inefficient for a generalization to two dimensions. Therefore, the Taylor-series approach was replaced by a PWL discretization for the 3D algorithm in [152] (Chapter 8). With the PWL discretization, a function $f(\tau_x(k_x), \tau_y(k_y))$ is approximated by a list of function values $f_{n_x,n_y} = f(\tau_x(n_x A/N_p), \tau_y(n_y A/N_p))$ as

$$\begin{aligned} f(\tau_x(k_x),&\tau_y(k_y)) \approx \\ &\sum_{n_x=-N_p}^{N_p} \sum_{n_y=-N_p}^{N_p} \Lambda_{n_x}(k_x)\Lambda_{n_y}(k_y) f_{n_x,n_y}, \end{aligned} \tag{9.6}$$

where

$$\Lambda_n(k) = \max\{0, 1 - |xN_p/A - n|\}. \tag{9.7}$$

With these basis functions most terms in the sum of Eq. 9.6 are zero. Therefore, the multiplication operation simply becomes a pointwise multiplication, which scales linearly with the number of basis functions in Region M, instead of quadratically in Eq. (9.5).

9.2.5 Transformation to the spatial domain

Transformation of PWL functions

To transform the representation on Region M in Eq. (9.6) to the spatial domain as needed for the electric field, we use the Fourier integrals over the complex spectral path restricted to M, i.e.

$$L_n(x) = \frac{1}{2\pi} \int_{-A}^{A} dk\ \Lambda_n(k) e^{j(1+j)kx}. \tag{9.8}$$

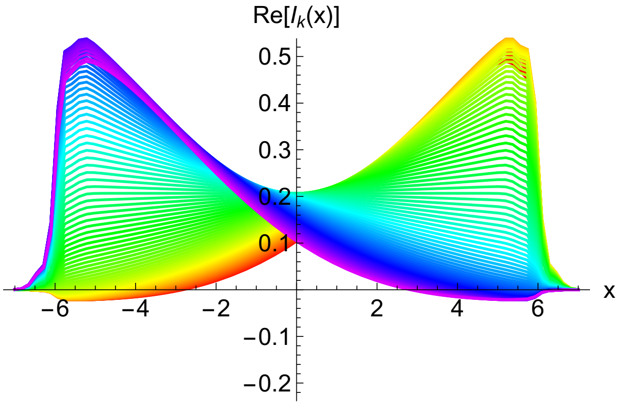


Figure 9.6: A set of $L_n(x)$ for $N_p = 20$, red signifying $n = 1$ and purple $n = 41$, and $A = 5.6 \times 10^{-3}$. These integrals are cut off at $x = \pm 6$ corresponding to the discretization used in the spatial domain.

Consequently, $f^M(x)$, i.e. the contribution of region M to the spatial domain, can be written as

$$f^M(x) = \sum_{n_x=-N_p}^{N_p} \sum_{n_y=-N_p}^{N_p} f_{n_x,n_y} L_{n_x}(x) L_{n_y}(y). \tag{9.9}$$

Now we rewrite this into a matrix formulation. First, let $\underline{f}$ represent the size $(2N_p+1)^2$ vector of f_{n_x,n_y} coefficients. The spatial domain is discretized via the discretization operator $\mathcal{S}$ into N_s basis functions $b_m(x,y)$. For example, we denote the discretized version of the arbitrary function $t(x,y)$ as $t_m = \mathcal{S} \circ t(x,y)$, such that $t(x,y) = \sum_{m=0}^{N_s} t_m b_m(x,y)$. We use this discretization operator to calculate $L_{m,(n_x,n_y)} = \mathcal{S} \circ (L_{n_x}(x) L_{n_y}(y))$, such that

$$L_{n_x}(x) L_{n_y}(y) = \sum_{m=1}^{N_s} L_{m,(n_x,n_y)} b_m(x,y), \tag{9.10}$$

where we will use the notation $\underline{\underline{L}}$ for the $N_s \times (2N_p+1)^2$ matrix, with in general $N_s >> N_p^2$. A discretized counterpart of Eq. (9.9) can be computed by computing $\underline{f}^M = \underline{\underline{L}} \cdot \underline{f}$, which yields a vector of length N_s. This matrix-vector product has to be computed, which requires $O(N_s N_p^2)$ operations. Especially for large, but realistic, N_p this becomes the dominating contribution to the computation time, as will be shown in Section 9.2.6.

Singular-value decomposition

In Figure 9.6, a set of Fourier transforms of the PWL basis functions, L_n, as defined in Eq.(9.8) are shown. Clearly, there is considerable redundancy in this set. For a large distance from $x = 0$, these integrals $L_n(x)$ are less redundant. However, they are only required for small x where the scattering object is located. This suggests that the summation in Eq. (9.9) can be accelerated by a truncated singular-value decomposition (SVD) [38, Chapter 2.6] at initialization.

The matrix $\underline{\underline{L}}$ can be decomposed into

$$\underline{\underline{L}} = \underline{\underline{U}} \cdot \underline{\underline{\Sigma}} \cdot \underline{\underline{V}}^T, \tag{9.11}$$

where many entries in the diagonal $(2N_p + 1)^2$ matrix $\underline{\underline{\Sigma}}$ are negligible. The elements are said to be negligible when they are smaller than threshhold ϵ, and will then be set to zero. Afterwards, $N_t \leq (2N_p + 1)^2$ significant entries in the diagonal matrix remain. Now $\underline{\underline{L}}$ can be approximated by $\underline{\underline{L}} \approx (\underline{\underline{\tilde{U}}} \cdot \underline{\underline{\tilde{\Sigma}}}) \cdot \underline{\underline{\tilde{V}}}^T$, where $\underline{\underline{\tilde{V}}}^T$ is an $N_t \times (2N_p + 1)^2$ matrix and $(\underline{\underline{\tilde{U}}} \cdot \underline{\underline{\tilde{\Sigma}}})$ is an $N_s \times N_t$ matrix.

Because there is a significant redundancy in the system, usually we choose $N_t << (2N_p + 1)^2$, computing

$$\underline{f}^M = (\underline{\underline{\tilde{U}}} \cdot \underline{\underline{\tilde{\Sigma}}}) \cdot (\underline{\underline{\tilde{V}}}^T \cdot \underline{f}) \tag{9.12}$$

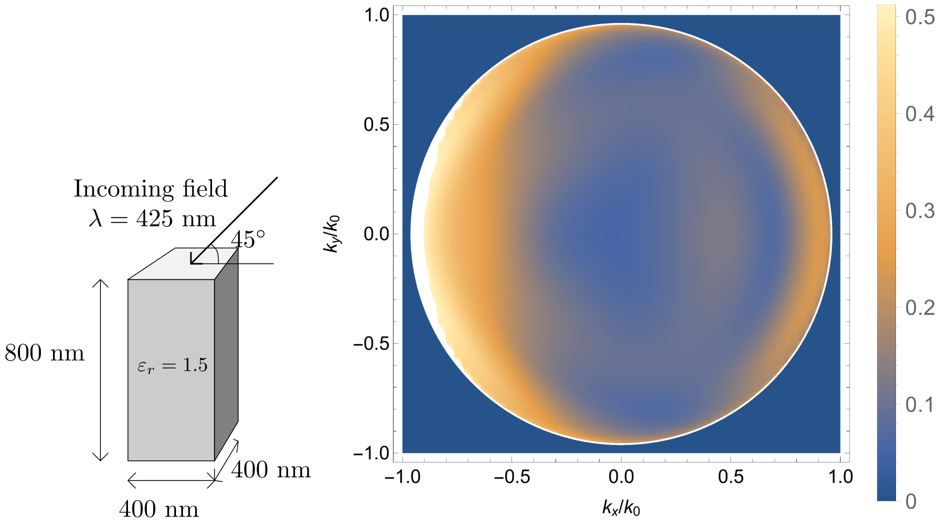
will require only $O(N_t(N_p^2 + N_s))$ operations instead of $O(N_p^2 N_s)$. Since the spectral Region M is of small size, compared to the complete spectral range included in the simulation, it represents only a small amount of information compared to the number of spatial basis functions N_s. For this reason N_t does not show a significant increase after a certain number PWL basis function N_p, as will be shown next.

9.2.6 Impact on accuracy

The idea to apply the SVD on the region M in the spectral domain is tested by computing the scattering from a dielectric block of 400 × 400 × 800 nm illuminated by an incident plane wave with wavelength $\lambda = 425$ nm as shown in Figure 9.7(a). The amplitude of the scattered electric field, $|\mathbf{E}|$, in the far field is plotted against the transverse part of the wavenumber (k_x, k_y), and $k_0 = 2\pi/\lambda$ in Figure 9.7(b). For large N_p, the rank of the truncated SVD increases to a value that is independent of N_p and depends merely on the error level ϵ as is shown in Figure 9.8(a). Without the SVD, the computation time increases significantly with large N_p. However, when the truncated SVD is used, the computation time does not depend strongly on N_p as can be observed in Figure 9.8(b). We did not include the computation time of the SVD in these results, but this number was small compared to the total computation time and in principle the SVD-decomposition can be cached. To show how the truncated SVD influences the accuracy of the far field, we have plotted the $\mathcal{L}^2$-norm of the relative difference in the far-field data for different truncation thresholds ϵ and numbers of PWL basis function N_p. A reference was calculated with $\epsilon = 10^{-7}$ and $N_p = 40$. In Figure 9.8(c), it can be observed that the error due to the use of the SVD can be made small, when N_p is chosen large enough. For an error level of 10^{-3} a truncation threshold $\epsilon = 10^{-2}$ is already sufficient, since region M contributes only a part of the complete electric field, to which all regions in Figure 9.5 contribute.

9.2.7 Conclusion

The Fourier transformation from the spectral domain to the spatial domain of a small patch of the spectral domain can take up more than half of the computation time. A



(a) (b)

Figure 9.7: (a) Scattering setup, (b) the upwards-directed electric-field amplitude in the far field, $|\mathbf{E}(k_x, k_y)|$ (a.u.) from this scattering setup.

truncated singular-value decomposition was used to speed up the computation of this Fourier transform, so the time spent on this part was reduced to less than 10%, with an error level of 10^{-3}.

Numerical evidence was shown that the rank of the singular-value decomposition and therefore the computation time depends weakly on the fineness of the discretization on this part of the spectral domain.

9.3 Scaling to large longitudinal distances

We will now demonstrate how the accuracy in the present algorithm scales to large simulation regions in the z direction. To save space, we will assume in this section $z_{min} = 0$ and a background permittivity $\varepsilon_{rb,i} = 1$. We deal with the lossless case, since the propagation function becomes smoother with increasing losses, so the lossless case is the hardest.

The main difficulties with the spectral discretization is that for large z_{max} the vertical propagation function, $e^{-\gamma z_{max}}$, which exhibits many oscillations in the domain $k_x \in [-k_0, k_0]$ and a sharp cutoff for $|k_x|$ larger than k_0. Although the complex spectral path smooths these effects, their behavior for large z must be investigated. In the spatial domain, this propagation function is very well behaved. At a large distance, the number of oscillations within the simulation domain will even decrease to an almost constant function since the wave front is almost flat. The contributions of the spectral oscillations are only visible at horizontal values in the xy-plane that are outside our simulation domain.

The oscillations in the spectral domain are caused by γ being purely imaginary (for non dissipative media) in the range $k_x \in [-k_0, k_0]$. On the complex spectral path and for

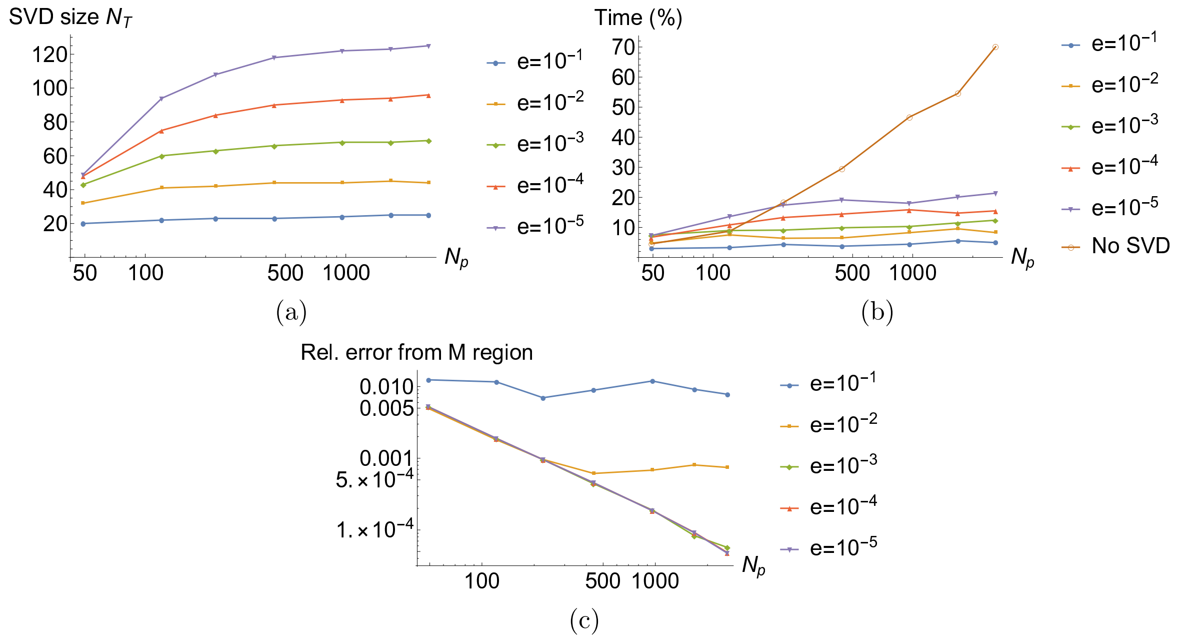


Figure 9.8: (a) The size of the significant part of the SVD truncated at ϵ with N_p basis functions. (b) The percentage of the total computation time that is spent on region M. With SVD the computation time was between 290 and 350 seconds on a single core of a 3.1 GHz Xeon E5-2687 processor, without SVD this increased up to 950 seconds with $N_p = 25$. (c) The error in the region M discretization relative to the complete far-field result as in Figure 9.7(b)

small A, γ can be approximated as

$$\gamma = \sqrt{(k_x \pm jA)^2 - k_0^2} \approx j\gamma_r - A|k_x|/\gamma_r, \tag{9.13}$$

where γ_r equals $\gamma(Re(k_x))/j$, which is real-valued for $|k_x| < k_0$. The A/γ_r term is completely real and therefore it is a factor that dampens

$$\exp(-\gamma z) \approx e^{-j\gamma_r z} \quad e^{-A|k_x|z/\gamma_r}. \tag{9.14}$$

The approximation in Eq. (9.13) allows us to clearly identify the two main influencing factors of the absolute discretization error as is shown in Figure 9.9. The first is that for large z the propagation function $\exp(-j\gamma_r z)$ in Eq. (9.14) becomes increasingly oscillatory, which makes the discretization with a Gabor frame exhibit a large error. The second factor is that this function is dampened exponentially by $\exp(-A|k_x|z)$ in Eq. (9.14) for large z and nonzero k_x, the absolute discretization error decreases with the exponential damping of the function itself. For small k_x the damping is small, but this falls in the middle region of the complex spectral path, the shaded region in the Figure 9.5, where a different discretization is employed. The question is whether the damping will compensate for the increased number of oscillations for large propagation distances to keep the absolute error small enough.

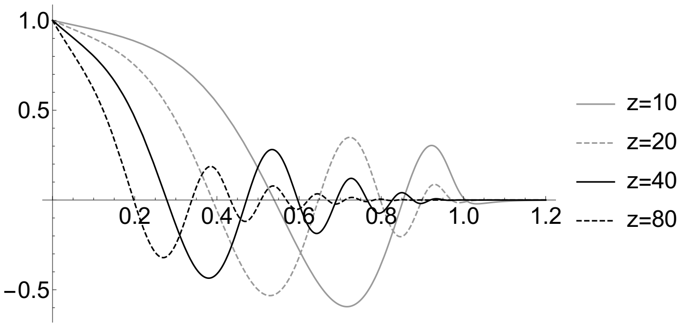


Figure 9.9: The propagation function $\exp(-\gamma z)$ for $k_0 = 1$ plotted for different z_{max} values.

When we fix the simulation range W and increase the propagation distance z, we observe that first the approximation error increases because the number of oscillations increases. When the propagation distance is increased even further we observe that the propagation function as a whole tends to zero, except in a decreasing area around $k_x = 0$. However, Gabor coefficients are not used in the middle part around $k_x = 0$, so at $k_x = \pm A$, where the Gabor-represented part starts, the whole propagation function converges to zero, including its error. From this observation we expect there to be a maximum in the absolute approximation error for $\exp(-\gamma z)$ at a certain z for each width W of the domain. In Figure 9.10 (a) and (b) this maximum is clearly visible. Here we see the absolute error in the discretization of $exp(-\gamma z)$ on the left part of the complex spectral path, for different sizes of the simulation domain as a function of the propagation distance z. These plots have been made for $\lambda = 425$nm, a Gabor frame with $T = 300$nm, $\alpha = \beta = q/p = 2/3$ oversampling in the Gabor frame, and the Gabor coefficients were truncated at $M = \{4, 6, 8, 12, 16, 24\}$ and $N = 9$ for $m \in \{-M, \cdots, M\}$ and for $n \in \{-N, N\}$ in Eq. (4.3). The $\mathcal{L}^2$ norm of the propagation function on the real k_x-axis, to which all these absolute errors should be compared, is of the order $1.0 \cdot 10^7 \cdots 1.4 \cdot 10^7$ and depends weakly on A. It is important to find the position $z = p(W)$ of the maximum in the error as a function of W, in order to test whether there are certain combinations of z and W where an increase in both W and z_{max} can increase the maximum error in the propagation function. When W increases, the maximum absolute error of the approximation is found at $z = p(W)$ and the error should decrease along this curve for a stable scaling to large simulation domains. This would mean that, when the simulation width (W) increases, the absolute error in the propagation function is bounded by a decreasing function for any longitudinal distance z.

Although we were not able to find an analytical expression for this function $p(W)$, we have two strong arguments and numerical evidence concerning the behavior of this function. The first argument is a renormalization argument. Assume that A^*, Z^* and W^* are such that the maximum error is reached for simulation width W^*. Now we would like to find the propagation distance that maximizes the error for simulation width $2W^*$. From the assumption that Z^* and A^* have values at which the damping starts to take over, it follows that $\exp(-A^*kz)$ has a strong damping for larger k. We can assume that the

input			Damping		Oscillations		
W	Z	k-range	damping exponent	significant	oscillations	sampling	significant
W^*	Z^*	$A^* \dots 2A^*$	$-A^{*2}Z^* \cdots - 2A^{*2}Z^*$	yes	reference	reference	yes
$2W^*$	Z^*	$A^* \dots 2A^*$	$-A^{*2}/2Z^* \dots \ (-A^{*2}Z^*)$	yes	same	double	no
$2W^*$	$2Z^*$	$A^* \dots 2A^*$	$-A^{*2}Z^* \dots \ (-2A^{*2}Z^*)$	yes	double	doube	yes
$2W^*$	$4Z^*$	$A^* \dots 2A^*$	$-2A^{*2}Z^* \dots \ (-4A^{*2}Z^*)$	no	quadruple	double	yes
$2W^*$	Z^*	$A/2^* \dots A^*$	$-A^{*2}/4Z^* \dots \ (-A^{*2}/2Z^*)$	yes	half	double	no
$2W^*$	$2Z^*$	$A/2^* \dots A^*$	$-A^{*2}/2Z^* \dots \ (-2A^{*2}Z^*)$	yes	same	doube	no
$2W^*$	$4Z^*$	$A/2^* \dots A^*$	$-A^{*2}Z^* \dots \ (-2A^{*2}Z^*)$	yes	double	double	yes

Table 9.1: Table identifying where approximation errors can be expected during renormalization

largest part of the error originates from $k \in [A^*, 2A^*]$, since the damping for $k > 2A^*$ will be dominating anyway.

In Table 9.1 we show where significant contributions to the approximation error can be expected. The first row shows the reference at W^*, A^* from which we start and the k_x interval $A^*, 2A^*$ where the discretization error is the largest. We assume that these values represent a maximum in error. The columns 'significant' indicate whether or not a maximum error can be expected based on the damping or based on the number of oscillations per sample, respectively.

The rows with $(2W^*, Z^*)$ show that the error will be significantly lower because of the increased sampling, even though there is less damping. The rows with $(2W^*, 2Z^*)$ show a possibly significant contribution for $k \in [A^*, 2A^*]$, but not for $k \in [A^*/2, A^*]$. However, the largest contribution to the error should be in $k \in [A^*/2, A^*]$, otherwise some further propagation in Z will still increase the error and hence this is not the maximum. The rows with $(2W^*, 4Z^*)$ show a significant contribution to the error for $k \in [A^*/2, A^*]$, with significant damping for larger k. Since this is the same situation as for (W^*, Z^*) we conclude from the renormalization that the largest error for $2W^*$ can be expected at $Z = 4Z^*$. Consequently, the maximum error can be expected at $Z = \alpha W^2$, with α a constant that we can determine numerically and which depends on the details of the discretization.

A second argument why the maximum of the error can be expected near $Z = \alpha W^2$ follows from a spatial perspective. Spatially, the propagation function can be looked upon as an electric field point source that radiates uniformly. Our approach divides the spectral domain in three parts, of which f_L and f_R represent left- and right-propagating waves, respectively. For k_x on the real line, we can rigorously pose this, but since we have moved a little bit into the complex plane, this is no longer entirely true. However, waves moving to the right are still heavily damped for negative k_x and vice versa.

On the complex spectral path, we can distinguish between contributions propagating largely in the z direction, which are mainly represented in the middle part of the integration path and waves moving to the left or to the right for a negative or positive real part of k_x, respectively. An angle φ can be defined between which waves are represented by the middle part or by the left and right parts of the complex integration path. This angle depends on A according to

$$\sin\varphi = \frac{A}{k_0}, \tag{9.15}$$

since the left and right parts start at $Re(k_x) = \pm A$.

The left and right parts of the propagation function discretization are damped at large propagation distances because their contribution is outside the simulation domain, as illustrated in Fig. 9.11. At a large longitudinal distance, the left and right parts are negligible, so it is the middle part of that carries the information of the propagation function and in the middle part $|k_x| < \sqrt{2}A$.

When the simulation width W increases, A decreases and with it the angle φ. Since a doubling of W leads to halving A and therefore also to halving φ (when φ is sufficiently small), the propagation distance z over which the outer parts contain a significant contribution is quadrupled. Consequently, the significant z-range for the left and right parts of the propagation function again behaves as $Z = \alpha W^2$. Since the largest approximation error can again be expected towards the end of this range, because there the highest number of oscillations can be observed in the propagation function, the maximum in the error is again expected to be reached around $Z = \alpha W^2$.

In Figure 9.10(a) and (b), we show the absolute error in the propagation function for different simulation region sizes and different propagation distances with and without oversampling. In Figure 9.10(c) we show that for larger simulation regions (where $\sin(\varphi) \approx \varphi$) the maximum error can indeed be found along a line $Z \propto W^2$. It is also clearly visible in Figure 9.10(d) that the absolute error decreases along this line when W increases. From this trend we can conclude that an increase of the simulation region with a constant sampling density of Gabor coefficients will not deteriorate the accuracy of this representation of the Green function for the homogeneous medium, Eq. (2.38).

Throughout this whole discussion we ignored the number of coefficients needed for the middle part. For the middle part there is a need for an increasing number of derivatives when Z_{tot} increases. We observed that for realistic simulation sizes from a few to tens of wavelengths the time needed to evaluate the middle part is negligible compared to the outer parts for 2D simulations, when the SVD decomposition in Section 9.2 is employed.

