Efficient Maxwell solver using Gabor frames and preconditioning with applications to optical metrology and integrated photonics

Efficient Maxwell solver using Gabor frames and preconditioning with applications to optical metrology and integrated photonics

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus prof.dr. S.K. Lenaerts, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op dinsdag 24 oktober 2023 om 13:30 uur

door

Ligang Sun

geboren te Shandong, China

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

voorzitter:	prof.dr.ir. J.P.M. Voeten
promotor:	prof.dr.ir. M.C. van Beurden
co-promotor:	dr. R.J. Dilz
leden:	prof.dr.ir. G. Gerini
	prof.dr. I.D. Setija
	prof.dr.ir. I.M. Vellekoop (Universiteit Twente)
	dr. X.J.M. Leijtens
	dr. F. Alpeggiani (ASML Netherlands B.V.)

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

This work was funded by NWO-TTW as part of the HTSM program under project number 16184.

In-cash and in-kind co-funding was provided by ASML Netherlands B.V. with further in-kind support from VSL.

Efficient Maxwell solver using Gabor frames and preconditioning with applications to optical metrology and integrated photonics
Ligang Sun
Eindhoven University of Technology, 2023

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-5844-5
NUR: 959

This thesis was prepared with the $\LaTeX\,2_{\varepsilon}$ documentation system
Reproduction: Ipskamp Printing, Enschede, The Netherlands
Cover design: Xiaoshi Li and Ligang Sun
Image courtesy of Hangzhou Huitu Network Technology Co., Ltd.
The legal permission to reproduce this image was obtained by Ligang Sun

『折而不撓，終不為下。』

——《三國志・先主傳》

The aphorism on the preceding page is quoted from the Records of the Three Kingdoms, an official and authoritative Chinese imperial history book written by Chen Shou (233-297 CE). This book covers the end of the Han dynasty (184–220 CE) and the following Three Kingdoms period (220–280 CE).

The direct translation of this aphorism is "Never give up despite repeated setbacks. Ultimately never yield to others." These words have been used in the final assessment of Liu Bei (161-223 CE), the founding emperor of Shu Han, one of the Three Kingdoms of China.

Dedicated to my motherland.

Contents

Summary

In modern wafer metrology, there is a strong demand for a Maxwell solver that is precise, efficient, and sufficiently flexible to tackle complex geometries. The recently developed spatial spectral method has been successfully applied to solve both 2D and 3D scattering problems in a layered medium. Gabor frames are used to perform discretization in the directions perpendicular to the stratification of the background medium. The Gabor frames yield a fast connection between the spatial domain and the spectral domain. The spatial spectral method incorporates the spectral-domain Green function for a stratified medium with a special path to avoid poles and branch cuts. Additionally, a normal-vector-field formulation is employed to enhance accuracy. Numerical experiments demonstrate that this spatial spectral method maintains an $O(N \log N)$ complexity for the matrix-vector product in computation time while maintaining good accuracy.

Electromagnetic scattering problems can become much more difficult for more complex geometries and higher optical contrasts. Therefore, it is advantageous to push the boundaries of this spatial spectral method such that more challenging real-world and industry-based problems can be simulated and solved. This is the goal of the research work under consideration. We focus on the following challenges. The first challenge is how to efficiently represent scattering objects with complex geometry in terms of Gabor frames. The corresponding research question can be stated as how to compute the pertaining Gabor coefficients accurately and efficiently. We concentrate on computing the Gabor coefficients of a 2D indicator function supported on a polygon, because polygon has the flexibility to accurately approximate an arbitrary 2D shape. The second challenge in the spatial spectral method faces is to deal with scatterers with high contrast or negative permittivity. The pertaining matrix system becomes hard to solve iteratively due to its unfavorable eigenvalue distribution, which results in a large number of iterations when using Krylov subspace methods. The third challenge concerns cases where the scatterers are distributed over a large-scale transverse structure and for more realistic industry-based applications.

To approximate the cross section of a 3D object with complex geometry, an N-sided polygon can be used and the approximation can be refined by increasing the number of edges N. Therefore, to increase the geometrical flexibility of the spatial spectral solver, we focus on the computation of Gabor coefficients of a 2D indicator function supported on an N-sided polygon. Two analytical methods are proposed to compute the fundamental integrals associated with the Gabor coefficient after applying Gauss's theorem. In the first method, the fundamental integral is formulated with the complex error function in the integrand. The complex error function is approximated by its truncated Taylor series expansion. Aiming at accelerating the computation, a second-order inhomogeneous difference equation is derived and solved by Olver's algorithm. However, this method needs a very high working precision, which results in limited practical use. In the second method we overcome the disadvantages caused by the Taylor series expansion. We reformulate the fundamental integral with the Faddeeva function in the integrand and a rational expansion for the Faddeeva function is used. We take advantage of its global fast convergence on the whole complex plane and employ the fast Fourier transform. A second-order inhomogeneous difference equation is again derived, which is also solved by Olver's algorithm. Numerical examples demonstrate that the rational-expansion-based method surpasses numerical quadrature in computation time while preserving accuracy.