9.4 Numerical example: large grating

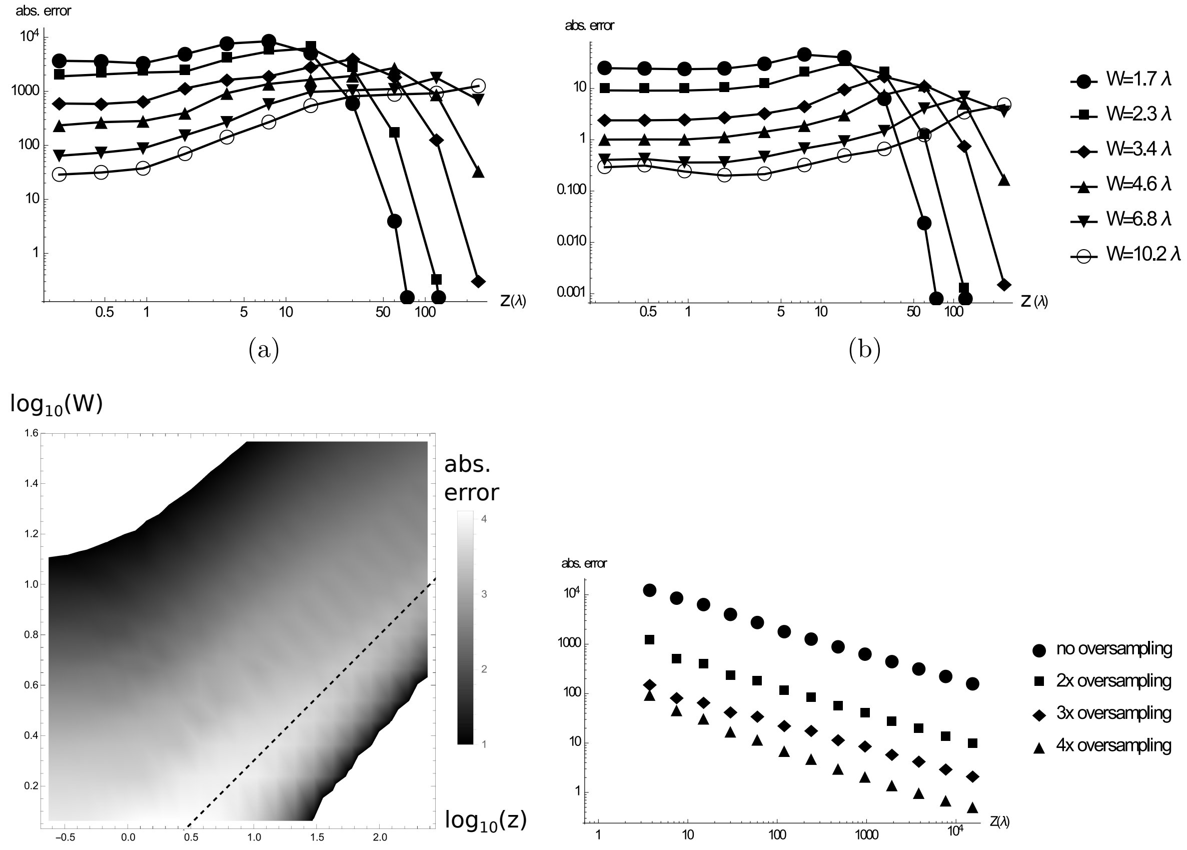
To conclude this chapter, we show the result for scattering from a dielectric grating with a varying number of repeating elements. The scattering setup is shown in Figure 9.12. We are interested to know the number of repeating elements after which an infinitely repeating scatterer is a good approximation of this scattering setup. The far-field scattering is computed for $N \in \{2, 4, 6, 10, 16, 25, 36\}$ repeating elements. We have chosen a Gabor frame with a window width $X = 500$ nm. The range of Gabor coefficients was chosen in the x direction $m_x \in \{-M_x, \cdots, M_x\}$ and $n_x \in \{-N_x, \cdots, N_x\}$ in Eq. (4.3) and similarly in the y direction. In the x direction M_x is varied with the number of repeating elements $M_x \in \{4, 7, 7, 13, 16, 25, 40\}$, the other numbers are constant: $M_y = 4$, $N_x = N_y = 7$.

For $N = 6$ we have compared the far field with a reference result that was computed by using the finite element method implementation JCMWave [134], with an estimated error of $3 \cdot 10^{-3}$. An $\mathcal{L}^2$ difference of $3 \cdot 10^{-3}$ was observed between the present algorithm and the results computed with FEM. The optimizations discussed in Chapter 10 were also employed.

In Figure 9.13(a), (b) and (c) the absolute value of the x, y and z-components of the electric field are shown for $N = 36$ repeating elements, for a cross section computed at the layer interface. In the middle, the electric field around every block looks very similar. At the end, some waves scattering from the edges are visible. To get a better idea of the far-field scattering from this setup, we show the far field for $N = 6$ and $N = 36$ in Figure 9.14(a) and (b), respectively. The zeroth and first diffraction orders can be clearly recognized as indicated.

With an increasing number of repeating elements, the maxima become more localized and the peak amplitude increases as can be seen in Figure 9.14. To see how the far field converges when the number of repeating elements increases, we look at the ratio between the first and zeroth order maximum, since that converges for large N. In Figure 9.15(a) the convergence of this ratio is clearly visible. Although we do not know to which value this ratio converges, we have found a fitted value of 0.4977 as the result for an infinite grating. In Figure 9.15(b) the deviation of E_1/E_0 from the fitted value is shown for varying N. In Figure 9.15(c) the k_x value of the first-order maximum is shown. From the dimensions of the problem it can be deduced that for the infinite case the maximum should be located at $k_x = 425/700k_0 = 8.976 \cdot 10^6\text{nm}^{-1}$. The convergence to this value for a large number of repeating elements is illustrated in Figure 9.15(d). We conclude that at around 10 repeating elements the difference between the far-field scattering from a finite or infinite grating with these dimensions becomes smaller than the accuracy of the simulations, i.e. $3 \cdot 10^{-3}$.

We conclude by remarking that the computation of the largest scattering case was done on a laptop with an i7-4600U processor in less than 15 minutes, for a simulation domain larger than $60\lambda \times 5\lambda \times \lambda/5$. The algorithm required a total memory of 13 Gb.



(a)

(b)

(c)

(d)

Figure 9.10: The absolute error in the discretization of the right part of the complex plane spectral integration path for the propagation function $\exp(-\gamma z)$ over range z with simulation width W. These errors should be compared to $1.0 \cdot 10^7 \cdots 1.4 \cdot 10^7$, which is the $\mathcal{L}^2$ norm of the propagation function itself. (a) Without oversampling, (b) with a factor 4 oversampling, i.e. 4 times as much Gabor coefficients are computed. (c) The absolute error in a $\log_{10} Z$, $\log_{10} W$. The line is a fit to the maximum values in the direction $Z \propto W^2$. (d) The absolute discretization error along the line for different oversampling factors

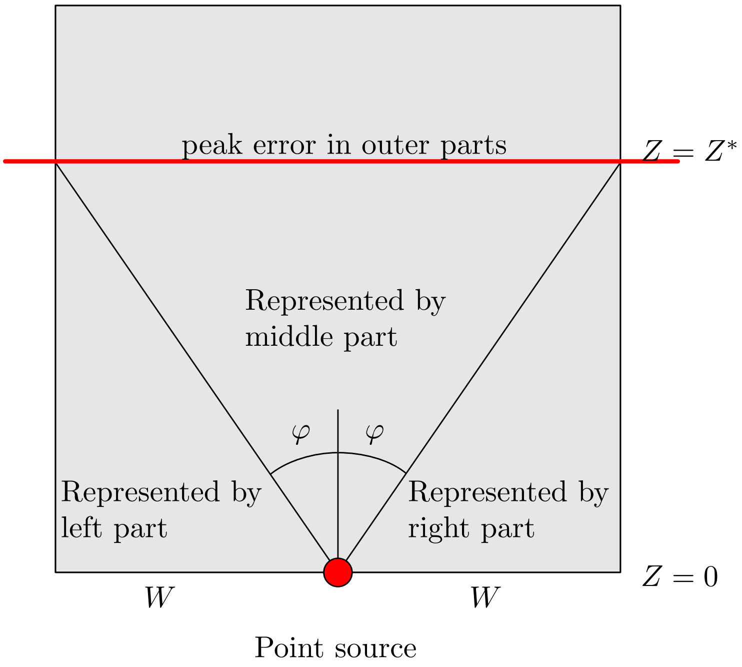


Figure 9.11: Illustration to clarify which part of the representation dominates which part of the simulation domain. Because of the move into the complex domain, the transition between the regions is smoother than implied in the picture.

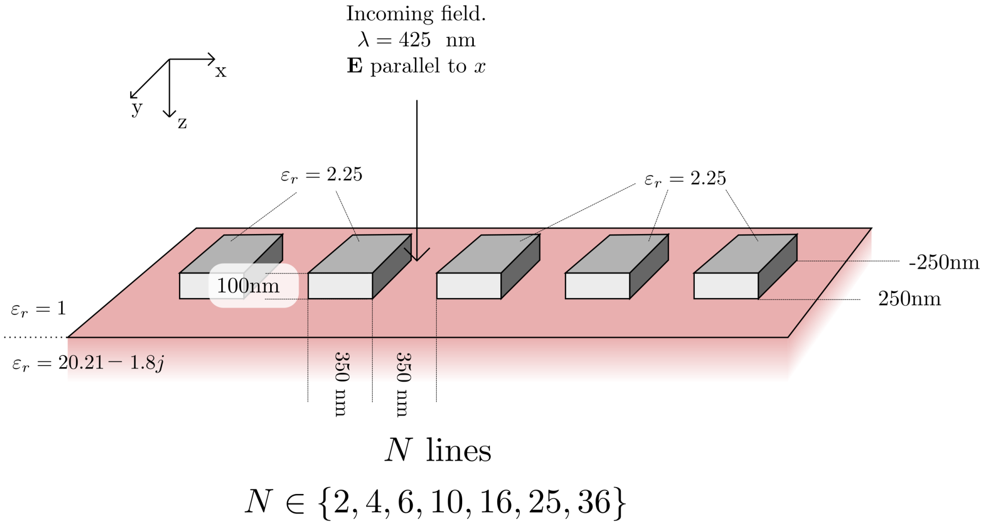


Figure 9.12: The scattering setup. The number of repeating N elements is varied.

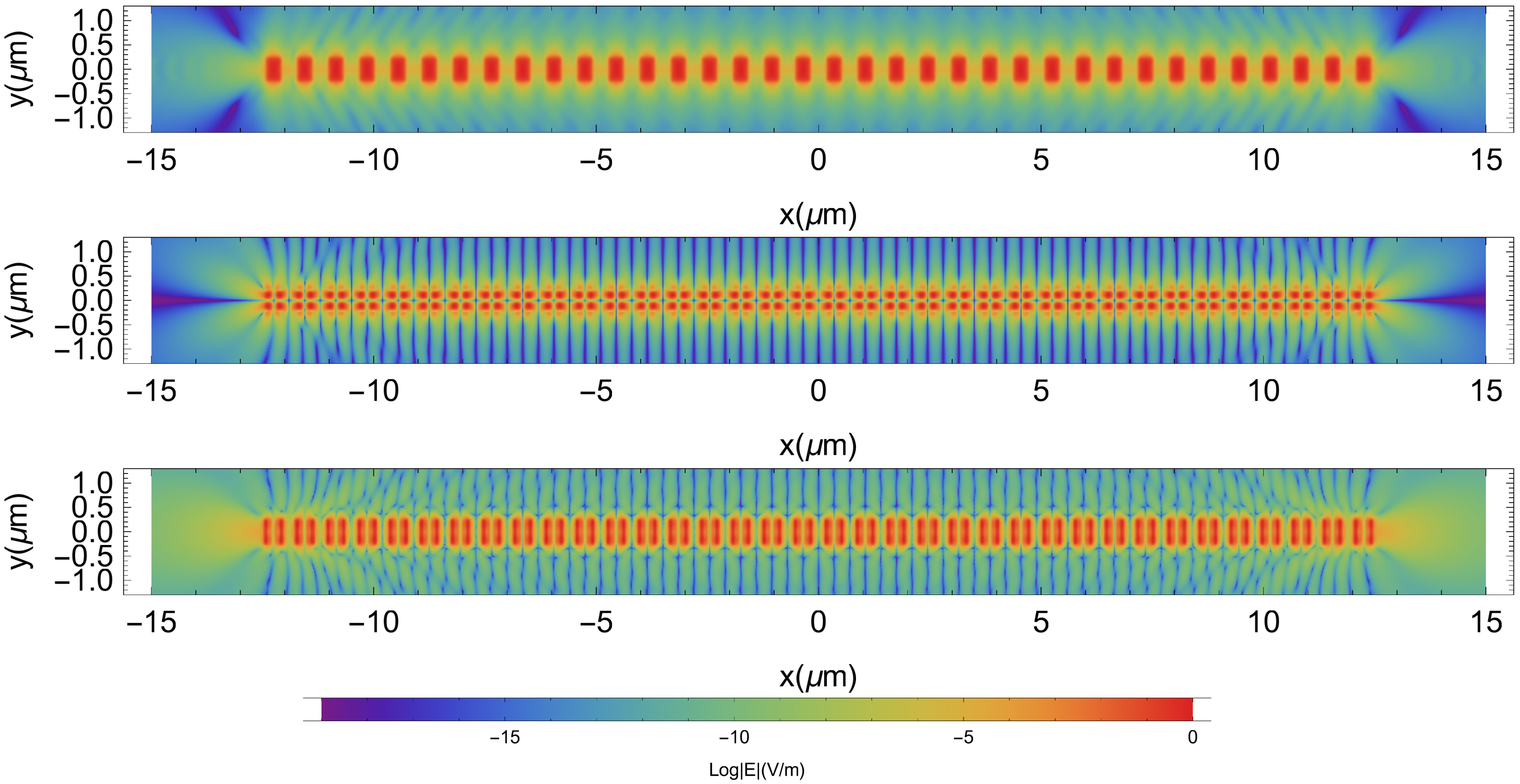


Figure 9.13: The value of the x, y and z components of the electric field $\log|E_x(x,y)|$.

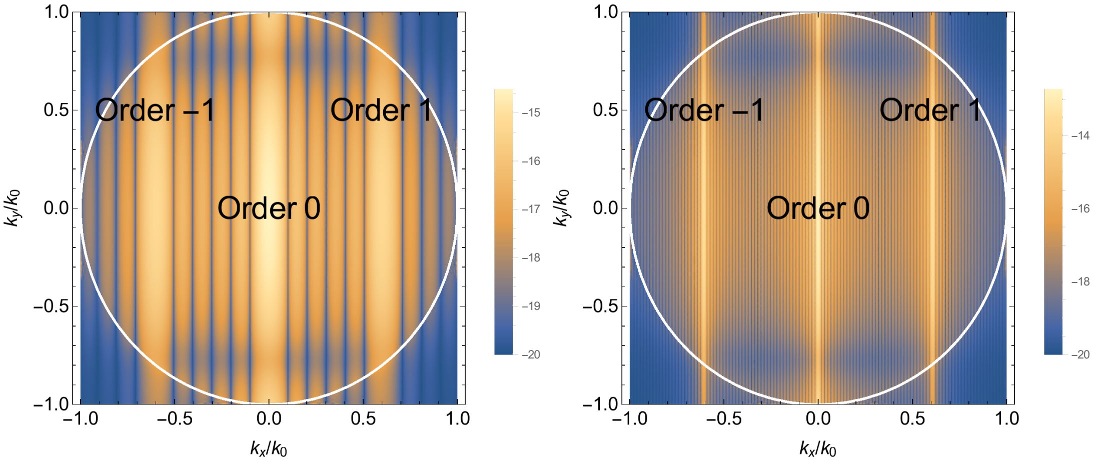


Figure 9.14: The $\log |E_\alpha(k_x, k_y)|$ for the radiating wavenumbers in the Ewald circle $\mathbf{k}^2 < k_0^2$. In (a) for $N = 6$ repeating elements and in (b) for $N = 36$ repeating elements.

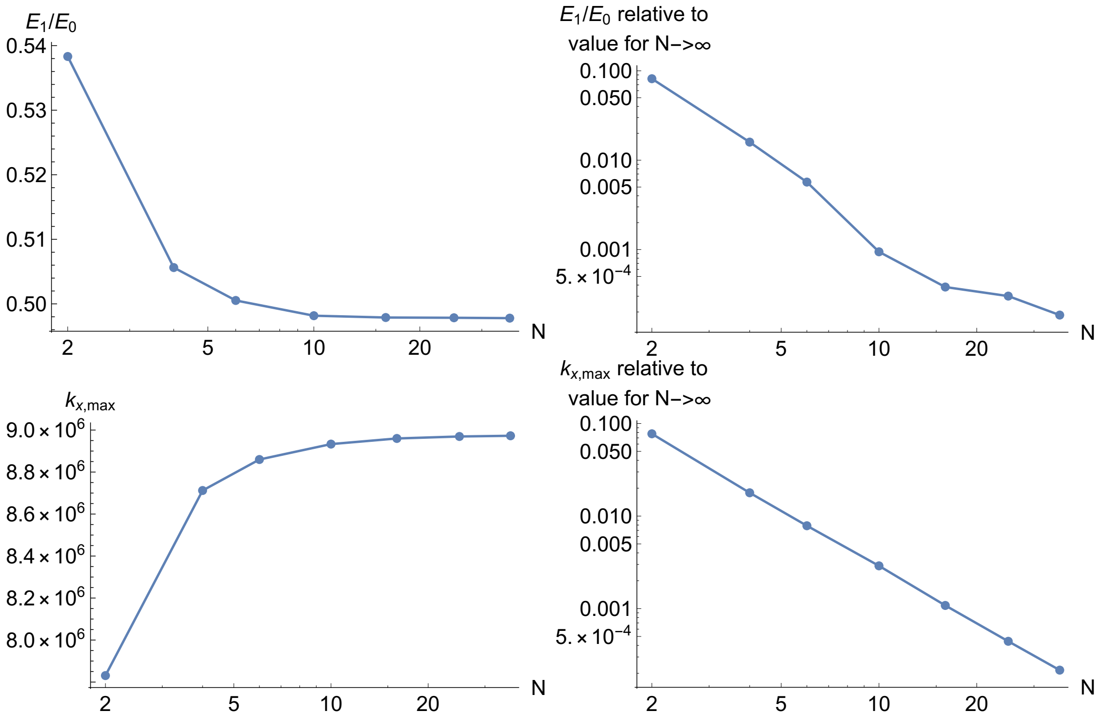


Figure 9.15: (a) The ration between the zeroth and first order maxima, E_1/E_0. (b) The relative difference of this ratio to a fit for an infinite number of repeating elements ($N \to \infty$), given by $|E_1/E_0 - 0.4977|/0.4977$. (c) The k_x coordinate of the maximum. (d) The relative difference of this coordinate to the value for an infinitely repeating scatterer ($N \to \infty$), given by $8.976 \cdot 10^6 \mathrm{nm}^{-1}$.

Chapter 10

Fast operations for a Gabor-frame-based integral equation with equidistant sampling [1]

10.1 Abstract

We consider the computation time of a 3D Gabor-frame based spatial spectral integral equation solver for scattering from dielectric objects embedded in a multi-layer medium. Based on the Gabor frame, a new set of basis functions is proposed, together with a set of equidistant Dirac-delta test functions. Using this construction, we approximate the operations of Fourier transformation and pointwise multiplication by a method that is significantly faster than the original method. A numerical example is included where the computation time is reduced by a factor of 15, while preserving accuracy.

10.2 Introduction

Spatial spectral solvers for computing the scattering from dielectric objects embedded in a multilayered medium are presented in [135, 147, 152] (Chapters 6-8). These methods rely on the Gabor frame and a discretization in both the spatial and the spectral domain. In the spectral domain, a deformation to a complex manifold is employed on which the Green function is smooth enough to allow for a Gabor-frame representation. The main advantage of the Gabor frame is that the Fourier transformation is represented analytically by a simple transposition of the Gabor-coefficient matrix. The downside of the Gabor frame is that the operation of multiplication is represented through an operator that contains a large number of small-sized FFTs and considerable overhead in reordering coefficients [118].

[1] This chapter was submitted as an article for the journal IEEE Antennas and Wireless Propagation Letters [160].

In general, a fast matrix-vector product[2] requires a suitable discretization method. In particular, this method should allow for both a rapid multiplication of functions and a rapid Fourier transformation to and from the spectral domain. A fast multiplication can be achieved when product of basisfunctions are bi-orthogonal to the test functions, since that allows for a coefficient-wise multiplication. Conversely, when they are not bi-orthogognal, each multiplication between two basis functions requires testing with multiple test functions, which is undesirable. The products of Gabor frame functions and the dual Gabor frame functions as test function are not bi-orthogonal. Secondly, a fairly rapid Fourier transformation can be achieved when the test functions are all identical and spaced uniformly, since that allows the use of FFTs. Additionally, the basis functions should also decay sufficiently fast, to allow for a truncation to a small region in the spatial and spectral domain.

We show how the electric field and contrast current density can be represented by a set of basis functions that are related to a Gabor frame and a set of Dirac-delta test functions related to the same Gabor frame that together satisfy the above conditions. In a numerical example, we demonstrate the decrease in computation time and compare results for both discretization methods with comparable accuracy.

10.3 Gabor frame - definitions

We define the Gabor frame [99], starting with the Fourier transform

$$\hat{\varphi}(k) = \int_{-\infty}^{\infty} dx\, \varphi(x) e^{-jkx}. \tag{10.1}$$

For the Gabor frame we follow the definition in [100], i.e.

$$g_{mn}(x) = g(x - m\alpha X) e^{j\beta K n x}, \tag{10.2}$$

with $m, n \in \mathbb{Z}$ and where we use the window function

$$g(x) = 2^{\frac{1}{4}} e^{\left(-\pi \frac{x^2}{X^2}\right)}. \tag{10.3}$$

Here, $X = \frac{2\pi}{K}$ is the spacing of the window functions in the spatial domain for an exact frame. In this article, we assume a rational oversampling with $\alpha = \beta = \sqrt{p/q}$, and choose $p = 2$ and $q = 3$. The dual window, $\eta(x)$, is calculated with the aid of the Moore Penrose pseudo-inverse and the method described in [99, 100]. When we have chosen a frame and a dual window, we can calculate the Gabor coefficients of a (square-integrable) function f as

$$f_{mn} = \int_{\mathbb{R}} dx\, \eta_{mn}(x) f(x) \tag{10.4}$$

[2] In the sence of Section 3.4

and function values from Gabor coefficients via

$$f(x) = \sum_{m,n \in \mathbb{Z}} f_{mn} g_{mn}(x). \tag{10.5}$$

In practice, the number of Gabor coefficients is truncated to $m \in \{-M, \cdots, M\}$ and $n \in \{-N, \cdots, N\}$, which yields a total number of $2L + 1$ coefficients. By taking the Fourier transformation of the frame function $g_{mn}(x)$, a spectral frame is defined as

$$\hat{g}_{nm}(k) = \hat{g}(k - n\beta K) e^{j\alpha X mk} e^{-2\pi j \alpha \beta mn}. \tag{10.6}$$

10.4 Basis functions

10.4.1 Representation using lists

In [100], Bastiaans describes the fast Gabor transformation $\mathcal{B}$, that calculates the Gabor coefficients of a function from an uniformly sampled function. This algorithm can also be inverted to obtain $\mathcal{B}^{-1}$, to calculate a list of uniformly sampled function values from a set of Gabor coefficients. The uniformly sampled lists will be denoted in boldface. Since the lists are defined in connection with a particular Gabor frame, the sampling is restricted, i.e. the sampling operator $\mathcal{S}$ samples a function according to that Gabor frame

$$\mathbf{f} = \mathcal{S} \circ f = \{f(\ell \Delta_x), \ell \in -L, \cdots, L\}, \tag{10.7}$$

with L as defined above and where Δ_x depends on the parameters defining the Gabor frame via

$$\Delta_x = \frac{X}{\beta(2N+1)}. \tag{10.8}$$

10.4.2 Shape of the basis functions

A continuous function is approximated by a set of weighted basis functions. In the context of Gabor frames, the most obvious choice for a basis are the frame functions of Eq. (10.2), which were used in [135, 147, 152, 118]. However, here we will not use the Gabor frame directly as a basis. Instead, we derive the basis functions from the fast Gabor transformation $\mathcal{B}$ of entries to uniformly sampled lists. Since the list representation is tied to the Gabor frame via $\mathcal{B}$, we can compute the Gabor coefficients for each list. The coefficients (in $\ell^2(\mathbb{R})$) are related to a continuous function (in $\mathcal{L}^2(\mathbb{R})$) via Eq. (10.5). Therefore, we define basis functions corresponding to the list $\mathbf{b}_i = \{0, \cdots, 1, \cdots, 0\}$ with a one at position i, that can be found by means of

$$b_i(x) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} g_{mn}(x) \{\mathcal{B} \circ \mathbf{b}_i\}_{mn}, \tag{10.9}$$

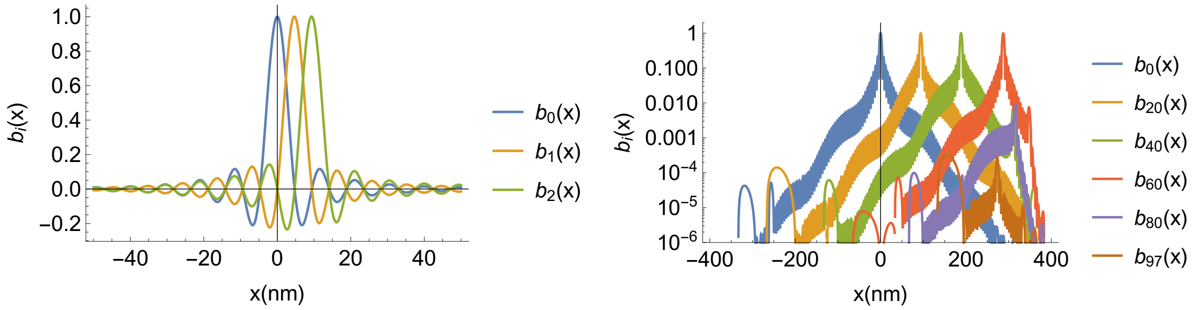


Figure 10.1: Several basis functions for a Gabor frame with $X = 50\text{nm}$, $M = 7$, $N = 6$, $L = 97$.

where the circle denotes the application of an operator.

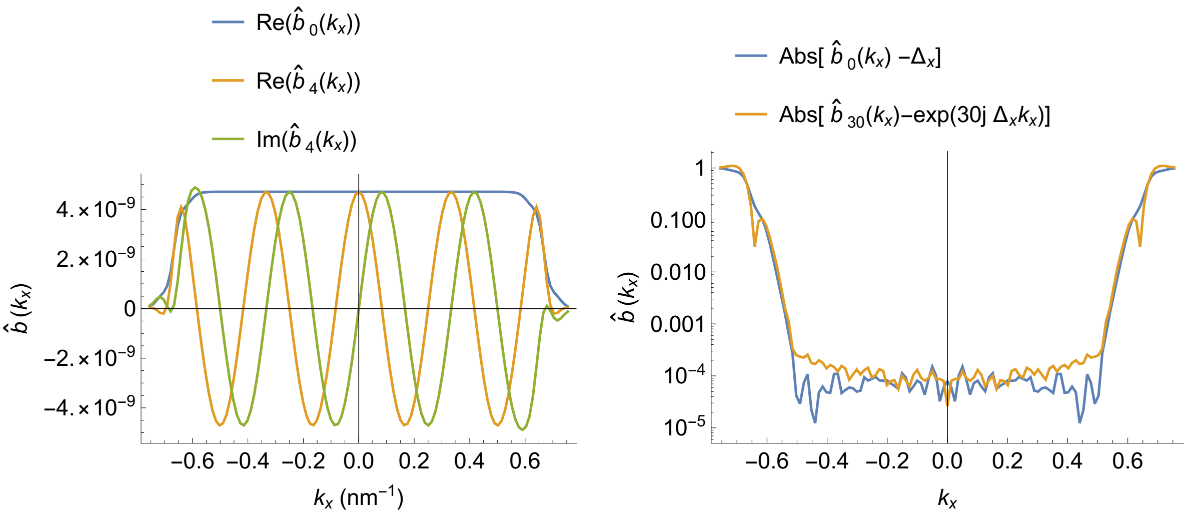
In Fig. 10.1, we have plotted several of these basis functions. Several interesting features can be observed about these functions. The basis functions have zeroes at the Δ_x grid, except at $x = i\Delta_x$, where they are one for $i < pN/q$. The reason is that they are produced by $\mathcal{B}$ and $\mathcal{B}^{-1}$ on the $\mathbf{b}_i$ list defined for the Δ_x grid. Another observation is that, for high index i, the basis functions are very small. This is caused by the redundancy in the Gabor frame. We also observe that, although these basis functions look similar, they are not one function that is merely shifted in position. There are subtle differences between the basis functions. The final observation that we mention is that $b_i(x)$ resembles a sinc function, which decays slowly. However, outside the simulation domain, $b_i(x)$ decays much faster than the sinc function. In Fig. 10.2(a) and (b), we have also plotted several basis functions in the spectral domain, where it is clearly visible how these basis functions $\hat{b}_i(k_x)$ resemble truncated complex exponentials, since they are produced by functions resembling Dirac-Delta distributions. It is interesting to notice that the truncation has a smooth transition, so the b_i functions rapidly decay to zero at the ends of the simulation domain in the spatial domain.

10.4.3 Testing functions and inner products

In the Gabor-frame discretization we used the dual Gabor frame as test functions, which works well, since the dual Gabor frame is dual to the Gabor frame with respect to the $\mathcal{L}^2(\mathbb{R})$ norm. For the set of $b_i(x)$ basis functions we use Dirac delta test functions on the Δ_x lattice, a set which is dual with respect to the $\mathcal{L}^2(\mathbb{R})$ norm as well.

An $\mathcal{L}^2(\mathbb{R})$-based inner product was employed for the Gabor-frame based method. The computation of an $\mathcal{L}^2$-based inner product was not used here, since all basis functions are slightly different, i.e. they are not simply shifted copies of each other. This means that $\langle b_i | b_j \rangle_{\mathcal{L}^2(\mathbb{R})}$ has a different value for each i and j and there is for example no translation symmetry in the sense that $\langle b_i | b_j \rangle_{\mathcal{L}^2(\mathbb{R})} \neq \langle b_{i+m} | b_{j+m} \rangle_{\mathcal{L}^2(\mathbb{R})}$.

With the Dirac-delta testing procedure, the test functions are $t_i(x) = \delta(x - i\Delta_x)$. As we mentioned before, the basis functions are such that that $\langle b_i | t_j \rangle_{\mathcal{L}^2(\mathbb{R})} = \delta_{ij}$, i.e. they are bi-orthogonal for $i, j < \ p/qL$. Since $b_i(x)$ and $t_j(x)$ are biorthogonal, we choose the ℓ^2



(a) (b)

Figure 10.2: (a) The Fourier transform of basis functions of the frame in Figure 10.1. (b) The difference between two spectral basis functions and their corresponding complex exponentials.

inner product

$$\langle \mathbf{f}|\mathbf{h}\rangle = \Delta_x \sum_{i=-L}^{L} f_i \overline{h}_i \approx \langle \mathbf{f}|\mathbf{h}\rangle_{\mathcal{L}^2} = \int_{-\infty}^{\infty} dx f(x)h(x) \tag{10.10}$$

with $\mathbf{f}$ and $\mathbf{h}$ discretized from smooth functions $f(x)$ and $h(x)$, respectively by Eq. (10.7). ¿ Clearly, Eq. (10.10) is equivalent to numerically evaluating the integral in the $\mathcal{L}^2$ inner product between $f(x)$ and $h(x)$ by N equidistant samples. ¿ Therefore, Eq. (10.10) converges to the $\mathcal{L}^2$ inner product when N increases.

10.5 Operations

10.5.1 Multiplication

First, we emphasize that the multiplication operation is non-linear. Consequently, when two functions can be represented well in a Gabor frame, their product is not necessarily well represented in the same Gabor frame. The reason for this is that a spatial multiplication is equivalent to a spectral convolution. For that reason the product of two functions potentially has twice the spectral support of the original functions.

An approximation has to be made to fit the product in the space spanned by the Gabor-frame of the original functions. In the Gabor-frame formulation [118] an (almost) exact multiplication was implemented, but in the end the spectral range is truncated. This is equivalent to an exact multiplication followed by testing with a finite number of test functions.

We apply this procedure to find a multiplication operation with the basis $b_i(x)$, and test functions $t_i(x)$. The procedure for multiplying lists $\mathbf{f}$ and $\mathbf{g}$ then becomes

$$(\mathbf{fg})_i = \int_{x=-\infty}^{\infty} dx \sum_{j,k} \mathbf{f}_j \mathbf{g}_k b_j(x) b_k(x) t_i(x) = \mathbf{f}_i \mathbf{g}_i, \tag{10.11}$$

where we used the property that $b_i(j\Delta_x) = \delta_{ij}$. We would like to emphasize that it is the choice of the test function that yields this simple form of multiplication.

10.5.2 Fourier transformation

The advantage of a uniformly sampled list-based approach is that multiplication is a very fast operation. However, now the Fourier transformation is slower. It is possible to implement the Fourier transformation by succesively applying a fast Gabor transformation $\mathcal{B}$, a Fourier transformation $\mathcal{F}$ on Gabor coefficients and then an inverse Gabor transformation $\hat{\mathcal{B}}^{-1}$, i.e.

$$\hat{\mathbf{f}} = \hat{\mathcal{B}}^{-1} \circ \mathcal{F} \circ \mathcal{B} \circ \mathbf{f}. \tag{10.12}$$

The main drawback is the use of the relatively slow operations $\mathcal{B}$ and $\hat{\mathcal{B}}^{-1}$. For a more optimized method, we exploit the fact that without truncation

$$\sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} g_{mn}(x') \eta_{mn}(x) = \delta(x - x'), \tag{10.13}$$

and therefore, using Eq. (4.31), we can write its Fourier transformation in x' as

$$\sum_{m,n=-\infty}^{\infty} \hat{g}_{nm}(k_x) e^{2\pi j \alpha\beta mn} \eta_{mn}(x) = e^{-jk_x x}. \tag{10.14}$$

Now, the Fourier transformation of a function f can be approximated by

$$\begin{aligned} \hat{f}(k_x) &= \int_{-\infty}^{\infty} dx \sum_{m,n=-\infty}^{\infty} f(x) \hat{g}_{nm}(k_x) e^{2\pi j \alpha\beta mn} \eta_{mn}(x) \\ &\approx \Delta_x \sum_{m,n=-\infty}^{\infty} \sum_{\ell=-L}^{\ell=L} f(\ell\Delta_x) \eta_{mn}(\ell\Delta_x) \hat{g}_{nm}(k_x) e^{2\pi j \alpha\beta mn}. \end{aligned} \tag{10.15}$$

In this expression, we recognize the Gabor transformation, Eq. (10.4), as the integral over x, which is replaced by a sum in the second line. This approximation holds when the sampling in Eq. (10.7) approximates f well. This discretized Gabor transformation equals the operator $\mathcal{B}$. The fact that we wrote $\hat{g}_{nm}(k_x)e^{2\pi j\alpha\beta mn}$ instead of $g_{mn}(x)$ exactly represents the Fourier transformation operator $\mathcal{F}$. And finally, the summation over m, n represents the inverse Gabor transformation of Eq. (10.5). The integral in Eq. (10.15) is evaluated on the uniform grid $k_x \in \{-L\Delta_k, \cdots, L\Delta_k\}$ and therefore the m, n summation equals $\hat{\mathcal{B}}^{-1}$, i.e. the inverse Gabor transformation operator in the spectral domain. Here $\Delta_k = K/\alpha(2N+1)$ is the spectral-domain counterpart of Δ_x. Hence we identified all operators of Eq. (10.12) in Eq. (10.15) and both are equal up to the discretization error in $\mathbf{f}$.

When we apply the discretized version of Eq. (10.14) to the second line of Eq. (10.15), we can write

$$\hat{\mathbf{f}}_m = \sum_{n=-L}^{L} \mathbf{f}_n e^{-j\Delta_x\Delta_k mn} = \sum_{n=-L}^{L} \mathbf{f}_n \exp(-2\pi jmn\frac{p}{q(2L+1)}), \tag{10.16}$$

which looks similar to a discrete Fourier transformation. However, it is the oversampling factor p/q makes the difference.

In case $(2L+1)/p \in \mathbb{N}$, this can be calculated as the Fast Fourier Transform (FFT) of size $q/p(2L+1)$ of $\mathbf{f}^\dagger$ in

$$\mathbf{f}^\dagger_n = \sum_{n=1}^{N} \mathbf{f}_{(n \mod (2L+1)q/p)}. \tag{10.17}$$

Since this FFT is of smaller size than the list $\mathbf{f}$, it is extended to the full size, $2L+1$, by periodically expanding $\hat{\mathbf{f}}^\dagger$.

There is a subtle difference between Eq. (10.12) and Eq. (10.16). The difference is that in Eq. (10.13) an infinite sum is taken over m and n. When this sum is truncated, as is done in Eq. (10.12), this yields a good approximation of a Dirac delta function only for a part of the domain of x and x' that is as wide as the region where the $b_i(x)$ peak is close to one (see Fig. 10.1(b)). Hence, this sets the coefficients $\mathbf{f}_i$ for large $|i|$ to zero. The main cause of differences between the Gabor-based Fourier transformation and the FFT-based Fourier transformation is the periodic continuation of functions at the edges of the domain in the region where the basis functions are close to zero.

To demonstrate the range over which the approximated Gabor transform is accurate, we apply the Fourier transformation operator on the modulated and shifted Gaussian pulse $\exp\{(x-x_0)^2 + jk_0x\}$. This pulse function is localized around (x_0, k_0) in spatial-spectral plane. We compare the discretized pulse function with a pulse function transformed to and from the spectral domain via Eq. (10.16). In Figure 10.3, the relative error is shown as a function of the location of the pulse (x_0, k_0) in the $x-k$-plane. Clearly, there is a four-digit accuracy over most of the domain, which corresponds to the accuracy up to which the dual window $\eta(x)$ was computed. For large k_0 and x_0 the accuracy is lower, because of the oversampling and the periodic continuation of functions. Therefore, we conclude that modulated Gaussian pulses can be accurately transformed back and forth from the spectral domain and since a discretization based on Gabor frames that consist of Gaussian pulses delivers accurate results, this method will be accurate as well.

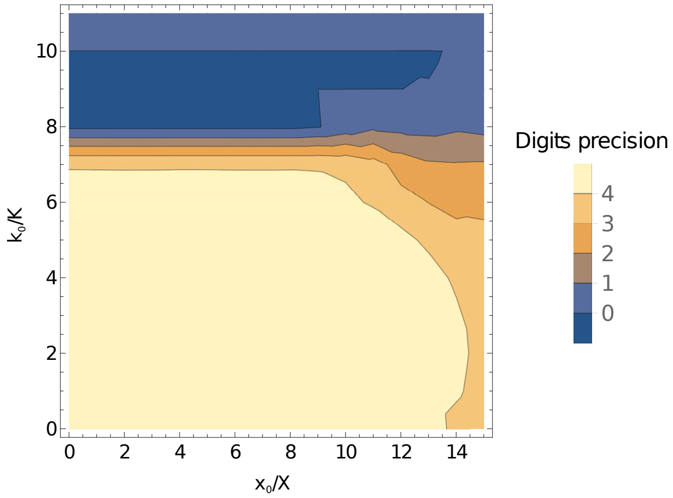


Figure 10.3: The relative error for the Fourier transform in a Gabor frame with $M = 15$ and $N = 13$ for a Gaussian pulse function located at a position in the spatial-spectral $x - k$-plane.

10.6 Application to a three-dimensional scattering problem

We exchanged the Gabor-frame discretization in [152] (Chapter 8) with the proposed list-based discretization. The Green function and the contrast function are not continuous enough to discretize by simply sampling them according to Eq. (10.7). Note that list-based representations of these functions are only calculated during the initialization of the algorithm, not during each iteration of the iterative solver. Therefore, the computation time of these lists is not very critical. We find a better approximation for these functions by taking the Gabor coefficients as they were calculated in [152] and transform them to the list-based representation through $\mathcal{B}^{-1}$.

The convergence of this method was tested on the first test case of [152], a dielectric cube embedded in a layered medium. Timing and accuracy results are displayed in Figure 10.4 for the two methods of Fourier transformation in Eq. (10.12) and in Eq. (10.16). Both employ the multiplication in Eq. (10.11). Clearly, the latter Fourier transformation requires much less computation time, while the results are very close in terms of accuracy. In the example, the difference between the results with different Fourier transformations is smaller than the error from the simulations itself, indicated by the circles in Figure 10.4. This implies that it is the discretization that governs the accuracy of the result, not the type of Fourier transformation that is being applied.

To show the applicability of this method to a larger problem, we compute the far field due to scattering from a finite grating consisting of 12 bars of relative permittivity $\varepsilon_r = 2.25$ in vacuum placed on a half space with $\varepsilon_r = 20.21 - 1.8j$ with a normal incident plane wave of unit amplitude as depicted in Figure 10.5(a). A Gabor frame with a window width of 1 micron was used that was truncated at $M = 7$ and $N = 16$ in both the x

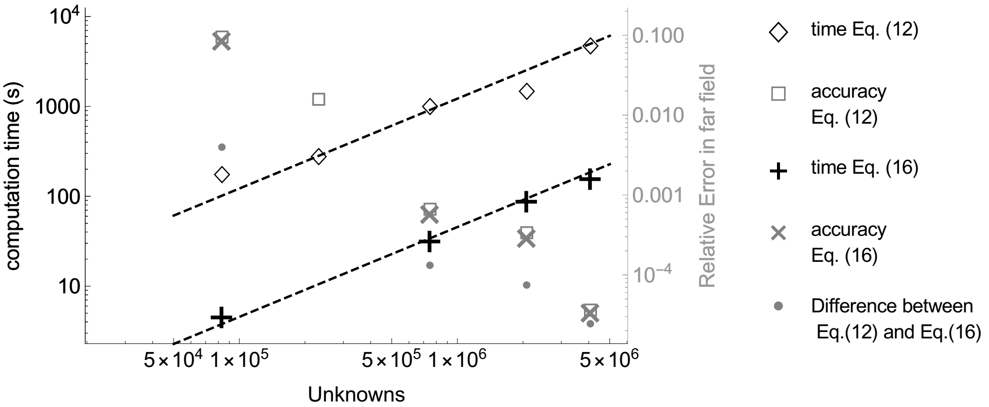
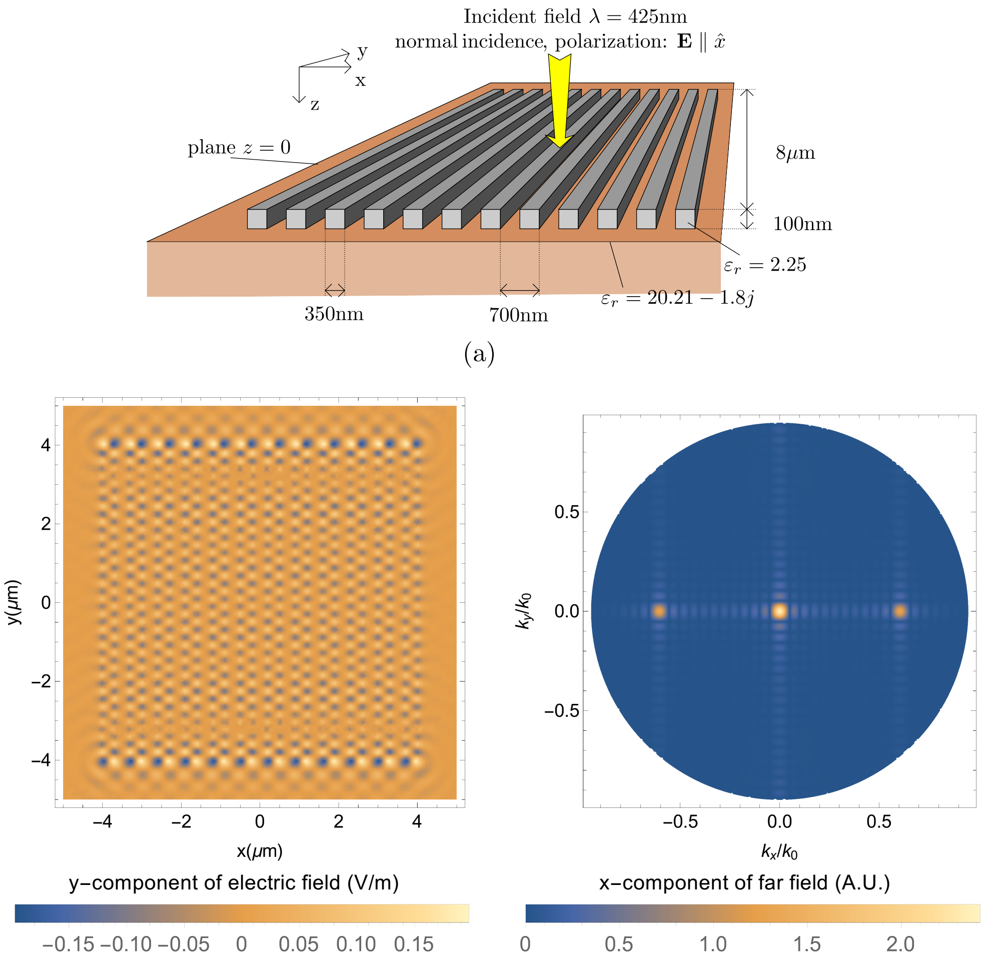


Figure 10.4: Scaling of the computation time and accuracy with the number of unknowns for different samplings in the xy plane, realized by truncating the Gabor frame at $N \in \{1, 4, 7, 10\}$. The relative error is computed against a JCMWave [134] reference calculation as described in [152] (Chapter 8). The dashed lines indicate linear scaling of computation time.

and y directions. In Fig. 10.5(b) and (c), the near and far field are depicted. In the far field, the relative $\mathcal{L}^2(\mathbb{R}^2)$ difference between results obtained with Eq. (10.16) and the algorithm in [152] was estimated at settings where the latter reached a 3×10^{-3} relative error. The computation time required with the present method was 105 minutes, using the Fourier transform in Eq. (10.12) it was 54 hours, and the time required by employing a Gabor-based multiplication [135] was estimated to be longer than 40 days based on the time required for a single multiplication. All computation times pertain to a single core of a 3.1 GHz Intel Xeon E5 2687W processor.

Since our complex path formulation works with a Gabor frame, and when Gabor-frame functions (modulated gaussians) can be accurately transformed with this newly developed Fourier-transform it will work. So this discretization should hold for discretizing the complete EM-scattering problem from dielectric objects. To show this, we show the electric field generated by a line source $J(x, z) = \delta(z)\Pi(x/100\mathtt{nm})$ observed at $z = 250$nm and $\lambda = 245$nm calculated using frame r' in Figure 10.6. The relative error in this plot is a bit larger than 10^{-3}. The reason is that a coarser spectral sampling is used compared to the direct Gabor coefficient multiplication. With this algorithm oversampling is not possible in the spectral domain, therefore the sampling in the spectral domain is much coarser than with the Gabor coefficient method. However, choosing a large oversampling has a larger penalty for the number of operations than simply choosing a larger spatial simulation domain. We verified that choosing the simulation domain twice as large does indeed lower the relative error below 10^{-3}.