When extending the application of the spatial spectral method to high-contrast scattering problems, numerically unreliable results were observed. We identify that the original multiplication operator for two sets of Gabor coefficients causes this problem when the contrasts become large. Hence, one of the primary objectives is to construct a stable multiplication operator, as it plays a crucial role in the spatial spectral method. An improved multiplication function is proposed to multiply two sets of Gabor coefficients, and its analog for two equidistantly sampled discrete functions is provided. The key step is to identify the range of the Gabor coefficients that is accurate and apply zero-padding and restriction operations. With this improved multiplication operator, we are able to solve a scattering problem with a contrast of 16 with a relative error of 10^{-3} in the far field, against an independent reference.

An unfavorable eigenvalue distribution of the matrix system occurs when the scatterers have high contrast or negative permittivities. To tackle this problem, we propose a normal-vector-field-based block-diagonal-preconditioner in to be used with Krylov subspace methods BiCGstab(ℓ) or IDR(s). A spectral analysis reveals that the preconditioned system possesses a more clustered distribution of eigenvalues as compared to the unpreconditioned system. In numerical experiments involving high contrast or negative permittivity, it is observed that the preconditioned system outperforms the unpreconditioned system in terms of the number of iterations. Consequently, the total computation time is reduced while maintaining accuracy.

The thesis is completed by three applications. The first application results from optical metrology: a single-pad resist-only metrology target illuminated by more than 9000 optical beams. This scattering problem has been simulated with the spatial spectral solver. The solutions are analyzed and compared with a periodic solver. Numerical results suggest that the improved spatial spectral Maxwell solver is able to solve relevant scattering problems in optical metrology without artifacts due to periodicity. The second application comes from integrated photonics: a computer-generated waveguide hologram device illuminated by a beam-type source. The hologram area contains 10126 bar-type scatterers in a $36.8\lambda \times 33\lambda \times 0.03\lambda$ domain with λ being the free-space wavelength. The third application concerns scattering by a metasurface. Numerical results from the spatial spectral method suggest that the height of the scatterer is a critical parameter for the number of iterations of the Krylov subspace solver. The full-scale problem is currently still beyond the capability of the solver.

List of abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
ABC	absorbing boundary condition
AFM	atomic force microscopy
BEM	boundary element method
BTTD	block-Toeplitz-Toeplitz-block
BVP	boundary value problem
CBFM	characteristic basis function method
CCD	charge-coupled device
CD	critical dimension
CD-SAXS	critical dimension small-angle X-ray scattering
CD-SEM	critical dimension scanning electron microscopy
CDU	uniformity of the critical dimension
CFIE	combined field integral equation
CFL	Courant–Friedrichs–Lewy
CG	conjugate gradient
CGWH	computer-generated waveguide hologram
DBO	diffraction-based overlay
DUV	deep ultraviolet
EFIE	electric field integral equation
EUV	extreme ultraviolet
FDTD	finite difference time domain
FEM	finite element method
FFT	fast Fourier transformation

FIT	finite integration technique
FMM	fast multipole method
FT	Fourier transformation
GMRES	generalized minimum residual
ICs	integrated circuits
InP	indium phosphide
MFIE	magnetic field integral equation
MINRES	minimum residual
MLFMA	multilevel fast multipole algorithm
MoM	method of moments
MVP	matrix-vector product
NVF	normal vector field
NVF-BD	normal-vector-field block-diagonal
ODE	ordinary differential equation
PDE	partial differential equation
PMCHWT	Poggio-Miller-Chang-Harrington-Wu-Tsai
PML	perfectly matched layer
PWL	piecewise linear
RCWA	rigorous coupled-wave analysis
RE-DE	rational-expansion and difference-equation based
RWG	Rao–Wilton–Glisson
SIE	surface integral equation
TE	transverse electric
TM	transverse magnetic
VIE	volume integral equation

Chapter 1

Introduction

1.1 Lithography and optical metrology

Microchips, or integrated circuits, were invented by Jack Kilby and Robert Noyce in the end of the 1950s [1], and they have become the heart of modern technology nowadays. A microchip is a module of packaged electronic circuits on one small flat piece of semiconductor material, usually silicon. On the chip, transistors are used as electrical switches to turn a current on or off, or as means to amplify or modulate a signal. In general, a larger number of transistors integrated on the chip or a smaller feature size of each transistor, yields improvements in almost every aspect of a microchip [2, 3], such as lower power consumption, a larger memory capacity, and a higher processing speed. In the early 1970s, a microprocessor usually contained thousands of transistors on a single microchip. Today, a microchip the size of a human fingernail can contain billions of transistors, and its features are measured on a nanometer scale [4, 5].