We have benchmarked the calculation-time within Mathematica for a Fourier transform with Gabor frame r. The FFT-based method was 15 times faster with our Mathematica based algorithm, however, this is not a good indicator for performance in more optimized programming languages such as Fortran.



(a)

(b)

(c)

Figure 10.5: (a) A scattering setup of 12 dielectric lines on a dielectric halfspace. In (b) y-component of the electric field is plotted at $z = 20$ nm. In (c) the x-component of the far field is given that scatters back.

10.7 Conclusion

A point-wise multiplication and FFT-based Fourier transform operation were proposed based on a discretization by Gabor frames. An improvement to the algorithm in [152] is proposed that is at least 15 times faster for two represntative computational examples. Numerical evidence was shown that the approximation error is negligible compared to the discretization error.

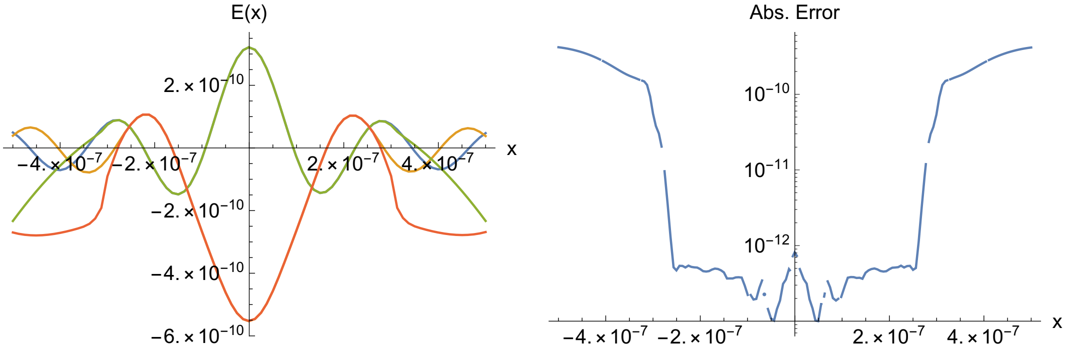
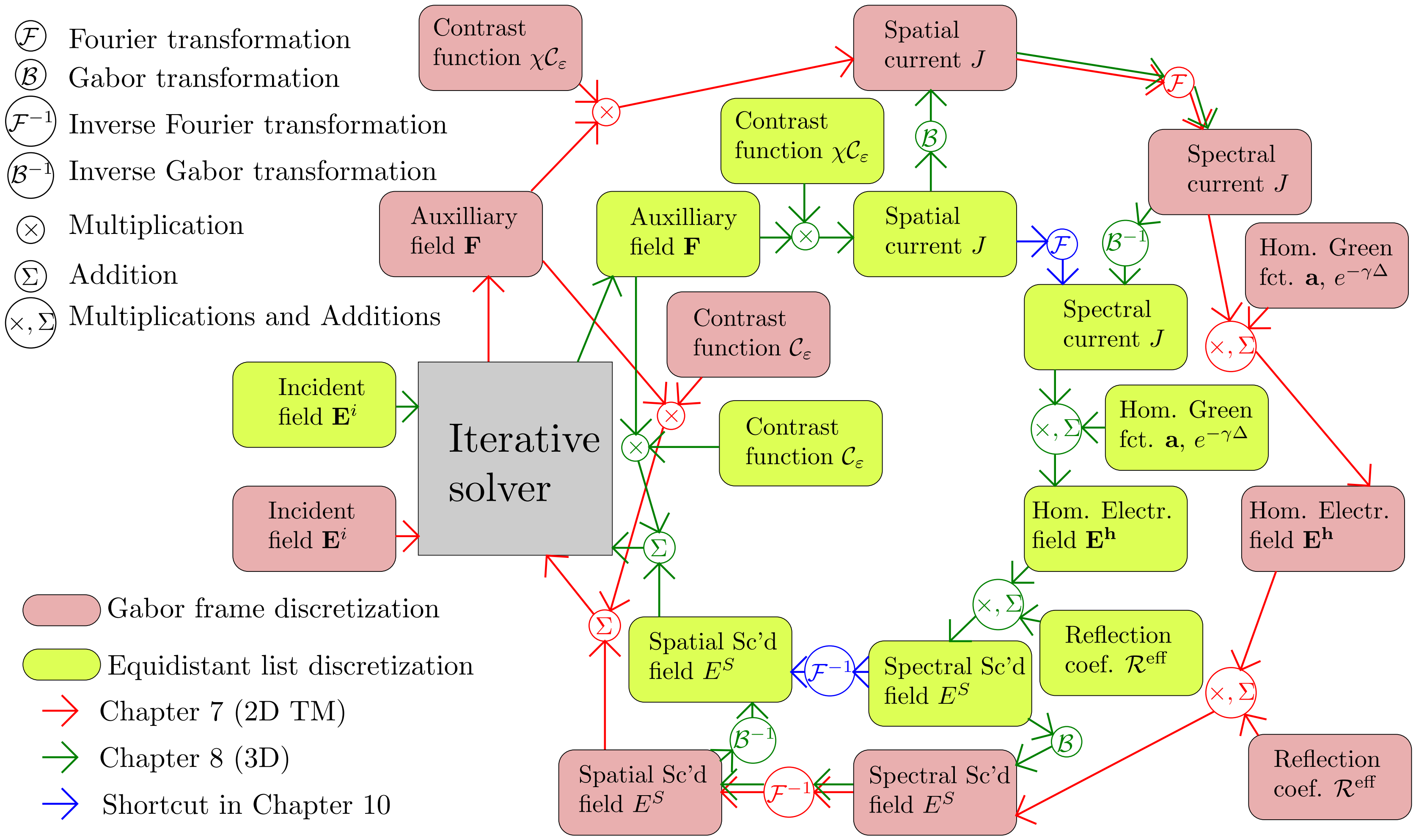


Figure 10.6: Performance of a caculation using frame r'. In (a) Blue and yellow are the real and imaginary part of numerical integration. Green and orange are the results generated with the presented method.

Addendum

To clarify the differences between the algorithm for 2D TM scattering in Chapter 7, the algorithm for 3D scattering in Chapter 8, and the algorithm presented in this chapter we present a schematic overview of the employed discretizations in Figure 10.7. The algorithm in Chapter 7 only employs the Gabor-frame discretization. In the description of this algorithm different names are used for the contrast function, $\mathcal{L}_\chi$ corresponds to $\mathcal{C}_\varepsilon$ and $\mathcal{M}_\chi$ corresponds to $\chi\mathcal{C}_\varepsilon$. The algorithm in Chapter 8 uses an equidistant sampled list for all operations except the Fourier transformation and the inverse Fourier transformation. A detour is taken via Gabor transforms to apply the Fourier transformation on Gabor coefficients. Initially, all components of the Green function, the reflection coefficients and contrast functions are all discretized via the Gabor frame. Afterwards, an inverse Gabor transform is employed to transform them to lists with equidistant sampling. The only thing that has changed in this chapter is that the Fourier transformation is carried out directly on the lists with equidistant samples with the method described in Section 10.5.2. The method in this chapter does not follow the detour via the Gabor transformations, it employs the direct Fourier transformation that is described in this chapter.



$\mathcal{F}$ Fourier transformation
$\mathcal{B}$ Gabor transformation
$\mathcal{F}^{-1}$ Inverse Fourier transformation
$\mathcal{B}^{-1}$ Inverse Gabor transformation
$\times$ Multiplication
Σ Addition
$\times, \Sigma$ Multiplications and Additions
Gabor frame discretization
Equidistant list discretization
Chapter 7 (2D TM)
Chapter 8 (3D)
Shortcut in Chapter 10
Contrast function $\chi\mathcal{C}_\varepsilon$
Spatial current J
$\mathcal{F}$
Spectral current J
Contrast function $\chi\mathcal{C}_\varepsilon$
$\mathcal{B}$
Auxilliary field $\mathbf{F}$
Auxilliary field $\mathbf{F}$
Spatial current J
$\mathcal{F}$
$\mathcal{B}^{-1}$
Hom. Green fct. $\mathbf{a}$, $e^{-\gamma\Delta}$
Contrast function $\mathcal{C}_\varepsilon$
Spectral current J
$\times, \Sigma$
Incident field $\mathbf{E}^i$
Iterative solver
Contrast function $\mathcal{C}_\varepsilon$
Hom. Green fct. $\mathbf{a}$, $e^{-\gamma\Delta}$
Incident field $\mathbf{E}^i$
Σ
Hom. Electr. field $\mathbf{E^h}$
Hom. Electr. field $\mathbf{E^h}$
$\times, \Sigma$
Σ
Spatial Sc'd field E^S
$\mathcal{F}^{-1}$
Spectral Sc'd field E^S
Reflection coef. $\mathcal{R}^{\text{eff}}$
$\times, \Sigma$
$\mathcal{B}^{-1}$
$\mathcal{B}$
Spatial Sc'd field E^S
$\mathcal{F}^{-1}$
Spectral Sc'd field E^S
Reflection coef. $\mathcal{R}^{\text{eff}}$

Figure 10.7: A schematic description of the steps required in the algorithms of Chapter 7, Chapter 8 and this chapter. By Sc'd field, the scattered field is meant.

Chapter 11

Alternatives to the discretization and the integration path

11.1 Motivation

Although our objective of building a fully functional solver for three-dimensional scattering from finite dielectric objects in a multilayered dielectric medium was reached, we show that there are alternative approaches available for some of the choices that were made during the development. Most of the development time was spent in a rush towards a fully three-dimensional solver. During the development there were sometimes several options available to continue the development, where it was not clear in advance which method would yield the best results. To make fast progress, the safest option was mostly chosen. In this section we show an alternative to the Gabor discretization of Chapter 4 and an alternative to the complex spectral path in Eq. (6.20) in Chapter 6.

The formulation in Chapter 2 is extended with a PWL-based discretization in the z direction in Chapter 3, a scheme that is particularly efficient for the Green function operator. We take this as the starting point for this chapter. We will present an alternative approach, which treats the transverse direction differently from Chapters 4 to 10.

In the transverse direction(s), a Gabor frame-discretization was proposed in Chapter 4. The reason we chose this discretization was that functions are represented in the spatial and spectral domain simultaneously and that a Fourier transformation is represented by merely a reordering of coefficients. In Section 11.2 a Hermite interpolation scheme will be presented as an example of an alternative discretization. We introduce this discretization to demonstrate that the Gabor frame, with its complicated implementation, is not required for this formulation.

The formulation requires a Fourier transformation of the contrast current density to the spectral domain, a discretization of the Green tensor in the spectral domain and an inverse Fourier transformation of the scattered electric field back to the spatial domain, where the field-material interaction is computed. Since the spatial Green tensor decays slowly for large distances, the spectral Green tensor contains singularities that are hard to discretize in the spectral domain. In Chapter 5, a coordinate scaling was employed to handle the

Green function in a homogeneous medium. In Chapters 6-10, a contour deformation into the complex plane was employed to handle the Green tensor for a multilayered medium. Compared to real spectral coordinates, it takes extra time to compute the transformation to and from the complex-plane spectral path, i.e. the number of Fourier transformations is doubled for 2D simulations and quadrupled for 3D simulations. When the number of Fourier transformations can be reduced, this can be advantageous for the computation time and memory requirements.

Within the current formulation, very large improvements (factors of 10 or more) in computation time cannot be expected from choosing alternative complex paths, since a large part of the computation time is already spent on the FFTs and multiplications for the Green function. However, the algorithm can be made considerably more elegant by finding an alternative to the spectral representation in the nine regions of Chapter 8. In Section 11.3 we focus on alternative integration paths and transformations. The main goal of such an alternative path is to devise a more elegant algorithm. A more elegant algorithm often allows for more flexibility and a faster development of further improvements.

11.2 Alternative transverse basis

11.2.1 General thoughts about the discretization

The discretization in the transverse directions is vital for the efficiency of the algorithms developed in Chapters 5 to 10, since both the accuracy and the computational complexity strongly depend on it. Several properties can be identified that a discretization should exhibit.

1. The number of required basis functions should be small.

2. An accurate implementation of Fourier transformation should be available.

3. A rapid means of Fourier transformation should be available.

4. Fast addition and multiplication operators should be available for the discretized equation.

The analytical Gabor-frame discretization satisfies point 1 for large scatterers, although the broad window functions makes it less efficient for small scatterers. Since the Gabor frame is a simultaneous discretization of the spatial and spectral domain, the second and third point are very well satisfied. The largest downside is that it is hampered by a slow multiplication operator in point 4.

The list-based representation of Chapter 10 improves the multiplication time at the cost of a slightly slower Fourier transformation. This considerably reduces the computational complexity, with negligible approximations made.

Although there is a significant difference in computation time, both the aforementioned discretizations are closely related in terms of discretization via a Gabor frame. Entirely different discretizations are also possible. A discrete translation invariance in the basis functions is advantageous, since that often allows the use of FFTs in the numerical implementation. As an example we will consider Hermite interpolation in this section and argue that it also satisfies the four points mentioned above.

11.2.2 Hermite interpolation

The Hermite interpolation [161, Section 2.11] interpolates a function where the function values and (a number of) derivatives are known. In this section we assume that the function is sampled on an equidistant Δ_x lattice and that at each of the lattice points the function value and derivatives are known, such that a function is represented by coefficients f_{nr} given by

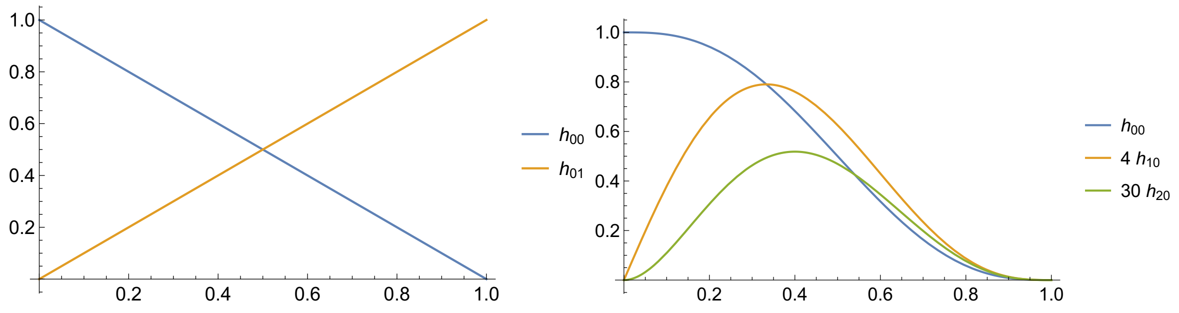
$$f_{nr} = f^{(r)}(n\Delta_x), \tag{11.1}$$

with $n \in \{-N_x, \cdots, N_x\}$ and degree $r \in \{0, \cdots, R-1\}$. To produce the interpolation between two sampling points, e.g. $x = 0$ and $x = \Delta_x$, scaled polynomials $h_{rj}(x)$ are used on this interval defined as

$$h_{rj}(x) = \sum_{\ell=0}^{2R-2} \eta_{rj,\ell} x^{\ell}, \tag{11.2}$$

with indices $r \in \{0, \cdots, R-1\}$ and $j \in \{0, 1\}$. The coefficients $\eta_{rj,\ell}$ are chosen such that all $h_{rj}^{(q)}(0) = 0$ and $h_{rj}^{(q)}(1) = 0$, with $q, r \in \{0, \cdots, R-1\}$ with the exception that $h_{rj}^{(r)}(j) = 1$. In Figure 11.1, some examples are shown for these $h_{rj}(x)$ functions. From these $h_{rj}(x)$ polynomials, the basis functions for the interpolation $b_{nr}(x)$ can be obtained as

$$b_{nr}(x) = \begin{cases} 0 & \text{if } x < (n-1)\Delta_x \\ \Delta_x^r h_{r1}(x/\Delta_x - n - 1) & \text{if } (n-1)\Delta_x \leq x < (n)\Delta_x \\ \Delta_x^r h_{r0}(x/\Delta_x - n) & \text{if } n\Delta_x \leq x < (n+1)\Delta_x \\ 0 & \text{if } x > (n+1)\Delta_x. \end{cases} \tag{11.3}$$



(a) (b)

Figure 11.1: (a) The $h_{rj}(x)$ function for order R=1. (b) Scaled $h_{0j}(x)$ function for order R=3.

Here again $n \in \{-N_x, \cdots, N_x\}$ and $r \in \{0, \cdots, R-1\}$. In Figure 11.2 some example basis functions are shown. Now the interpolation with the coefficients from Eq .(11.1) is simply given by

$$f(x) \approx \sum_{n=-N_x}^{N_x} \sum_{r=0}^{R-1} f_{nr} b_{nr}(x). \tag{11.4}$$

The addition of two functions $a(x) = b(x) + c(x)$ is trivially represented by $a_{nr} = b_{nr} + c_{nr}$.

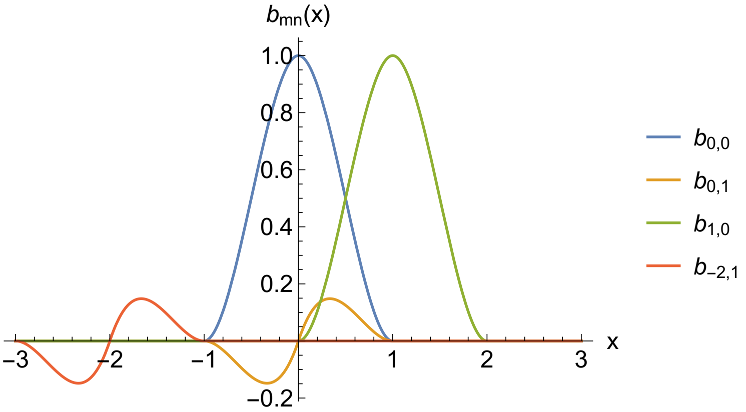


Figure 11.2: Hermite-interpolation basisfunctions for $\Delta_x = 1$ and $R = 2$.

The multiplication of functions, $a(x) = b(x)\, c(x)$ is represented by coefficients computed from the generalized Leibnitz rule

$$a_{nr} = \sum_{\ell=0}^{r} \frac{r!}{(r-\ell)!\ \ell!}\ b_{n,(r-\ell)}\, c_{n\ell}. \tag{11.5}$$

The computation of all coefficients a_{nr} requires $O(R^2 N_x)$ operations. Therefore, we conclude that the fourth point on the list in Section 11.2.1 is satisfied by the Hermite interpolation as long as R is restricted to a small integer, i.e., $R \in \{2, \cdots 5\}$.

11.2.3 Fourier transformation

In the spectral domain, we use Hermite interpolation as well. The Fourier transform of the approximated function in Eq. (11.4) is calculated analytically. The uniform sampling in the spatial domain is exploited in this Fourier transformation by employing FFTs. These FFTs dictate that (derivatives of) the analytic Fourier transformation are evaluated at an equidistant lattice in the spectral domain with lattice constant $\Delta_k = 2\pi/(2N_x + 1)$. From the analytically transformed values on this lattice in the spectral domain, a new Hermite interpolation can be devised in the spectral domain. Therefore, the analytic Fourier transform is only evaluated at an equidistant lattice in the spectral domain, and in between a Hermite-interpolated is used. Therefore, some approximation error is made. This is illustrated in Figure 11.3.

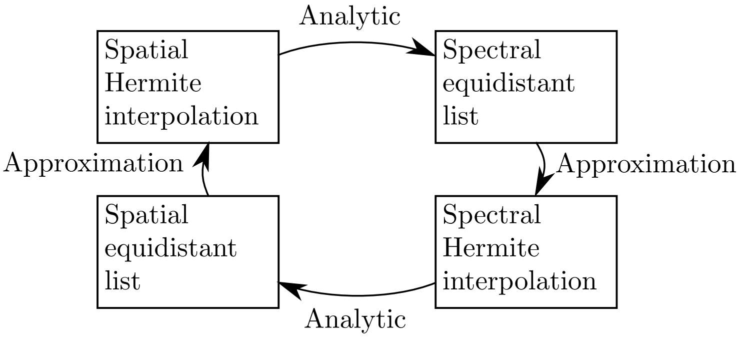

Analytic
Spatial Hermite interpolation
Spectral equidistant list
Approximation
Approximation
Spatial equidistant list
Spectral Hermite interpolation
Analytic

Figure 11.3: An analytic algorithm exists to compute function values and $R-1$ derivatives on an equidistant lattice. These values are then used in a new Hermite interpolation, from which some approximation error results.

In the spectral domain, the Hermite interpolation is computed at N_x lattice points. The equidistant sampling in the representation of Eq. (11.1) and Eq. (11.4) allows to compute the Fourier transform as

$$\hat{f}(k_x) \approx \int_{-\infty}^{\infty} dx \sum_{n=-N_x}^{N_x} \sum_{r=0}^{R-1} f_{nr} b_{nr}(x) e^{-jk_x x}, \tag{11.6}$$

where we used a hat on top of the function symbol to indicate the spectral domain. We propose to use Hermite interpolation in the spectral domain as well, with sampling distance $\Delta_k = 2\pi/(N_x\Delta_x)$. The coefficients in the spectral domain, $\hat{f}_{st}$ through Eq. (11.1), can be computed efficiently as

$$\begin{aligned} \hat{f}_{st} = \hat{f}^{(t)}(s\Delta_k) &= \int_{-\infty}^{\infty} \sum_{n=-N_x}^{N_x} \sum_{r=0}^{R-1} f_{nr} b_{nr}(x)(-jx)^t e^{-js\Delta_k x} \\ &= \sum_{n=-N_x}^{N_x} \sum_{r=0}^{R-1} [(-jn\Delta_x)^t f_{nr}] e^{-2\pi j\, sn/N_x} \int_{-\infty}^{\infty} (jx)^t e^{-js\Delta_k x} b_{0r}(x). \end{aligned} \tag{11.7}$$

Here we recognize a discrete Fourier transform in the sum over n, which can be calculated rapidly with the FFT algorithm. The integrals can be calculated during initialization, which yields a discrete Fourier transformation from n to s for each r and t value. Therefore, the computational efficiency scales as $O(R^2 N_x \log N_x)$. Since the number of derivatives should be chosen rather small, $R \in \{2, \cdots, 5\}$ and N_x large, this algorithm scales well to large numbers of unknowns. We conclude that also the third point on the list in Section 11.2.1 is satisfied by the Hermite interpolation.

11.2.4 Tests of the Hermite interpolation

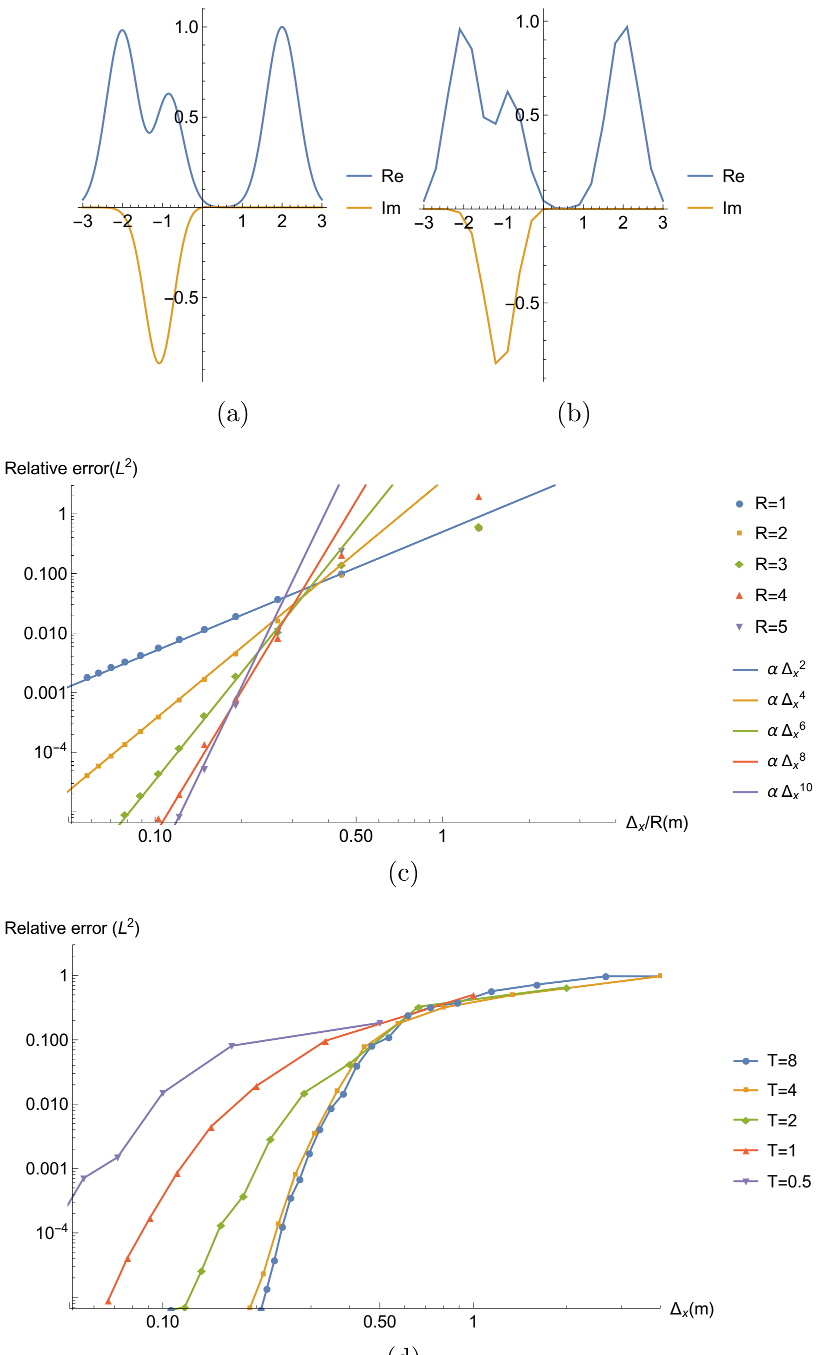
Showing that the Hermite interpolation satisfies the third and fourth point on the list in Section 11.2.1, leaves the first two points. We now provide numerical evidence that the first two points are also satisfied.

We test the Hermite interpolation on a continuous function consisting of three modulated Gaussians, which is shown in Figure 11.4(a). In Figure 11.4(b) we show the function approximated by a Hermite interpolation of very low order $R = 1$. In Figure 11.4(c) the $\mathcal{L}^2[-3, 3]$ error is shown on the double logarithmic scale for functions that are Hermite interpolated. The lines show the trend for dense sampling (small Δ_x). A polynomial convergence is observed with a convergence of order $2R$. For a relative error of 10^{-3}, the $R = 1$ sampling requires more than 10 times more samples than the sampling with $R = 5$. Although the high-R sampling is not as efficient as the Gabor-frame discretization shown in Figure 11.4(d), both methods perform well. It should be noted that the large window widths, where the Gabor frame has a clear advantage, are only efficient for very large simulation domains. Since the Gabor frame decays slowly to zero at the ends of the simulation domain, several extra windows are required to allow for this decay. Clearly, the Hermite interpolation is a competitive discretization and point one on the list in Section 11.2.1 is satisfied.

It remains to be determined how accurate the Fourier transformation associated with Hermite interpolation performs. As explained in Section 11.2.3, the forward Fourier transformation and inverse Fourier transformation contain an approximation where a new Hermite interpolation basis is used in the spectral and spatial domain, respectively. For this test, a simulation domain with $N_x = 200$ and $\Delta_x = 1$ is chosen. This corresponds to a $\Delta_k = 2\pi/401$ in the spectral domain. We check how a modulated Gaussian is transformed from the spatial to the spectral domain and back with Eq. (11.7) and its inverse, i.e., a full round in Figure 11.3. The Gaussian is chosen as

$$g_{X,K}(x) = \exp\left(-(x - X)^2/401 + jKx\right), \tag{11.8}$$

and we will use the symbol $\tilde{g}_{X,K}(x)$ to denote the back-and-forth transformed Gaussian. This function $\tilde{g}_{X,K}(x)$ approximates $g_{K,X}(x)$ best for small X and small K, since then $g_{X,K}$ exhibits the slowest oscillation in the spectral and spatial domain, respectively. A sample of $\tilde{g}_{X,K}$ and its approximation error are shown in Figure 11.5. As a measure of the total error in the approximation, we use the relative error based on the $\mathcal{L}^2[-X, X]$-norm. In Figure 11.6, we show how the approximation error depends on the position in the XK-plane. When a relative error of 10^{-3} is required, the useful domain for $R = 1$ is negligible. For larger R, the domain expands quickly to almost the complete simulation domain for $R = 4$. Therefore, we conclude that an accurate approximation also satisfies point two in Section 11.2.1 and therefore all points on the list in Section 11.2.1 are satisfied, when the proper value for R is chosen. Note that a small R corresponds to a low accuracy and that a large R corresponds to slow Fourier transformations, so a trade off between both is necessary. For our purpose, $R = 4$ is a proper choice. Strictly speaking, we have only shown numerical evidence for modulated Gaussian functions, but since these are the basis functions for the Gabor frame, any function that is important within the current formulation can be transformed.

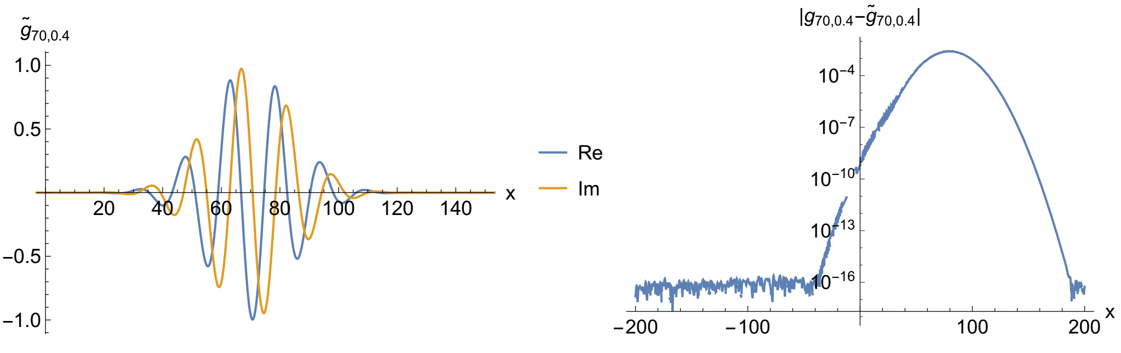


(a) (b)

(c)

(d)

Figure 11.4: (a) Function used for testing, given by $\exp(-\pi(x+2)^2)+\exp(-\pi(x-2)^2)+\exp(-\pi(x+1)^2+jx)$. (b) Function approximated with $\Delta_x = 0.3$ and $R = 1$. (c) Relative $\mathcal{L}^2$ error of approximation with Hermite interpolation for different R and Δ_x values. (d) Relative $\mathcal{L}^2$ error of approximation with a Gabor frame. Here $\Delta_x = \alpha T/N$, with N the range of the n-sum in Eq.(4.3), which is equivalent to a sample spacing of Δ_x/R in (c). Again we used an $\alpha = \sqrt{2/3}$ oversampling for the Gabor frame.



(a)

(b)

Figure 11.5: (a) A plot of $\tilde{g}_{70,0.4}(x)$ for $R = 2$. (b) The absolute error in $\tilde{g}_{70,0.4}(x)$, compared with the original function $g_{70,0.4}(x)$.

11.3 A continuous path in the spectral domain

11.3.1 Introduction

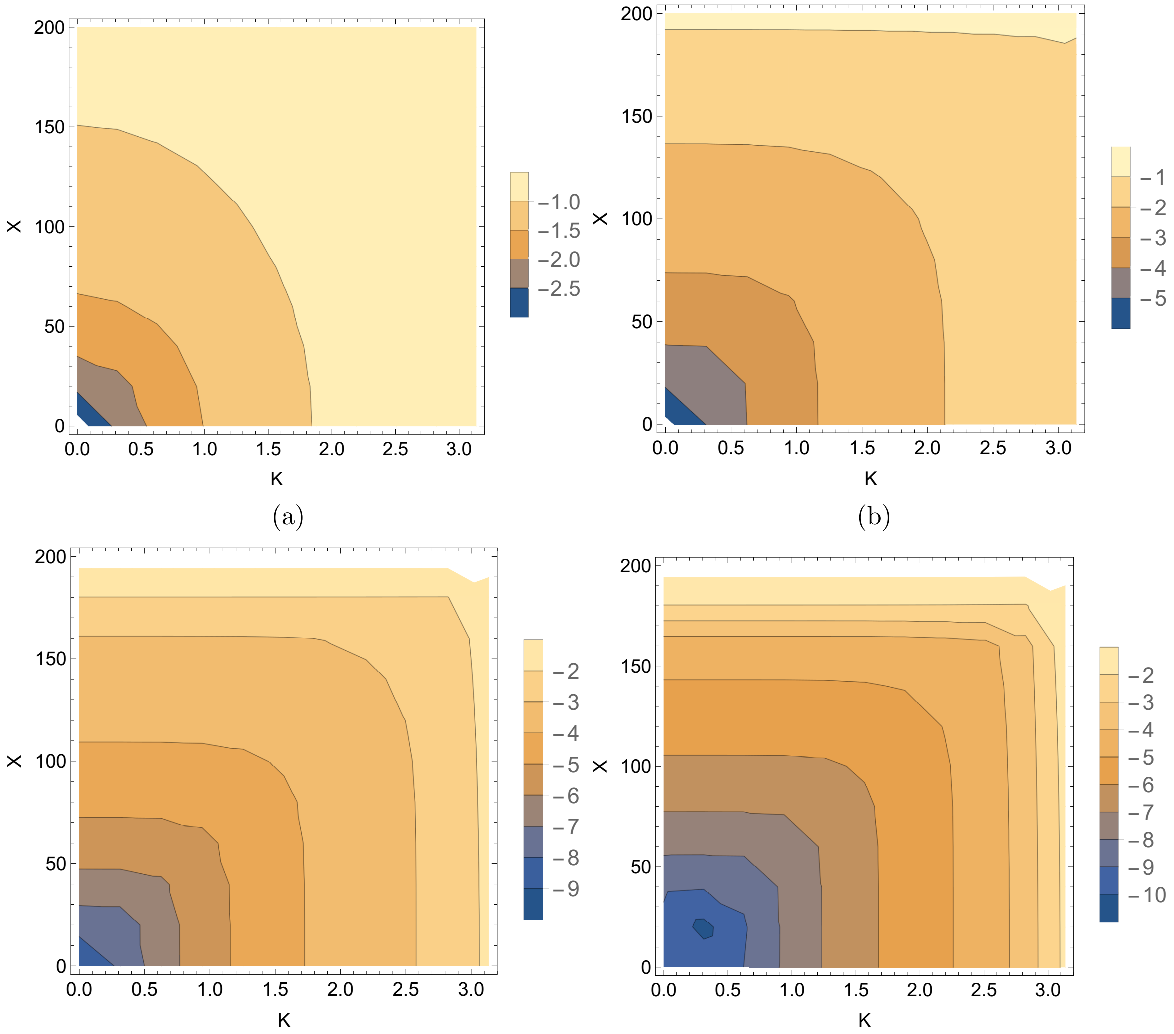
The main challenge with the spectral path is the transformation from the spatial domain to the complex-plane integration path in the spectral domain and back. Until now, we have used a path consisting of three distinct parts, each with a separate discretization. A major part of the work in this thesis consists of finding ways to transform back and forth to this path and representing functions accurately on this path. We now try to generalize the methods of transformation to methods that are applicable to more general choices for the path in the complex plane.

We begin by showing a more general way to represent a class of fast transformations between the spatial domain and the complex spectral path. This class of fast transformations assumes the availability of a rapid means of Fourier transformation, multiplication and summation. Then we will shown how the transformations to and from the complex spectral path that was used in Chapters 6 to 10 can be represented in this form. Afterwards, several alternative approaches to such a transformation are proposed and discussed. Based on these alternative transformations, a continuous path in the spectral domain is proposed and tested.

11.3.2 Transformations to and from a complex spectral path

We are interested in small path deformations $k \to \tau(k)$, with real-valued k, that can be described by $\tau(k) = k + jc(k)$. We assume $c(k)$ to be real valued and to have values, bounded by a number of the order of Δ_k, the resolution of the discretization in the spectral domain. The transformations that interest us are the transformation of a function $f(x)$ from the spatial domain to the complex spectral path

$$f(\tau(k)) = \int_{-\infty}^{\infty} dx\, f(x) e^{-jxk} e^{c(k)x} \tag{11.9}$$



(a) (b) (c) (d)

Figure 11.6: The $\log_{10}$ of the relative error in the approximation of the Fourier transform of the modulated Gaussian in Eq. (11.8) as a function of the location of its peak in the spatial and spectral domain. (a) $R = 1$ (b) $R = 2$ (c) $R = 3$ (d) $R = 4$. Notice the different scales in the legends.

and the transformation from the spectral path back to the spatial domain

$$f(x) = \int_{-\infty}^{\infty} dk\, (1 + jc'(k)) f(\tau(k)) e^{jxk} e^{-c(k)x}, \tag{11.10}$$

which is found by means of a substitution $k \to \tau(k)$ in the Fourier integral of $f(k)$. The difference between the integrals in Eqs. (11.9) and (11.10) is minor, i.e., the multiplication with the factor $(1 + jc'(k))$ does not pose a problem and then we are left with two integrals of the same form. Both integrals resemble a Fourier integral, except for the factor $\exp(\pm c(k)x)$. Since we assume that a rapid Fourier transformation is available (cf. Section 11.2.1), we will approximate Eqs. (11.9) and (11.10) as a sum of Fourier transformations. When we approximate

$$\begin{aligned} e^{c(k)x} &\approx \sum_{n=1}^{N} a_n(x) b_n(k) \\ e^{-c(k)x} &\approx \sum_{n=1}^{N} a_n^i(x) b_n^i(k), \end{aligned} \tag{11.11}$$

where the functions $a_n(x)$, $a_n^i(x)$, $b_n(k)$, and $b_n^i(k)$ remain to be defined and where the superscript ^{i} indicates the functions for the inverse transformation. Such approximations are available, since $c(k)$ is assumed to be small, i.e., on the order of the spectral resolution $2\pi/W_x$, and in the spatial domain functions are not evaluated outside of the simulation domain, i.e. we have $|x| < W_x$. This means that $\exp(\pm c(k)x)$ is bounded. With the path deformation as in Chapter 6, this approximation of the integral in Eq. (11.9) is usually chosen such that $0.05 < \exp(\pm c(k)x) < 20$ for x in the simulation domain. The transformations of a function f to and from the spectral path in Eq. (11.9) and Eq. (11.10) can then be rewritten as

$$f(\tau(k)) = \sum_{n=0}^{N} b_n(k) \int_{-\infty}^{\infty} dx\, a_n(x) f(x) e^{-jxk} \tag{11.12}$$

$$f(x) = \sum_{n=0}^{N} a_n^i(x) \int_{-\infty}^{\infty} dk\, (1 + jc'(k)) b_n^i(k) f(k) e^{jxk}. \tag{11.13}$$

These are sums of ordinary Fourier integrals. So when the functions $f(x)$ or $f(k)$ are represented by a Hermite interpolation, they can be computed efficiently, with the aid of the Fourier transformation in Eq. (11.7) and its inverse. The Hermite interpolation of $a(x)$, $a^i(x)$, $b(k)$ and $b^i(k)$ can be computed during initialization. Hence, these transformation consist of $N + 1$ FFTs and therefore a small N is highly important. The challenge is to find an efficient approximation Eq. (11.11) in a small number of terms.

11.3.3 The piecewise path

We cannot completely fit the piecewise path that was introduced in Chapter 6 into the picture of Section 11.3.2. The reason is that this method utilizes three distinct representations with distinct discretizations, f_L, f_M, and f_R, for each of the three parts of the complex path in the spectral domain, whereas the method in Section 11.3.2 uses only a single representation with a single discretization in the spectral domain. We now associate the $_L$ subscript with the spectral representation on the left horizontal part of the integration path in Figure 6.2(a) and the first line in Eq. (6.20). We associate the $_R$ subscript with the right horizontal part in Figure 6.2(a) and the third line in Eq. (6.20). For the middle part an optimized Fourier transformation was not used, so we will not dwell further on the second line in Eq. (6.20). This notation is applied in Eqs. (11.13) and (11.12) to yield expressions Eq. (6.19) and Eq. (6.22)

$$\begin{aligned} a_L(x) = a^i_R(x) =& e^{Ax} \\ a_R(x) = a^i_L(x) =& e^{-Ax} \\ b_L(k) = b_R(k) =& 1 \\ b^i_L(k) = b^i_R(-k) =& \begin{cases} 1 \text{ if } (k < A) \\ 0 \text{ if } (k \geq A). \end{cases} \end{aligned} \tag{11.14}$$

For the middle part (M) of the path, fast methods of Fourier transforming are not applied. The transformation to the spectral domain is approximated from the fit of a Taylor series to data points and the transformation back to the spatial domain is carried out by computing the integrals in Eq. (11.10) directly. This method is not very optimized, but since the M-part is small, not much computation time is used. A more optimized method was presented in Section 9.2.

11.3.4 Approximation by Taylor series

It is possible to approximate $\exp(\pm c(k)x)$ via a Taylor series as

$$\begin{aligned} e^{c(k)x} \approx& \sum_{n=0}^{N} c^n(k)\frac{x^n}{n!} \\ e^{-c(k)x} \approx& \sum_{n=0}^{N} c^n(k)\frac{(-x)^n}{n!}. \end{aligned} \tag{11.15}$$

When compared to Eq. (11.11), we can identify $a_n(x) = x^n/n!$, $a^i_n(x) = (-x)^n/n!$ and $b_n(k) = b^i_n(k) = c^n(k)$. An advantage of this method is that it allows for more general path shapes, whereas the piecewise path in Section 11.3.3 uses the fact that $c(k)$ is constant over most of the domain. It is advantageous to use a smooth path deformation, since that will yield a smooth $f(\tau(k))$ and therefore splitting up the spectral domain in different parts is no longer needed. Throughout the rest of this chapter we will therefore use the continuously prametrized path

$$\tau(k) = k + jc(k) = k + jA \text{ erf}(s\sqrt{\pi/2}\ k/A), \tag{11.16}$$

where the parameter A again has the meaning of the amplitude of the deformation and s defines the slope of $c(k)$ around 0. In Figure 11.7 $c(k)$ is plotted for Eq. (11.16).

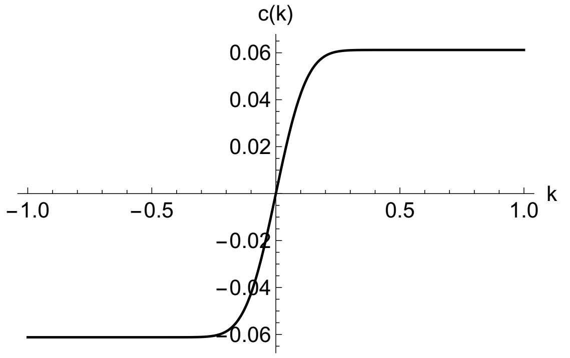


Figure 11.7: The imaginary part $c(k)$ of $\tau(k)$ in Eq. (11.16) for $s = 0.5$ and $A = 0.06$.

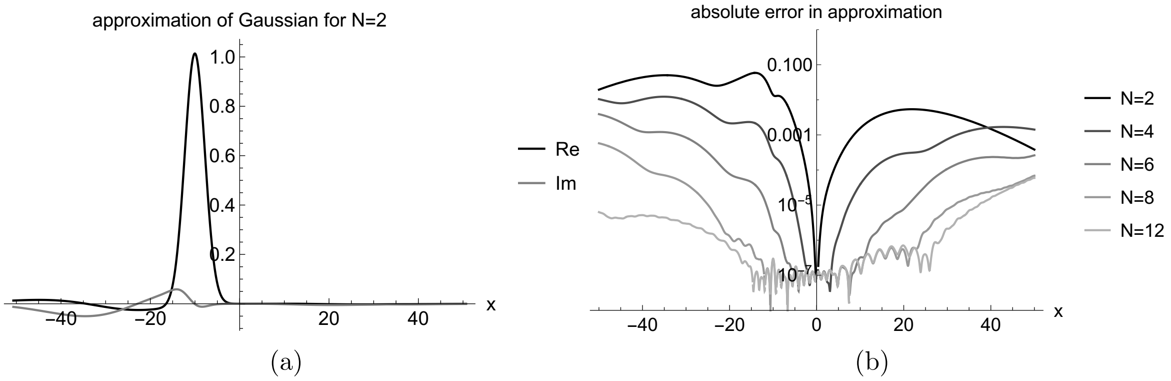


Figure 11.8: (a) The analytic Fourier transform of Gaussian pulse $f(k) = \mathcal{F}_k[\exp(0.1(x - 10)^2)](x)$ is sampled and then numerically transformed to the spatial domain using transformation Eq. (11.13) with approximation Eq. (11.15) for $N = 2$. (b) Approximation error for the same Gaussian pulse, transformed with higher-order Taylor approximations.

Next, we evaluate the accuracy of the Taylor-based version of the transformation to the spatial domain, i.e. Eq.(11.13), for the analytic Fourier transform of a Gaussian pulse that equals $f(k) = \mathcal{F}_k[\exp(0.1(x - 10)^2)](x)$. The numerical data is based on Hermite interpolation-based Fourier transforms with $N_x = 38$, $\Delta_x = 4/3$, $R = 4$, $\Delta_k = A = 0.612$. In Figure 11.8(a), the result of the transformation to the spatial domain is shown. A very low order Taylor-approximation with $N = 2$ was applied, which results in a visible error. In Figure 11.8(b), the absolute error is shown for transforming this same Gaussian pulse with higher-order Taylor approximations. Clearly, the error is larger for large x. This was to be expected, since the truncated Taylor series in Eq. (11.15) loses accuracy for large x.

Instead of using the Taylor series in Eq. (11.15) as a power series to approximate $\exp(-c(k)x)$, we have also used a fitted power series that is more accurate for large arguments, at the cost of the accuracy for small arguments. In Figure 11.9(a) we see the approximation of $\exp(Ax)$, which corresponds to $\exp(-c(k)x)$ for large values of k. Clearly, the fitted approximation has a wider range of validity. This is tested by transforming the spectral representation of a set of Gaussian pulses to the spatial domain. This set of pulses

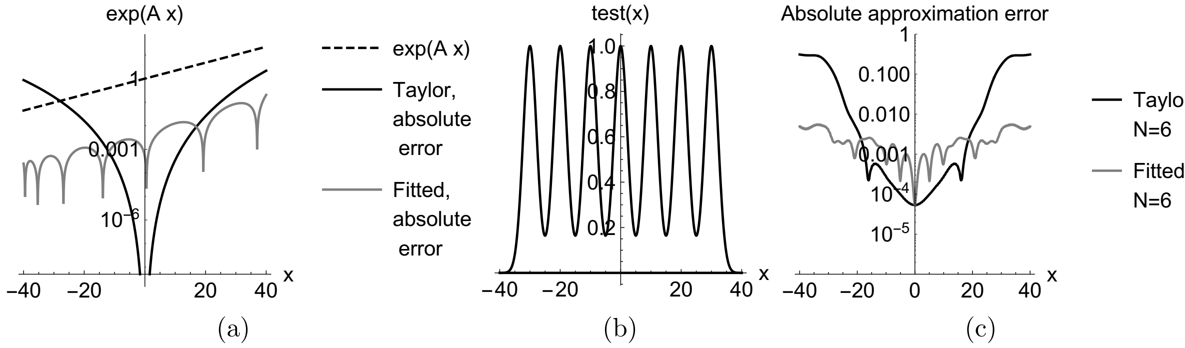


Figure 11.9: (a) An approximation of $\exp(Ax)$ valid for $x \in [-40, 40]$ compared to a Taylor series with the same number of terms. (b) A set of Gaussian pulses, whose representation on the spectral path is used as an input for the transformation from the path in Eq. (11.16) to the spatial domain. (c) The error made in the transformation to the spatial domain using Eq. (11.13) for a fitted polynomial and a Taylor series.

$\sum_{n=-3}^{3} \exp(0.1(x-10n)^2)$ is shown in Figure 11.9(b). In Figure 11.9(c), the fitted approximation performs better for large x. With higher numbers of terms the accuracy of the fit and the Taylor series improves.

11.3.5 Approximation by analytical expansion

A second approach to transform to and from the complex spectral integration path employs the assumption that the Fourier transform of a function is an analytic function and can therefore be continued analytically. This is especially useful for calculating the transformation from the spatial to the spectral domain in Eq. (11.12). For the transformation of the scattered electric field back to the spatial domain this method is less reliable, since the Green function is not analytic close to real k_x-axis for lossless background media. Therefore, we focus on the transformation from the spatial to the spectral domain.

Assume that we would like to calculate $f(k^*)$ and its derivatives at $k^* \in \mathbb{C}$, while values or derivatives of f are known only at some nearby points $\{f^{(d_1)}(k_1), f^{(d_2)}(k_2), \cdots, f^{(d_N)}(k_N)\}$. For example, the values are computed with the aid of the rapid Hermite-interpolation-based Fourier transformation Eq. (11.7), which yields values and derivatives up to order $R-1$ at an equidistant grid, where we then select values closest to k^*, from which we approximate f. The analyticity of $f(k)$ implies that a Taylor expansion t_f around k^* can be made, i.e.

$$t_f(k) = \sum_{n=0}^{N_t} w_n \frac{(k-k^*)^n}{n!}. \tag{11.17}$$

The weights in the Taylor expansion can be found from solving the associated Vandermonde system

$$t_f^{(d_n)}(k_n) = f^{(d_n)}(k_n)$$

for w_n. This technique is closely related to the Taylor series for the middle part in Eq. (6.29), where more information about solving such Vandermonde systems is given. When the Vandermonde system is solved, the coefficients w_n are then simply a list of derivatives. Again, since this is a small Vandermonde system, problems with a poorly conditioned system are avoided.

In the case of Hermite interpolation, this analytical continuation is especially useful, since functions are represented by an equidistant list of the function values and $R-1$ derivatives. The Hermite interpolation already uses the analyticity of the functions, since the basis functions are a local power-series expansion. This implies that an analytical expansion holds up to a distance on the order of the resolution Δ_k. Therefore, the result of a single Fourier transformation, which yields values on the real k axis, can be analytically expanded to any complex deformation, provided that the imaginary part of the path deformation is of the same order as Δ_k.

A numerical example is calculated for a modulated Gaussian pulse

$$g(x) = e^{-0.1(x+15)^2+0.4jx}. \tag{11.18}$$

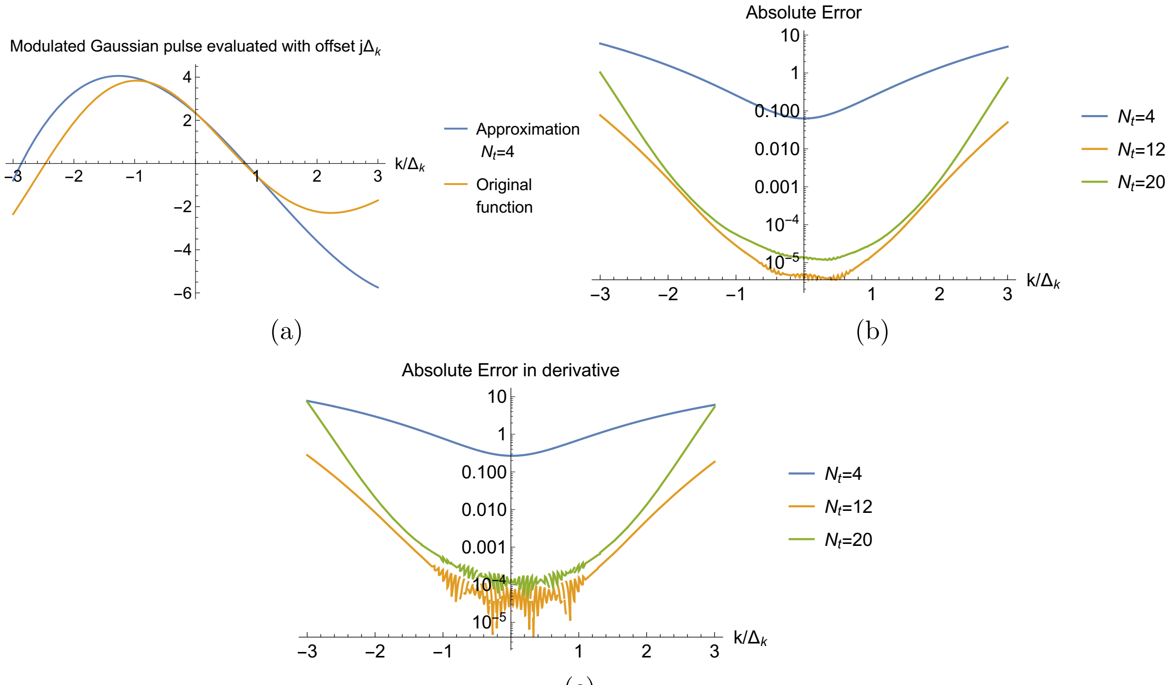
We have chosen $k^* = j\Delta_k$, and the Hermite interpolation of Figure 11.8 is used. The Fourier transformation of Eq. (11.7) is applied to find a numerical approximation of $g(k)$ and the result is analytically expanded into the complex plane to $g(k + j\Delta_k)$. Figure 11.10(a) shows $g(k + j\Delta_k)$ and an approximation where only $g^{(0)}(0), \cdots, g^{(3)}(0)$ were used to find the approximation for $N_t = 4$. In Figure 11.10(b), the difference between this numerical representation and two higher-order approximations, where $g^{(0)}(-\Delta_k), \cdots, g^{(3)}(-\Delta_k)$ and $g^{(0)}(\Delta_k), \cdots, g^{(3)}(\Delta_k)$ were added to the expansion for $N_t = 12$ for the middle line and values at $\pm 2\Delta_k$ were added for a total of 20 terms in the Taylor expansion for the lower line. In Figure 11.10(c) it can be observed that this approximation is also succesful for derivatives. The link between such an analytical expansion and the function $a_n(x)$ and $b_n(x)$ in Eq. (11.12) is that we can choose

$$\begin{aligned} a_n(x) &= (jx)^{d_n} e^{j(k_n - k^*)x} \\ b_n(x) &= 1. \end{aligned} \tag{11.19}$$

11.4 Numerical example: 2D transverse electric scattering in a homogeneous medium

Here, we demonstrate that the methods described in Sections 11.2 and 11.3 can be applied to an actual electromagnetic scattering problem. We apply the method to the problem of 2D TE scattering from a rectangular object with $\varepsilon_r = 4$ and dimensions 250 × 100 nm in the x and z directions, embedded in vacuum, as depicted in Figure 11.11. The incident field is propagating along the z-axis. Note that this case resembles the second example in Chapter 6 for a single scattering block.

A fourth-order Hermite interpolation is used ($R = 4$) to discretize the electric field in the x-direction. The Hermite interpolation is chosen such that $\Delta_x = 13.3$ nm, and



(a)

(b)

(c)

Figure 11.10: (a) A Fourier transformed modulated Gaussian pulse evaluated at complex coordinates, and its 4th order Taylor approximation. (b) The error for higher order approximations. (c) The error in the derivative for higher order approximations.

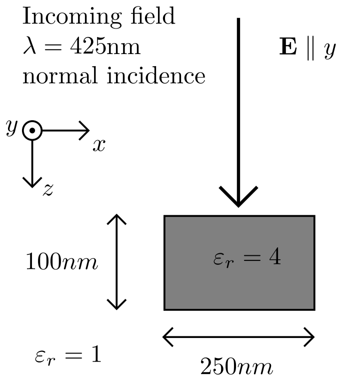


Figure 11.11: The scattering setup for 2D transverse electric scattering.

$N_x = 30$. This yields a simulation domain of 800 nm in width. Since the multiplication with the Green function in the spectral domain is comparable to a convolution in the spatial domain, the Green function has to be accurately discretized over a range that is twice as large as the scatterer, i.e. 500 nm in the present case. With the Gabor frame it was possible to choose a larger number of coefficients in the spectral domain, but this is not easily realizable with a Hermite interpolation, hence we choose to use the relatively large simulation domain of width 800 nm. In the z-direction 41 PWL basis functions were used with a $\Delta = 2.5$ nm interval as explained in Section 3.2. In the spectral, domain a continuous complex integration path was chosen as

$$\tau(k) = k + j\Delta_k \mathrm{erf}\left(\frac{\sqrt{\pi}}{4\Delta_k}k\right). \tag{11.20}$$

The transformation to the spectral domain was carried out by applying the method in Section 11.3.5, where N_t was chosen to be 12, with derivative 0 to $R-1$ at points $\{k - \Delta_k, 0, k + \Delta_k\}$. For the transformation to the spatial domain the method in 11.3.4 was used with a Taylor series of six terms. For simplicity, we preferred the Taylor series over the fitted polynomial.

To accurately approximate the contrast function, we discretized this function in the spectral domain, and multiplied it by the function $\exp((-.08k_x/k_0)^8)$. In this way, the Gibbs phenomenon at the edges of the object is somewhat suppressed.

In Figure 11.12 simulation results with this Hermite-interpolation method are shown, compared against a JCMWave [134] validation. Clearly, the results agree up to three digits precision, although some Gibbs phenomenon is visible at the edges of the object. However, this effect is smaller here than in earlier cases, since the contrast function was multiplied by a smoothing factor.

We have compared these results with the Gabor-frame based algorithm of Chapter 6. To reach a comparable accuracy, the contrast functions was discretized by a Gabor frame with window width $X = 150$ nm and the coefficients spanning $m \in \{-5, 5\}$ and $n \in \{-12, 12\}$ in Eq. (4.3). The total number of Gabor coefficients then equals 275, whereas the Hermite-interpolation based formulation required 244 basis functions. However, the Gabor frame

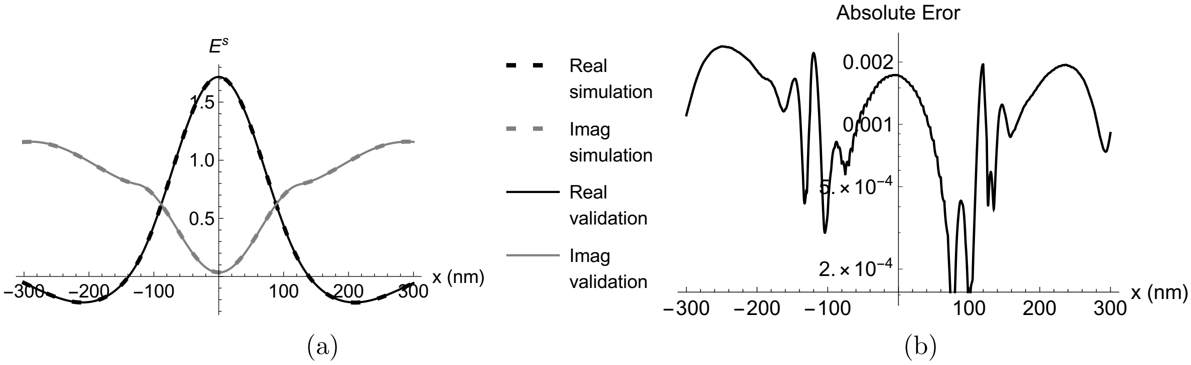


Figure 11.12: (a) The scattered electric field E^s at the bottom of the rectangular block. (b) The difference compared to a JCMWave validation.

requires a single coefficient per 4.9 nm, where the Hermite interpolation required a single unknown per 3.33 nm in the x-direction.

Clearly, both methods have have their advantages. The Gabor-frame based method needs a lower sampling resolution, as was already observed in Fig. 11.4(c) and (d). For smaller scatterering problems, such as this one, the Gabor frame is less efficient, since it requires a relatively large minimum number of coefficients. For completion we include timing results: with the mentioned settings both algorithms required 16 seconds. However, the Hermite-interpolation based method was implemented in Mathematica and the improvements of Chapter 10 were not implemented for the Gabor-frame based method, which hampers a true comparison of these computation times.

11.5 Discussion

In this chapter, an alternative for the Gabor-frame based discretization has been proposed: the Hermite interpolation. Although the Hermite interpolation is not as efficient as the Gabor frame for large structures, it requires fewer unknowns for smaller scattering problems such as in Fig. 11.12. Probably the most important lesson from this chapter is that the Gabor frame is certainly not the only possibility for the formulation in Chapter 2 and experiments with other discretizations should be considered. In Section 11.2.1, a list of requirements for an efficient discretization are mentioned.

A significant advantage of the Hermite interpolation is that it allows evaluating the interpolation quickly at any point, since there are only $2R$ basis functions contributing to each point in space. This is a significant disadvantage of the Gabor frame, i.e., the slow decay of the window functions often requires the sum in Eq. (4.3) to be evaluated over a hundred or more coefficients for 2D simulations and over 10,000 for 3D simulations. Therefore, it is much easier to extract data from Hermite-interpolated data than from Gabor-frame represented data.

Another important lesson from this chapter is that alternative integration paths are feasible. Implementing such a path requires different transformations to and from the complex-plane spectral integration path. Examples have been given of such transformations, but there will be other options. The major advantage of a path consisting of a single domain, instead of the three in Eq. (6.20), is that it can be handled more elegantly, especially for 3D problems. The spectral domain is then represented by just one discretization instead of nine distinct discretizations on the nine parts of the manifold for 3D in Fig. 9.5. However, for scatterers that are large in the z direction, where a lot of information is contained around $k_x = 0$, a more detailed discretization around $k_x = 0$ can still be advantageous, see Sections 9.2 and 9.3.

The transformation to the spectral domain in Section 11.3.5 is more efficient, in the sense that it requires only one Fourier transformation, instead of two. However, the transformation to the spectral domain in Section 11.3.4 is less efficient, since it required six Fourier transformations, which is more than the two required in the old formulation in Chapter 6. Since all information is now stored in a single spectral representation, the memory requirements have halved with respect to the old formulation where the spectral domain was divided in a left and right part, both of the size of the simulation domain.

Although the efficiency with respect to computational resources of the method in this chapter and that of Chapter 6 is roughly comparable, their elegance is not. The time spent on programming and debugging was about half a year for the method in Chapter 6 and about three days for the method in this chapter. This might not be a fair comparison, since the development of the method in this chapter largely benefited from the knowledge gained in Chapters 5 to 10, but that certainly does not compensate for the complete difference. The method of this chapter is still a good alternative, since new options can be implemented with significantly less effort, which can speed up scientific progress.

Chapter 12

Conclusions and outlook

12.1 Conclusions

The main result of this work is a new computational approach to compute the electromagnetic scattering from finite dielectric objects embedded in a dielectric multilayered medium. Algorithms were created for 2D scattering with both transverse electric and transverse magnetic polarization and for fully vectorial 3D scattering.

In the directions transverse to the layers, a translation symmetry is present in the multilayered background medium. This translation symmetry is exploited by applying the Green function multiplication in the spectral domain and the field-material interaction in the spatial domain. Hence, a discretization was needed for both the spatial and the spectral domain. By utilizing basis functions with a discrete translation symmetry, e.g. Gabor frames, the FFT algorithm could be applied to speed up operations on the data. The computational complexity of the present algorithm scales as $O(N \log N)$ and the memory requirements scale as $O(N)$, with N the number of unknowns.

The application of the Gabor frame for the discretization of this integral equation has had several advantages during the development of this algorithm. The exact Fourier transformation associated with the Gabor transformation allowed monitoring the error of functions and fields in both domains simultaneously. This gave much insight into the causes of large approximation errors. The Gabor frame allows for the truncation of functions in the spatial domain with control of its behaviour in the spectral domain and vice versa. Since some functions, among which the contrast function, are discontinuous, such control is vital for effective representations.

The spatial spectral method outlined in this thesis follows from two main ideas. The first idea behind this method is that the discretization should be accurate over a broad range around the origin in both the spectral and the spatial domain simultaneously. The second idea is that it should be closely monitored that any manipulation of functions in one domain does not deteriorate the accuracy in the other domain. For example, when two functions with complementary jumps are multiplied in the spatial domain, a very slow convergence is observed in the spectral domain. In the spectral domain this multiplication is represented by a poorly converging convolution, which violates the second idea.

The deformation of the discretization path into the complex plane allows for the use of Gohberg and Koltracht's efficient first-order recursion in the stacking direction of the multilayer medium. This method is extraordinarily efficient, since it does not require a Fourier transformation in the z-direction. The formulation with the complex spectral path and Gabor frames does not so much remove the need for Sommerfeld integrals, it samples them efficiently which allows us to include them into each iteration of the iterative solver, while maintaining a good computational efficiency.

The efficiency of the algorithm allows for accurate fully vectorial simulation of scattering from large dielectric objects in multilayered media. Throughout this thesis, several numerical examples have been shown where the present method reaches relative errors of 10^{-3} or better. The main challenge to validate the three dimensional method was the generation of accurate validation results via other methods, not the accuracy of the present algorithm.

12.2 Outlook

12.2.1 Validation against experiments

The algorithm for 2D TE polarized scattering in a multilayered medium is tested against two solvers in Chapter 6, FEM and RCWA. The other algorithms are tested against FEM only. In principle the FEM implementation we use [134], is a well-established implementation that has been tested against experimental data. Therefore, a good agreement between both methods implies that the presented method will also agree well with experimental data. For the two-dimensional algorithms the results obtained with a FEM implementation are computed with a very fine sampling and high number of polynomials and can be considered very accurate, at least six digits accuracy or more. However, for the three-dimensional algorithm such a fine discretization and such a high polynomial refinement was not achievable.

In Chapter 8 only one result is presented, where the FEM solver had converged to the accuracy of 10^{-3}, which is the accuracy goal that we strive for. On all other cases the memory requirements for such precise simulations exceeded the 256GB memory that was available on the computation server. This has made it impossible to compare the proposed solver up to a relative accuracy of more than $\sim 3{\cdot}10^{-3}$ against other data on all but one tiny example problem. In our opinion the best way to more rigorously test this method would be to test it directly against experimental data. Another approach for better validation would be to compare it against canonical objects for which analytical results are available, such as a sphere embedded in a homogeneous medium.

12.2.2 Scatterers in multiple layers

It is certainly possible to extend this method to scatterers that are embedded in multiple layers. For periodically repeated scatterers, such a solver was applied in [71]. From a theoretical point of view, such a generalization is not very challenging. However, it requires an update of the bookkeeping in the way the propagation in the z-direction is managed. Such an improved bookkeeping can then also include an non-uniform discretization in the z-direction, which can improve the convergence around corner singularities. Another improvement that can be added is a higher-order discretization in the z direction than the PWL discretization that is used in the present implementation. A higher order Hermite interpolation, or a Chebyshev interpolation such as in [162, 163], can then be included.

12.2.3 Error control

The present algorithm contains a large number of parameters, each of which has influence on the overall accuracy. Obtaining good results from the algorithm often includes some trial-and-error experiments to find simulation parameters that yield good results. It would be very beneficial to automate the choice of as many parameters as possible.

It would be worthwhile to search for a way to calculate an approximated contribution to the simulation error for each simulation parameter. With such estimations it is possible to generate a good set of simulation parameters in much less time. This would also make the algorithm useful for a user with a less detailed knowledge about the algorithm.

12.2.4 Improving the rate of convergence of the iterative solver

An important downside of the present formulation is that the iterative solver converges very slowly for large objects with high dielectric contrast. An examples of a large, high contrast object is a full model of the dielectric grating coupler that is found in [23] for integrated optics purposes. Including preconditioners to the solver can substantially increase the rate of convergence for an iterative solver [164, 165].

12.2.5 Other spectral manifolds

The present method decomposes the two-dimensional spectral plane into four large regions and five smaller connecting regions. The choice for this representation manifold was inspired by the discretization path for the two-dimensional scattering cases. We choose this manifold since the implementation was the most straightforward.

A simple change would be to change from the four regions NE, NW, SE and SW, each spanning 90 degrees of the $k_x - k_y$ plane, to three parts each spanning 120 degrees and four interconnecting regions. This could reduce both the memory requirements and computation time, at the cost of a marginal loss of accuracy. However, the implementation is less straightforward, since it no longer aligns with the Cartesian axes.

In Chapter 11, other methods of transformation and continuous paths are discussed. It is worthwile to look for different manifolds with more efficient transformations back and forth. The present method requires four full-size FFTs per transformation to or from the spectral domain. If this could be reduced, significant improvements are to be expected in both CPU time and memory requirements.

12.2.6 Locally refined sampling

It can be advantageous to use a refined sampling on certain locations in the spatial and/or spectral domain. In principle, an example is the sampling in the middle part for the complex spectral path in Eq. (6.20). In the spatial domain, singularities are found around corners of scatterers. A refined sampling can be advantageous at these corners for higher accuracy.

It is hard to implement such a refinement scheme efficiently with a Gabor-frame discretization. This is because evaluation of a discretized field at a single point is very slow, since many coefficients contribute. However, with the Hermite interpolation, the field can be evaluated locally very efficiently, and local refinements might be achievable.

12.2.7 Optimized programming

Currently, the programming of the algorithms has been performed with an emphasis on modularity and readability. This is perfect for prototyping, where the modularity is very advantageous for experimenting with e.g. different discretization methods. For readability, a functional-programming approach has been taken. Both these methods are certainly not optimal for execution speed. Once a certain approach has been settled and when large changes to the basis of the code are not expected, both the modularity and functional programming can be sacrificed for a more speed-optimized approach.

Appendix A

Gabor coefficients of a step function

Both for the cut function in the spectral domain Eq. (6.21) and for the contrast function in the spatial domain (2.28) it is desirable to find Gabor coefficients for the Heaviside step function $H(x)$. The problem with such a discontinuous function is that with an equidistant sampling the information about the precise position of the discontinuity gets lost. Therefore, the fast method to calculate Gabor coefficients from a sampled list does not produce accurate results. This can be resolved by using heavy oversampling, typically with a factor of thousand or more. However, this uses lots of computation time and memory. In principle, it is also possible to compute Gabor coefficients by using the integral (4.4). However, the integral has to be calculated for every individual Gabor coefficient and the integral converges slowly. Therefore, it is desirable to find a more efficient method.

A.1 Step function in one dimension

We divide the Heaviside step function $H(x)$ in a discontinuous part with an effectively finite support that will be handled in the spectral domain

$$H_k(x) = \begin{cases} e^{-\nu x^2} & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}, \tag{A.1}$$

and in a continuous part with infinite support that will be handled directly in the spatial domain

$$H_x(x) = \begin{cases} 1 - e^{-\nu x^2} & \text{if } x > 0 \\ 0 & \text{if } x < 0, \end{cases} \tag{A.2}$$

with ν a parameter that can be chosen for a balance between the continuity in Eq. (A.2) and the size of support in Eq. (A.1). The $H_x(x)$ function is continuously differentiable in the spatial domain and $H_k(x)$ is continuously differentiable in the spectral domain since it effectively has finite support in the spatial domain. For $H_x(x)$ there is no difficulty in calculating the Gabor coefficients.

The $H_k(x)$ function is discontinuous, but its Fourier transform is continuous. We can calculate its Fourier transform analytically via

$$H_k(k) = \int_0^\infty dk\, e^{-\nu x^2} e^{jxk} = \sqrt{\frac{\pi}{4\nu}} e^{-k^2/4\nu} \left[1 + j\text{erfc}(\sqrt{\nu}\frac{-jk}{2\nu})\right]. \tag{A.3}$$

This function is continuous in k, so we can readily calculate its Gabor coefficients over the range where they are needed. When we have its Gabor coefficients in the spectral domain, we can Fourier transform them and add them to the Gabor coefficients of $H_x(x)$. This yields an accurate representation of $H(x)$ in terms of Gabor coefficients. Of course it is possible to shift the position of the discontinuity by convolving with $\delta(X - x)$, which in the spectral domain corresponds to a multiplication by e^{-kX}, which is continuous and therefore straightforward to discretize.

A.2 Cut function in two dimensions

For the three-dimensional algorithm in Chapter 8 of this thesis, Gabor coefficients for a step function in two dimensions is required. We define a two-dimensional step function $H_{\mathbf{r}_1,\mathbf{r}_2}(\mathbf{x})$ in the two-dimensional coordinates $\mathbf{x} = (x, y)$ by $\mathbf{r}_1 = (x_1, y_1)$ and $\mathbf{r}_2 = (x_2, y_2)$ by the definition that $H_{\mathbf{r}_1,\mathbf{r}_2} = 1$ left of the line through $\mathbf{r}_1$, $\mathbf{r}_2$ and zero right of that line as illustrated in Figure A.1. We define new coordinates through the basis vectors

$$\begin{aligned} \mathbf{b} =& \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_2 - \mathbf{r}_1|} \\ \mathbf{a} =& \mathbf{b} \times \hat{\mathbf{z}}, \end{aligned} \tag{A.4}$$

so that $\mathbf{x} = (x_a + d_0)\mathbf{a} + x_b\mathbf{b}$, where x_a signifies the distance from $\mathbf{x}$ to the discontinuity as is illustrated in Figure A.1 and where d_0 is obtained from $d_0 = \mathbf{a} \cdot \mathbf{r}_1 = \mathbf{a} \cdot \mathbf{r}_2$.

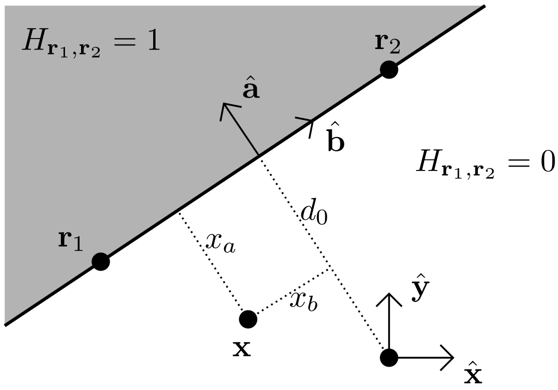


Figure A.1: A visualization of the variables involved in the calculation of the scaled coordinates.

In principle, we would like to use the same construction as we did in the previous section. So again, we decompose $H_{\mathbf{r}_1,\mathbf{r}_2}$ into a spatial part $H_{\mathbf{r}_1,\mathbf{r}_2,x}$ and a spectral part

$H_{\mathbf{r}_1,\mathbf{r}_2,x}$. The spatial part can be defined as

$$H_x(x_a, x_b) = \begin{cases} 1 - e^{-\nu x_a^2} & \text{if } x_a > 0 \\ 0 & \text{if } x_a < 0, \end{cases} . \tag{A.5}$$

where $x_a = \mathbf{x} \cdot \mathbf{a} - d_0$ and $x_b = x \cdot \mathbf{b}$, with $\mathbf{x} = (x, y)$, the position coordinate in the original system.

Special care has to be taken in the spectral domain, since in the spectral domain the function has to be continuous in both directions, where it was defined as a continuous function only in one direction in the previous section in (A.3). Therefore, we multiply this function in the spatial domain by $e^{-\mu x_b^2}$, which ensures that its Fourier transform is continuous and does not contain a δ function in the $\mathbf{b}$ direction. This spectral part in two dimensions in the spatial domain is then given by

$$H_{k,\mathbf{r}_1,\mathbf{r}_2}(k_a, k_b) = \sqrt{\frac{\pi^2}{4\nu\mu}} e^{-k_a^2/4\nu - k_b^2/4\mu} \left[1 + j\text{erfc}(\sqrt{\nu}\frac{-jk_a}{2\nu}) \right], \tag{A.6}$$

where $k_a = \mathbf{k} \cdot \mathbf{a}$ and $k_b = \mathbf{k} \cdot b$. This function is continuous in both the spatial and the spectral domain, and can be accurately discretized for a proper choice of μ without heavy oversampling. We choose $1/\mu = \max(W_x, W_y)$, since it yields good results for the cases we considered. To discretize this function in the spectral domain, we calculate four times more spatial samples (index m in Eq. (4.1)) for a finer sampling and therefore a higher accuracy. Afterwards we discard the excess Gabor coefficients.

The next step is to Fourier transform the Gabor coefficients back to the spatial domain. Since we used a $e^{-\mu x_b^2}$ to make Eq. (A.6) continuous, this has to be compensated again by a factor of $e^{\mu x_b^2}$. To summarize this, Gabor coefficients of the Heaviside step function are found through

$$H_{\mathbf{r}_1,\mathbf{r}_2}(\mathbf{x}) = e^{\mu(\mathbf{x}\cdot\mathbf{b})^2} \mathcal{F}_{\mathbf{k}} \left[H_{k,\mathbf{r}_1,\mathbf{r}_2}(\mathbf{k}\cdot\mathbf{a}, \mathbf{k}\cdot\mathbf{b}) \right](\mathbf{x}) + H_{x,\mathbf{r}_1,\mathbf{r}_2}(\mathbf{x}\cdot\mathbf{a} - d_0, \mathbf{x}\cdot\mathbf{b}). \tag{A.7}$$

A.3 Gabor coefficients of a disc

For completeness, we also include how we compute the Gabor coefficients of a disc with radius R and amplitude 1. Since this function is discontinuous in the spatial domain, it is discretized in the spectral domain by function $H_R^\circ(\mathbf{k}_T)$ via

$$H_R^\circ(\mathbf{k}_T) = 2\pi R \frac{J_1(|\mathbf{k}_t|R)}{\mathbf{k}_t}. \tag{A.8}$$

Here $J_1(k)$ denotes the Bessel function of the first kind of order one. Since this function is continuous, it can be easily discretized using a Gabor transform. Later the result can be transformed to the spatial domain through a Fourier transformation, to arrive at a discretized circle in the spatial domain. A displacement of the center of this disc by vector $\mathbf{x}_{T,0}$ can be achieved by multiplying by $\exp(-j\mathbf{x}_{T,0} \cdot \mathbf{k}_T)$ in the spectral domain.

Curriculum Vitae

Roeland Johannes Dilz was born in Utrecht, the Netherlands on March 17, 1987. He finished his secondary education in 2005, after which he enrolled for the study Theoretical Physics at the University of Antwerp. The next year he switched to Utrecht University, where he continued his studies in Physics and Mathematics. In 2009 he finished his B.Sc. Degree in mathematics and physics, to continue to pursue a M.Sc. Degree in Theoretical Physics. His M.Sc. Thesis involved modeling glass at temperatures close to the glass transition temperature under the supervision of Prof. Dr. Gerard T. Barkema. Starting at his B.Sc. education he simultaneously worked one to three days per week as a teaching assistant for different physics courses, most importantly for the experimental physics course.

After he graduated from university in 2012, his curiosity and love for a challenge drove him on his pushbike through the rain forests, snowy peaks and deserts of the Andes. A year of adventure later, in 2013, he started as a Ph.D. student at the Electromagnetics group of the Eindhoven University of Technology. The focus of his research is on developing an efficient spatial spectral solver for multilayered media. Other activities during his Ph.D. included the organizing the EM colloquium and teaching exercise classes at several courses including the basic course of Electromagnetics in the second year of the curriculum.

List of publications

Journal articles

- R. J. Dilz and M. C. van Beurden. The Gabor frame as a discretization for the 2D transverse-electric scattering-problem domain integral equation. *Progress in Electromagnetics Research B*, 69:117–136, 2016

- R. J. Dilz and M. C. van Beurden. An efficient complex spectral path formulation for simulating the 2D TE scattering problem in a layered medium using Gabor frames. *Journal of Computational Physics*, 345:528–542, 2017

- R. J. Dilz, M. G. G. M. van Kraaij, and M. C. van Beurden. The 2D TM scattering problem for finite objects in a dielectric stratified medium employing gabor frames in a domain integral equation. *Journal of the Optical society of America A*, 34(8):1315–1321, 2017

- R. J. Dilz, M. G. G. M. van Kraaij, and Martijn C. van Beurden. A 3D spatial spectral integral equation method for electromagnetic scattering from finite objects. *Optical and Quantum Electronics*, Under review

- Roeland J. Dilz and Martijn C. van Beurden. Fast operations for a Gabor-frame based integral equation with equidistant sampling. *IEEE Antennas and wireless propagation letters*, Accepted

Conferences

- R. J. Dilz and M. C. van Beurden. An efficient spatial spectral integral-equation method for EM scattering from finite objects in layered media. In *2016 International Conference on Electromagnetics in Advanced Applications (ICEAA), 19-23 September 2016, Cairns, Australia*, pages 509–511, 2016

- Roeland J. Dilz and Martijn C. van Beurden. A 3D spatial-spectral integral equation method for electromagnetic scattering from finite objects in a layered medium. In *Optical Wave and Waveguide Theory and Numerical Modelling (OWTNM), 5-6 April 2017, Eindhoven, The Netherlands*, 2017

- Roeland J. Dilz and Martijn C. van Beurden. Scaling of a spatial spectral integral-equation method for EM scattering in a stratified medium to large, finite objects. In *Progress in Electromagnetics research symposium (PIERS), 22-25 May 2017, Saint Petersburg, Russia*, 2017

- Roeland J. Dilz and Martijn C. van Beurden. Computational aspects of a spatial-spectral domain integral equation for scattering by objects of large longitudinal extent. In *2017 International Conference on Electromagnetics in Advanced Applications (ICEAA), 11-15 September 2017, Verona, Italy*, 2017

Acknowledgements

First of all, I would like to thank my father for teaching me how to experimentally build or repair whatever you set your mind on. He taught me the art of problem solving through trial and error. Even when you initially do not understand the details, the best way to start understanding is often (but not always!) to just try something and see what happens.

Another big thanks goes to Wim Westerveld from Utrecht University for letting me assist during his Experimental Physics course. Here, I gained much more than the freedom that comes with some money and a coffee card. Apart from creating a friendly open atmosphere, it was inspiring to see how he could make complicated things easily understandable. From him I have learned to find a good balance between theoretical and experimental approaches and an appreciation for teaching activities. Another important person during my studies was Gerard Barkema, supervisor of my M.Sc. Thesis, with whom I have had many interesting discussions on numerical methods, statistics, and molecular dynamics.

As support for my Ph.D. thesis, the most important person was definitely Martijn van Beurden. I think he very well complemented my weaker points. Not only did he give me many ideas and mathematical tools to complete the technical part of this thesis, he has also been the first person with enough patience to force me to present the fruits of my research in a presentable form of some sort. Additionally, he was also kindly supportive on the moments that private issues were forcing my mind to think of other things than my project. It was great to work together with someone who is meticulous in both his scientific work as well as in getting all organizatorial stuff done, so I could spend my time on research. Without his support, this thesis would not have been the same.

Further, I would like to express my gratitude towards Anton Tijhuis for leading a friendly group that includes a strong interest in computational electromagnetics. His critical proofreading has greatly improved the quality of this thesis. Additionally, I express my gratitude to the reading committee for proofreading. I would also like to thank the NWO-TTW user-committee, consisting of Wim Coene, Marijn van Veghel, Peter van den Berg, and Jos van der Tol for their input and checking whether the project was headed in the right direction. Further, I am grateful to Mark van Kraaij and Maxim Pisarenco at ASML for their efforts in producing validation results, which have been instrumental in lifting this work from an elaborate piece of speculation into reality.

On the social side of Ph.D. life, I happily think back at the many happy times that I spent in het Walhalla. After my first introduction by Mojtaba, I was immediately a fan of that smelly cellar that is best enjoyed with cheap, high-quality beer and a handful of cheap low-quality nuts. It never took much convincing to get some colleagues to join me there, most notably, Bas, Pieter, Ali, Mojtaba, Bedilu, Shady, and Satoru. After four years, I think that I can identify every person in our group by their laughter, which is a good thing.

Luckily, I was also warmly supported by my friends and family, even though the work sometimes limited the time I was able to spend with them. A special thanks goes to my parents, for their trust and support during the whole process. I would also like to mention my brother Floris; I always liked it when he joined me after a long week of hardcore bugfixing for a well-earned drink and Steven Seagal movie. To my girlfriend Annemarie: you gave me the energy and motivation to really push through when needed. Without you I would never have finished this thesis in time. And finally of course Jasper, Gijs, Emma, Marie-anne, Martijn, Loes: all of you have in your own way significantly contributed to the vital distraction from work that has kept me productive, be it by ice-skating, cycling, tropical holidays, music, or just a nicely baked brownie.

Bibliography

[1] Hugh Aitken. *The Continuous Wave: Technology and American Radio, 1900-1932.* Princeton University Press, 1985.

[2] A. E. Sale. The rebuilding of colossus at bletchly park. *IEEE Annals of the History of Computing*, 27(3):61–69, 2005.

[3] T. K. Sarkar, Robert Mailloux, Arthur A. Oliner, M. Salazar-Palma, and Dipak L. Sengupta. *History of Wireless.* John Wiley & Sons, Inc, 2006.

[4] Mike Green. Dummer's vision of solid circuits at the UK royal radar establishment. *IEEE Annals of the History of Computing*, 35(1):56–66, 2011.

[5] David Lammers. Moore's law milestones. *IEEE Spectrum*, April 2015. http://spectrum.ieee.org/geek-life/history/moores-law-milestones.

[6] T. Singh, S. Rangarajan, D. John, C. Henrion, S. Southard, H. McIntyre, A. Novak, S. Kosonocky, R. Jotwani, A. Schaefer, E. Chang, J. Bell, and M. Co. Zen: A next-generation high-performance x86 core. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 52–53, Feb. 2017.

[7] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):33–35, 1965.

[8] Rachel Courtland. Intel finds moore's law's next step at 10 nanometers. *IEEE Spectrum*, December 2016. http://spectrum.ieee.org/semiconductors/devices/intel-finds-moores-laws-next-step-at-10-nanometers.

[9] Nader Jalili and Karthik Laxminarayana. A review of atomic force microscopy imaging systems: application to molecular metrology and biological sciences. *Mechatronics*, 14(8):907–945, 2004.

[10] Jon Orloff, editor. *Handbook of charged particle optics.* CRC Press, 1997.

[11] Benjamin Bunday. Small feature accuracy challenge for CD-SEM metrology physical model solution. In *Proc. SPIE 6152, Metrology, Inspection, and Process Control for Microlithography XX*, 2006.

[12] Kazuhisa Hasumi, Osamu Inoue, Yutaka Okagawa, Chuanyu Shao, Philippe Leray, Sandip Halder, Gian Lorusso, and Christiane Jehoul. Sem based overlay measurement between via patterns and buried m1 patterns using high voltage sem. In *Proc. SPIE 10145, Metrology, Inspection, and Process Control for Microlithography XXXI*, 2017.

[13] P.C. Logofatu, D. Apostol, V. Damian, V. Nascov, F. Garoi, A. Timcu, and I. Iordache. Scatterometry, an optical metrology technique for lithography. In *Semiconductor Conference, 2004. CAS 2004 Proceedings. 2004 International*, pages 517–520, 2004.

[14] H. Tompkins and E. A. Irene. *Handbook of Ellipsometry.* William Andrew, 2005.

[15] Sridhar Mahendrakar, Alok Vaid, Kartik Venkataraman, Michael Lenahan, Steven Seipp, Fang Fang, Shweta Saxena, Dawei Hu, Nam Hee Yoon, Da Song, Janay Camp, and Zhou Ren. Optical metrology solutions for 10nm films process control challenges. In *Proceedings of SPIE, Vol. 9778: Metrology, Inspection, and Process Control for Microlithography XXX*, 2016.

[16] Danielle E. Adams, Christopher S. Wood, Margaret M. Munrane, and Henry C. Kapteyn. Tabletop high harmonics illuminate the nano-world. *LaserFocusWorld*, 51(5):38–41, 2015.

[17] Mirko Holler, Manuel Guizar-Sicairos, Esther H. R. Tsai, Roberto Dinapoli, Elisabeth Müller, Oliver Bunk, Jörg Raabe, and Gabriel Aeppli. High-resolution non-destructive three-dimensional imaging of integrated circuits. *Nature*, 543:402–406, March 2017.

[18] David M. Pozar and Daniel H. Schaubert. Scan blindness in infinite pphase aarray of printed dipoles. *IEEE Transactions on Antennas and Propagation*, 32(6):602–611, 1984.

[19] Nanfang Yu, Patrice Genevet, Mikhail A. Kats, Francesco Aieta, Federico Capasso, and Zeno Gaburro. Light propagation with phase discontinuities: Generalized laws of reflection and refraction. *Science*, 334:333–337, Oct. 2011.

[20] F. Silvestri, G. Gerini, E. Pisano, and V. Galdi. High numerical aperture all-dielectric metasurface micro-lenses. In *IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, 2015*, pages 1030–1031, 2015.

[21] Miguel Ruphuy, Omar Siddiqui, and Omar M. Ramahi. Electrically thin flat lenses and reflectors. *Journal of the Optical Society of America A*, 32(9):1700–1706, 2015.

[22] Amir Arbabi, Yu Horie, Mahmood Bagheri, and Andrei Faraon. Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission. *Nature Nanotechnology*, 10:937–946, Nov. 2015.

[23] Dirk Taillaert, Frederik Van Laere, Melanie Ayre, Wim Bogaerts, Dries Van Thourhout, Peter Bienstman, and Roel Baets. Grating couplers for coupling between optical fibers and nanophotonic waveguides. *Japanese Journal of Applied Physics*, 45(8a):6071–6077, 2006.

[24] D.O. Dzibrou, J.J.G.M. van der Tol, and M.K. Smit. Tolerant polarization converter for InGaAsP-InP photonic integrated circuits. *Optics Letters*, 38(18):3842–3484, Sep 2013.

[25] Lingyun Wang, Youmin Wang, and Xiaojing Zhang. Embedded metallic focus grating for silicon nitride waveguide with enhanced coupling and directive radiation. *Optics Express*, 20(16):17509–17521, 2012.

[26] Kurt L. Shlager and John B. Schneider. A selective survey of the finite-difference time-domain literature. *IEEE Antennas and Propagation Magazine*, 37(4):39–57, 1995.

[27] Atef Z. Elsherbeni and Veysel Demir. *The Finite Difference Time Domain Method for Electromagnetics: With MATLAB Simulations.* SciTech Publishing, 2009.

[28] Dennis M. Sullivan. *Electromagnetic Simulation Using the FDTD Method, 2nd edition.* Wiley, 2013.

[29] Gerrit Mur. Absorbing boundary conditions for the finite-difference approximation of the time-domain. *Transactions on Electromagnetic Compatibility*, 23(04):377–382, 1981.

[30] Ilker R. Capoglu. *Techniques for handling multilayered media in the FDTD method.* 2007.

[31] T. Weiland. Time domain electromagnetic field computation with finite difference methods. *International Journal of Numerical Modelling: Electronic networks, devices and fields*, 9(4):295–319, 1996.

[32] Mats G. Larson and Fredrik Bengzon. *The Finite Element Method: Theory, Implementation, and Applications.* Springer, 2013.

[33] Bangda Zhou and Dan Jiao. Direct finite-element solver of linear complexity for large-scale 3-D electromagnetic analysis and circuit extraction. *IEEE Trans. Microwave Theory Tech*, 63(10):3066–3080, 2015.

[34] Jan Pomplun, Sven Burger, Frank Schmidt, Frank Scholze, Christian Laubis, and Uwe Dersch. Metrology of EUV masks by EUV-scatterometry and finite element analysis. In *Photomask and Next-Generation Lithography Mask Technology XV, Proc. of SPIE Vol. 7028*, 2008.

[35] Frank Scholze, Christian Laubis, Gerhard Ulm, Uwe Dersch, Jan Pomplun, Sven Burger, and Frank Schmidt. Evaluation of EUV scatterometry for CD characterization of EUV masks using rigorous FEM-simulation. In *Emerging Lithographic Technologies XII, SPIE 6921*, 2008.

[36] Sven Burger, Roderick Kohle, Lin Zschiedrich, Weimin Gao, Frank Schmidt, Reinhard Marz, and Christoph Nolscher. Benchmark of FEM, waveguide and FDTD algorithms for rigorous mask simulation. In *25th Annual BACUS Symposium on Photomask Technology Proc. SPIE 5992*, 2005.

[37] P. Zwamborn and P. M. van den Berg. The three-dimensional weak form of the conjugate gradient FFT method for solving scattering problems. *IEEE Trans. Microwave Theory Tech*, 40(9):1757–1766, Sep 1992.

[38] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes: The Art of Scientific Computing.* Cambridge University Press, 2007.

[39] Henk A. van der Vorst. *Iterative Krylov Method for Large Linear Systems.* Cambridge University Press, 2003.

[40] Roger F. Harrington. *Field Computation by Moment Methods.* Wiley-IEEE Press, 1993.

[41] S. M. Rao, D. R. Wilton, and A. W. Glisson. Electromagnetic scattering by surfaces of arbitrary shape. *IEEE Transactions on Antennas and Propagation*, 30(3):409–418, 1982.

[42] E. Bleszynski, M. Bleszynski, and T. Jaroszewicz. AIM: Adaptive integral method for solving large-scale electromagnetic scattering and radiation problems. *Radio Science*, 31(5):1225–1251, 1996.

[43] J. R. Philips and J. K. White. Efficient capacitiance extraction of 3D structures using generalized precorrected FFT methods. In *Electrical Performance of Electronic packaging, 1994., IEEE 3rd Topical Meeting on*, pages 253–256, November 1994.

[44] Joel R. Phillips and Jacob K. White. A precorrected-FFT method for electrostatic analysis of complicated 3-D structures. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 16(10):1059–1072, 1997.

[45] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73:325–348, 1987.

[46] V. Rokhlin. Rapid solution of integral equations of scattering theory in two dimensions. *Journal of Computational Physics*, 86:414–439, 1990.

[47] J. Song, Cai-Cheng Lu, and Weng Cho Chew. Rapid solution of integral equations of scattering theory in two dimensions. *IEEE Transactions on Antennas and Propagation*, 45(10):1488–1493, 1997.

[48] C.C. Lu, J.M. Song, and W.C. Chew. A multilevel fast multipole algorithm for solving 3D volume integral equations of electromagnetic scattering. In *Antennas and Propagation Society International Symposium, 2000. IEEE*, pages 1864–1867, 2000.

[49] Norbert Geng, Anders Sullivan, and Lawrence Carin. Multilevel fast-multipole algorithm for scattering from conducting targets above or embedded in a lossy half space. *IEEE Transactions on Geoscience and Remote Sensing*, 38(4):1561–1573, 2000.

[50] B. Hu, W. C. Chew, E. Michielssen, and J. S. Zhao. Fast inhomogeneous plane wave algorithm (FIPWA) for the fast analysis of 2D scattering problems. *Radio Science*, 34(4):759–772, 1999.

[51] Bin Hu, Weng Cho Chew, and Sanjay Velamparambil. Fast inhomogeneous plane wave algorithm for the analysis of electromagnetic scattering. *Radio Science*, 36(6):1327–1340, 2001.

[52] Bin Hu and Weng Cho Chew. Fast inhomogeneous plane wave algorithm for electromagnetic solutions in layered medium structures: Two-dimensional case. *Radio Science*, 35(1):31–43, 2000.

[53] Krzysztof A. Michalski and Juan R. Mosig. Multilayered media Green's functions in integral equation formulations. *IEEE Transactions on Antennas and Propagation*, 45(3):501–519, 1997.

[54] Arnold Sommerfeld. Uber der Ausbreitung der Wellen in der drahtlosen Telegraphie. *Annalen der Physik*, 1909.

[55] Weng Cho Chew. *Waves and fields in inhomogeneous media.* IEEE Press, 1995.

[56] Leopold B. Felsen and Nathan Marcuvitz. *Radiation and Scattering of Waves.* IEEE Press, 1973.

[57] J. A. Kong. *Theory of electromagnetic waves.* John Wiley & Sons, Inc, 1975.

[58] J. R. Wait. *Electromagnetic Waves in Stratified Media.* Pergamon Press, 1970.

[59] Amit Hochman and Yehuda Leviatan. A numerical methodology for efficient evaluation of 2D Sommerfield integral in the dielectric half-space problem. *IEEE Transactions on Antennas and Propagation*, 58(2):413–431, Feb. 2010.

[60] H. M. de Ruiter. Limits on the propagation constants of planar optical waveguide modes. *Applied Optics*, 20(5):731–732, 1981.

[61] E. H. Newman and D. Forrai. Scattering from a microstrip patch. *IEEE Transactions on Antennas and Propagation*, 35(3):245–251, March 1987.

[62] Krzysztof A. Michalski and Juan R. Mosig. Efficient computation of Sommerfeld integral tails – methods and algorithms. *Journal of Electromagnetic Waves and Applications*, 30(3):281–317, 2016.

[63] Samir F. Mahmoud and Ahmed D. Metwally. New image representation for dipoles near a dissipative earth 1. discrete images. *Radio Science*, 16(6):1271–1275, 1981.

[64] D. G. Fang, J. J. Yang, and G. Y. Delisle. Discrete image theory for horizontal electric dipoles in a multilayered medium. In *IEEE Proceedings Microwaves, Antennas and Propagation*, pages 297–303, 1988.

[65] E.P. Karabulut, A. T. Erdogan, and M. I. Aksun. Discrete complex image method with spatial error criterion. In *IEEE Transactions on Microwave Theory and Techniques*, volume 59, pages 793–802, 2011.

[66] Ming-Yao Xia, Chi Hou Chan, Yuan Xu, and Weng Cho Chew. Time-domain Green's functions for microstrip structures using Cagniard–de Hoop method. *IEEE Transactions on Antennas and Propagation*, 52(6):1578–1585, 2004.

[67] Mohsen Ghaffari-Miab, Felipe Valdés, Reza Faraji-Dana, and Eric Michielssen. Time-domain integral equation solver for planar circuits over layered media using finite difference generated Green's functions. *IEEE Transactions on Antennas and Propagation*, 62(6):3076–3090, June 2014.

[68] Mohsen Ghaffari-Miab, Reza Faraji-Dana, and Eric Michielssen. Time-domain Green's functions of layered media using modified complex-time method. In *10th European Conference on Antennas and Propagation, EuCAP 2016*, 2016.

[69] Xue Min Xu and Qing H. Liu. The BCGS-FFT method for electromagnetic scattering from inhomogeneous objects in a planarly layered medium. *IEEE Antennas and wireless propagation letters*, 1(1):77–80, 2002.

[70] M. C. van Beurden. Fast convergence with spectral volume integral equation for crossed block-shaped gratings with improved material interface conditions. *Journal of the Optical Society of America A*, 28(11):2269–2278, 2011.

[71] M. C. van Beurden. A spectral volume integral equation method for arbitrary biperiodic gratings with explicit Fourier factorization. *Progress in Electromagnetics Research B*, 36:133–149, 2012.

[72] I. Gohberg and I. Koltracht. Numerical solution of integral equations, fast algorithms and Krein-Sobolev equation. *Numerical Mathematics*, 47(2):237–288, 1985.

[73] M. G. Moharam and T. K. Gaylord. Rigorous coupled-wave analysis of planar-grating diffraction. *Journal of the Optical Society of America*, 73(4):811–818, 1981.

[74] I. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams, and J. R. Andrewartha. The dielectric lamellar diffraction grating. *Optica Acta*, 28(3):413–428, 1981.

[75] Evgeny Popov, Michel Neviere, Boris Gralak, and Gerard Tayeb. Staircase approximation validity for arbitrary-shaped gratings. *Journal of the Optical Society of America A*, 19(1):33–42, 2002.

[76] J. Chandezon, D. Maystre, and G. Raoult. A new theoretical method for diffraction gratings and its numerical application. *Journal of Optics*, 11(4):235–241, 1980.

[77] Lifeng Li, Jean Chandezon, Gerard Granet, and Jean-Pierre Plumey. Rigorous and efficient grating-analysis method made easy for optical engineers. *Applied optics*, 38(2):303–312, 1999.

[78] Tuomas Vallius. Comparing the fourier modal method with the C method: analysis of conducting multilevel gratings in TM polarization. *Journal of the Optical Society of America A*, 19(8):1555–1562, 2002.

[79] A. V. Tischenko and A. A. Shcherbakov. General analytical solution for the electromagnetic grating diffraction problem. *Optics Express*, 25(12):13435–13447, 2017.

[80] Lifeng Li and Charles W. Haggans. Convergence of the coupled-wave method for metallic lamellar diffraction gratings. *Journal of the Optical society of America A*, 10(6):1184–1189, 1993.

[81] Lifeng Li. Use of Fourier series in the analysis of discontinuous periodic structures. *Journal of the Optical Society of America A*, 13:1870–1876, 1996.

[82] G. Granet and B. Guizal. Efficient implementation of the coupled-wave method for metallic lamellar gratings in TM polarization. *Journal of the Optical Society of America A*, 13(5):1019–1023, 1996.

[83] Philippe Lalanne and G. Michael Morris. Highly improved convergence of the coupled-wave method for TM polarization. *Journal of the Optical Society of America A*, 13(4):779–784, 1996.

[84] E. Popov and M. Nevière. Maxwell equations in Fourier space: fast-converging formulation for diffraction by arbitrary shaped, periodic, anisotripc media. *Journal of the Optical Society of America A*, 18(11):2886–2894, 2001.

[85] John David Jackson. *Classical Electrodynamics*. Wiley, 2007.

[86] Max Born, Emil Wolf, A. B. Bhatia, P. C. Clemmow, and D. Gabor. *Principles of Optics*. Cambridge University Press, 7 edition, 1999.

[87] M. G. M. M. van Kraaij. *Forward Diffraction Modelling: Analysis and Application to Grating Reconstruction.* PhD thesis, Eindhoven University of Technology, 2011.

[88] M. C. van Beurden. Scattering by periodic dielectric media via a spectral domain-integral equation: maintaining efficiency and accuracy. In *Proceedings of the 2011 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, pages 967–968, sep 2011.

[89] Thomas Schuster, Johannes Ruoff, Norbert Kerwien, Stephan Rafler, and Wolfgang Osten. Normal vector method for convergence improvement using the RCWA for crossed gratings. *Journal of the Optical Society of America A*, 2007.

[90] Martijn C. van Beurden, Teis J. Coenen, and Irwan D. Setija. Parametric modeling of doubly periodic dielectric structures via a spectral-domain integral equation. In *Proceedings of the 2013 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, 2013.

[91] M. C. van Beurden and I. D. Setija. Local normal vector field formulation for periodic scattering problems formulated in the spectral domain. *Journal of the Optical Society of America A*, 34(2):224–234, Feb 2017.

[92] M. C. van Beurden. Fast convergence with spectral volume integral equation for crossed block-shaped gratings with improved material interface conditions. *Journal of the Optical Society of America A*, 28(11):2269–2278, 2011.

[93] D. Slepian and H. O. Pollack. Prolate spheroidal wave functions, Fourier analysis and uncertainty I. *Bell System Technical Journal*, 40(1):43–65, 1961.

[94] Christopher Kurcz. *Fast convolutions with Helmholtz Greens's functions and radially symmetric band-limited kernels.* ProQuest Information and Learning Company, 2008.

[95] Youcef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3), 1986.

[96] Y Saad. *Iterative methods for sparse linear systems.* PWS, 1996.

[97] G. L. G. Sleijpen and D. R. Fokkema. BiCGstab(ℓ) for linear equations involving unsymmetric matrices with complex spectrum. *Electron. Trans. Numer. Anal.*, 1(1):11–32, 1993.

[98] G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema. BiCGstab(ℓ) and other hybrid Bi-CG methods. *Numerical Algorithms*, 7(1):75–109, 1994.

[99] Hans G. Feichtinger and Thomas Strohmer. *Gabor Analysis and Algorithms: Theory and Applications.* Birkhauser, 1998.

[100] Martin J. Bastiaans. Gabor's expansion and the Zak transform for continuous-time and discrete-time signals: Critical sampling and rational oversampling (online booklet: alexandria.tue.nl/extra1/erap/publichtml/9505415.pdf), 1995.

[101] Dennis Gabor. Theory of communication. *J. Inst. Elec. Eng.*, 93(3):429–457, 1946.

[102] Martin J. Bastiaans. A sampling theorem for the complex spectrogram, and Gabor's expansion of a signal in Gaussian elementary signals. *Optical engineering*, 20(4):594–598, 1981.

[103] Ingrid Daubechies. The wave transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990.

[104] R. Balian. Un principe d'incertitude fort en theorie du signal ou en mecanique quantique. *Comptes Rendus de l'Académie des Sciences. Série II. Mécanique, Physique, Chimie, Sciences de la Terre et de l'Univers*, 1981.

[105] F. Low. *A Passion for Physics - Essays in Honor of Geoffrey Chew*, chapter Complete sets of wave packets., pages 17–22. World Scientific, Singapore, 1985.

[106] Guy Battle. Heisenberg proof of the Balian-Low theorem. *Letters in Mathematical Physics*, 15:175–177, 1988.

[107] John J. Benedetto, Christopher Heil, and David F. Walnut. Differentiation and the Balian-Low theorem. *The Journal of Fourier Analysis and Applications*, 1(4):355–402, 1995.

[108] G. K. Wilson. Generalized wannier functions. *unpublished*, 1987.

[109] Ingrid Daubechies, Stephane Jaffard, and Jean-Lin Journe. A simple Wilson orthonormal basis with exponential decay. *SIAM Journal on Mathematical Analysis*, 22(2):554–572, 1991.

[110] Martin J. Bastiaans. Optimum sampling distances in the Gabor scheme. *Proceedings of the ProRISC Workshop on Circuits, Systems and Signal Processing 1997*, pages 29–35, 1997.

[111] A. J. E. M. Janssen. Signal analytic proof of two basic results on lattice expansions. *Applied and Computational Harmonic Analysis*, 1(4):350–354, 1994.

[112] Z. Landau I. Daubechies, H. Landau. Gabor time-frequency lattices and the wexler-raz identiy. *Journal of Fourier Analysis*, 1995.

[113] Tobias Werther, Yonina C. Eldar, and Nagesh K. Subbanna. Dual gabor frames: Theory and computational aspects. *IEEE transactions on signal processing*, 53(11):4147–4159, Nov 2005.

[114] Martin J. Bastiaans. Gabor's expansion of a signal into Gaussian elementary signals. *Proceedings of the IEEE*, 68(4):538–539, 1980.

[115] J. Wexler and S. Raz. Discrete Gabor expansion. *Signal Processing*, 1990.

[116] Ingrid Daubechies, H. J. Landau, and Zeph Landau. Gabor time-frequency lattices and the wexler-raz identity. *Journal of Fourier Analysis*, 1(4):437–478, 1995.

[117] Amir Shlivinski, Ehud Heyman, Amir Boag, IEEE, and Christine Letrou. A phase-space beam summation formulation for ultrawide-band radiation. *IEEE Transactions on Antennas and Propagation*, 52(8):2042–2056, 2004.

[118] R. J. Dilz and M. C. van Beurden. The Gabor frame as a discretization for the 2D transverse-electric scattering-problem domain integral equation. *Progress in Electromagnetics Research B*, 69:117–136, 2016.

[119] Alexey A. Basharin, Maria Kafesaki, Eleftherios N. Economou, Costas M. Soukoulis, Vassili A. Fedotov, Vassili Savinov, and Nikolay I. Zheludev. Dielectric metamaterials with toroidal dipolar response. *Physical Review X*, 2015.

[120] Céline Ribot, Philippe Lalanne, Mane-Si-Laure Lee, Brigitte Loiseaux, and Jean-Pierre Huignard. Analysis of blazed diffractive optical elements formed with artificial dielectrics. *Journal of the Optical Society of America A*, 2007.

[121] Tilman Glaser, Siegmund Schroter, Hartmut Bartelt, Hans-Jorg Fuchs, and Ernst-Bernhard Kley. Diffractive optical isolator made of high-efficiency dielectric gratings only. *Applied Optics*, 2002.

[122] M. Clemens and T. Weiland. Discrete electromagnetism with the finite integration technique. *Progress in Electromagnetics Research*, 2001.

[123] A. Ye. Poyedinchuk, Yu. A. Tuchkin, , N. P. Yashina, J. Chandezon, and G. Granet. C-method: Several aspects of spectral theorry of gratings. *Progress in Electromagnetics Research*, 2006.

[124] Maxim Pisarenco, Joseph Maubach, Irwan Setija, and Robert Mattheij. Aperiodic Fourier modal method in contrast-field formulation for simulation of scattering from finite structures. *Journal of the Optical Society of America A*, 27(11):2423–2431, 2010.

[125] Peter L. Sondergaard. *Finite Discrete Gabor Analysis.* PhD thesis, Institut for Matematik - DTU, 2007.

[126] John J. Maciel and Leopold B. Felsen. Discretized Gabor-based beam algorithm for time-harmonic radiation from two-dimensional truncated planar aperture distributions - i : Formulation and solution. *IEEE Transactions on Antennas and Propagation*, 50(12):1751–1759, 2002.

[127] John J. Maciel and Leopold B. Felsen. Discretized Gabor-based beam algorithm for time-harmonic radiation from two-dimensional truncated planar aperture distributions - ii : Asymptotics and numerical tests. *IEEE Transactions on Antennas and Propagation*, 50(12):1760–1768, 2002.

[128] P. D. Einziger, S. Raz, and M. Saphira. Gabor representation and aperture theory. *Journal of Journal of the Optical Society of America*, 3(4):508–522, 1986.

[129] John J. Maciel and Leopold B. Felsen. Systematic study of fields due to extended apertures by Gaussian discretization. *IEEE Transactions on Antennas and Propagation*, 37(7):884–892, 1989.

[130] S. J. Floris and B. P. de Hon. Electromagnetic field expansion in a Wilson basis. In *Proceedings of the 42nd European Microwave Conference (EuMC), October 29-November 1 2012, Amsterdam (NL)*, 2012.

[131] Delphine Lugara and Christine Letrou. Printed antennas analysis by a Gabor frame-based method of moments. *IEEE Transactions on Antennas and Propagation*, 50(11):1588–1597, 2002.

[132] G. Szegö. *Orthogonal Polynomials*. Royal American Mathematical Society, 1975.

[133] Constantine A. Balanis. *Advanced Engineering Electromagnetics*. John Wiley & Sons, Inc, 1989.

[134] Sven Burger, Lin Zschiedrich, Jan Pomplun, and Frank Schmidt. Finite-element based electromagnetic field simulations: benchmark results for isolated structures. In *Proc. SPIE 8880 Photomask Technology*, volume 8880, 2013.

[135] R. J. Dilz and M. C. van Beurden. An efficient complex spectral path formulation for simulating the 2D TE scattering problem in a layered medium using Gabor frames. *Journal of Computational Physics*, 345:528–542, 2017.

[136] Alain C Diebold, editor. *Handbook of Silicon Semiconductor Metrology*. CRC Press, 2001.

[137] Young-Nam Kim, Jong-Sun Paek, Silvio Rabello, Sangbong Lee, Jiangtao Hu, Zhuan Liu, Yudong Hao, and William McGahan. Device based in-chip critical dimension and overlay metrology. *Optics Express*, 17(23):21336–21343, 2009.

[138] Saman Jahani and Zubin Jacob. All-dielectric metamaterials. *Nature Nanotechnology*, 11(1):23–36, Jan 2016.

[139] L. Gurel and M. I. Aksun. Fast multipole method in layered medium: 2-D electromagnetic scattering problems. In *Proceedings of the 1996 IEEE Antennas and Propagation Society International Symposium*, pages 1734–1737. IEEE Press, 1996.

[140] K. A. Michalski. Extrapolation methods for Sommerfeld integral tails. *IEEE Antennas and Propagation Magazine*, 46(10):1405–1418, Oct 1998.

[141] H. Derudder, F. Olyslager, and D. De Zutter. An efficient series expansion for the 2D Green's function of a microstrip substrate using perfectly matched layers. *IEEE Microwave and Guided Wave Letters*, 9(12):505–507, 1999.

[142] H. Rogier and D. De Zutter. Berenger and leaky modes in microstrip ssubstrate terminated by a perfectly matched layer. *IEEE Transactions on Microwave Theory and Techniques*, 49(4):712–715, 2002.

[143] Pasi Ylä-Oijala, Johannes Markkanen, Seppo Järvenpää, and Sami P. Kiminki. Surface and volume integral equation methods for time-harmonic solutions of maxwell's equations. *Progress In Electromagnetics Research*, 149(1):15–44, 2014.

[144] P Lalanne, M. Besbes, J.P. Hugonin, S. van Haver, O.T.A. Janssen, A.M. Nugrowati, S.F. Pereira M. Xu, HP Urbach, A.S. van de Nes, P. Bienstman, A Moreau G. Granet, M. Sukharev S. Helfert, T Seideman, B. Guizal F. Baida, and D. van Labeke. Numerical analysis of a slit-groove diffraction problem. *Journal of the European Optical Society - Rapid publications*, 2(0), 2007.

[145] Gerd Ehret, Bernd Bodermann, and Martin Woehler. Comparison of rigorous modelling of different structure profiles on photomasks for quantitative linewidth measurements by means of UV- or DUV-opitcal microscopy. In *SPIE Proceedings Vol. 6617: Modeling Aspects in Optical Metrology*, volume 6617, 2007.

[146] Manuel E. Solano, Muhammad Faryad, Akhlesh Lakhtakia, and Peter B. Monk. Comparison of rigorous coupled-wave approach and finite element method for photovoltaic device with periodic corrugated metallic backreflector. *Journal of the Optical Society of America A*, 31(10):2275–2284, 2014.

[147] R. J. Dilz, M. G. G. M. van Kraaij, and M. C. van Beurden. The 2D TM scattering problem for finite objects in a dielectric stratified medium employing gabor frames in a domain integral equation. *Journal of the Optical society of America A*, 34(8):1315–1321, 2017.

[148] Yi-Sha Ku, Hsiu-Lan Pang, Weite Hsu, and Deh-Ming Shyu. Accuracy of diffraction-based overlay metrology using a single array target. *Optical Engineering*, 2009.

[149] M. Nevière, P. Vincent, R. Petit, and M. Cadilhac. Systematic study of resonances of holographic thin film couplers. *Optics Communications*, 9(1):48–53, 1973.

[150] Yia-Chung Chang, Guangwei Li, Hanyou Chu, and Jon Opsal. Efficient finite-element, Green's function approach for critical-dimension metrology of three-dimensional gratings on multilayer films. *Journal of the Optical Society of America A*, 2006.

[151] Teis J. Coenen and Martijn C. van Beurden. A spectral volume integral method using geometrically conforming normal-vector fields. *Progress in Electromagnetics Research*, 2013.

[152] R. J. Dilz, M. G. G. M. van Kraaij, and Martijn C. van Beurden. A 3D spatial spectral integral equation method for electromagnetic scattering from finite objects. *Optical and Quantum Electronics*, Under review.

[153] C. J. Raymond. *Handbook of Silicon Semiconductor Metrology.* CRC Press, 2001.

[154] Stephan Rafler, Peter Götz, Matthias Petschow, Thomas Schuster, Karsten Frenner, and Wolfgang Osten. Investigation of methods to set up the normal vector field for the differential method. In *Proc. SPIE 6995, Optical Micro- and Nanometrology in Microsystems Technology II, 69950Y*, 2008.

[155] P. Götz, T. Schuster, K. Frenner, S. Rafler, and W. Osten. Normal vector method for the RCWA with automated vector field generation. *Optics Express*, 16(22):17295–17301, 2008.

[156] Helmut Böcskei and Augustus J.E.M. Janssen. Gabor frames, unimodularity and window decay. *The Journal of Fourier Analysis and Applications*, 6(3):255–276, 2003.

[157] Roeland J. Dilz and Martijn C. van Beurden. Scaling of a spatial spectral integral-equation method for EM scattering in a stratified medium to large, finite objects. In *Progress in Electromagnetics research symposium (PIERS), 22-25 May 2017, Saint Petersburg, Russia*, 2017.

[158] Roeland J. Dilz and Martijn C. van Beurden. Computational aspects of a spatial-spectral domain integral equation for scattering by objects of large longitudinal extent. In *2017 International Conference on Electromagnetics in Advanced Applications (ICEAA), 11-15 September 2017, Verona, Italy*, 2017.

[159] R. J. Dilz and M. C. van Beurden. An efficient spatial spectral integral-equation method for EM scattering from finite objects in layered media. In *2016 International Conference on Electromagnetics in Advanced Applications (ICEAA), 19-23 September 2016, Cairns, Australia*, pages 509–511, 2016.

[160] Roeland J. Dilz and Martijn C. van Beurden. Fast operations for a Gabor-frame based integral equation with equidistant sampling. *IEEE Antennas and wireless propagation letters*, Accepted.

[161] I. S. Berezin and N. P. Zhidkov. *Computing Methods.* Pergamon Press, 1965.

[162] M. C. van Beurden. Methods and apparatus for calculating electromagnetic scattering properties of a structure and for reconstruction of approximate structures, 2013.

[163] M. C. M. C. van Beurden, T. Zacharopoulou, A. Roc'h, and M. G. G. M. van Kraaij. A pseudo-spectral longitudinal expansion in a spectral domain integral equation for scattering by periodic dielectric structures. In *Proceedings of the 2015 International Conference on Electromagnetics in Advanced Applications (ICEAA), 7-11 September 2015, Torino, Italy*, pages 1419–1422, 2015.

[164] Michael K. NG and Jianyu Pan. Approximate inverse circulant-plus-diagonal preconditioners for Toeplitz-plus-diagonal matrices. *SIAM Journal on Scientific and Statistical Computing*, 32(3):1442–1464, 2010.

[165] R. F. Remis. Preconditioning techniques for domain integral equations. In *Electromagnetics in Advanced Applications (ICEAA), 2013 International Conference on*, Torino, Italy, 2013. IEEE.

[166] Roeland J. Dilz and Martijn C. van Beurden. A 3D spatial-spectral integral equation method for electromagnetic scattering from finite objects in a layered medium. In *Optical Wave and Waveguide Theory and Numerical Modelling (OWTNM), 5-6 April 2017, Eindhoven, The Netherlands*, 2017.