More than one trillion microchips are manufactured around the world annually and the whole semiconductor industry forms a US$ 618 billion market in 2022 with continuing growth [6]. It is the advantage in mass production capability, reliability, and the building-block approach in design that made semiconductors revolutionize the world of electronics [7, 8]. From smartphones and computers to automotive and industrial equipment, from servers, data centers, and storage to diverse consumer electronics, semiconductors bring a host of applications and constitute the fundament of modern technology since the digital revolution began in the second half of the 20th century. Furthermore, semiconductor technology is playing a crucial role in the trends of shaping an interconnected and intelligent society, and contributing to the growing market of artificial intelligence, 5G, virtual reality, the internet of things and cloud computing. A range of application fields of the modern semiconductor industry are given in Fig. 1.1.

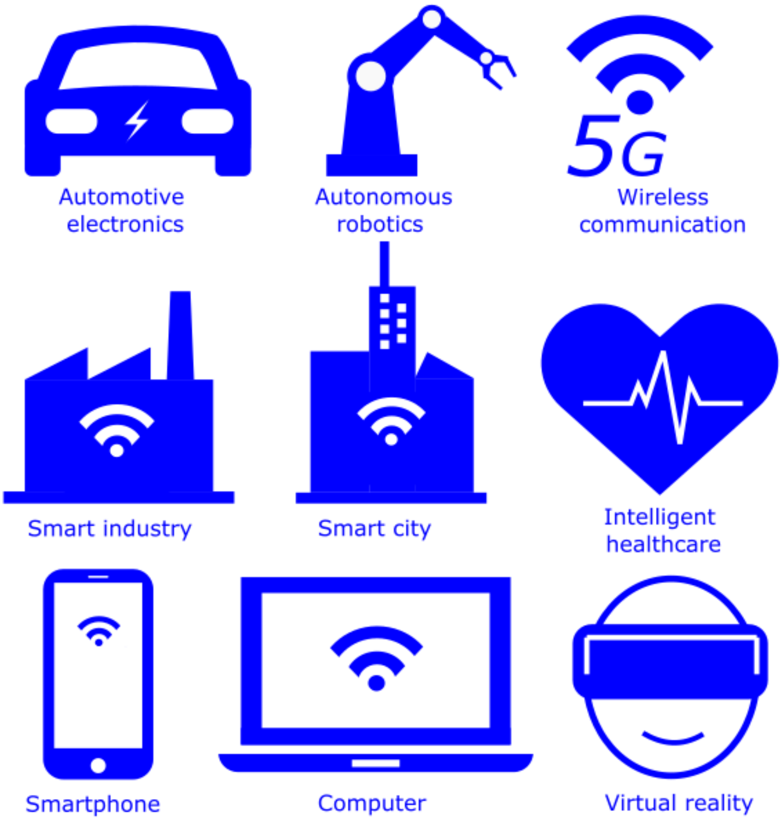


Figure 1.1: Semiconductor industry application fields.

All the above widespread semiconductor applications in Fig. 1.1 require reliable and efficient mass production of microchips. Microchips are manufactured by building up multiple patterned layers on a silicon wafer. When patterned layers are manufactured, the process can require hundreds of steps and takes up to several months [9]. Following the descriptions in [3, 10], we outline the main steps of microchip manufacturing as a recipe in Fig. 1.2. The manufacturing process begins with a silicon wafer, which is a thin slice of pure silicon polished to extreme smoothness. To manufacture a microchip, the first important step is deposition, where thin films of materials such as conductors, isolators, and semiconductors are deposited onto the silicon wafer. The wafer is then covered with a light-sensitive layer called a photoresist. There are positive and negative photoresists, and the positive resist is mainly used in semiconductor production due to its high-resolution capability. The next step is lithography. The intended details to be printed on a wafer are contained in a patterned plate made of glass or quartz, which is called a reticle or photomask. During this step, deep ultraviolet (DUV) or extreme ultraviolet (EUV) light is projected onto the wafer through the reticle. The lithography system's optics shrinks and focuses the patterns on the reticle onto the photosensitive resist layer of the wafer and this induces the pre-designed patterns in the resist layer by chemical changes. This step is repeated until the whole wafer is covered with these patterns. After the lithography step, the wafer is baked to make the resist insensitive to further illumination and subsequently developed such that printed patterns are represented by the resist being present or absent according to the printed pattern. Part of the resist layer is also washed away to generate a pattern of open space with material underneath. The next step is chemical processing called etching. The material that is not protected by the resist on top of it is etched away with either gases or chemical baths, so a 3D version of the pattern is produced. The following step is ion implantation. The wafer may be bombarded with positive or negative ions of doping material to create electrically conducting properties according to the patterns. Once the layer is doped, the remaining photoresist is removed. The above steps are repeated to make more layers with patterns until the whole structure is finished.

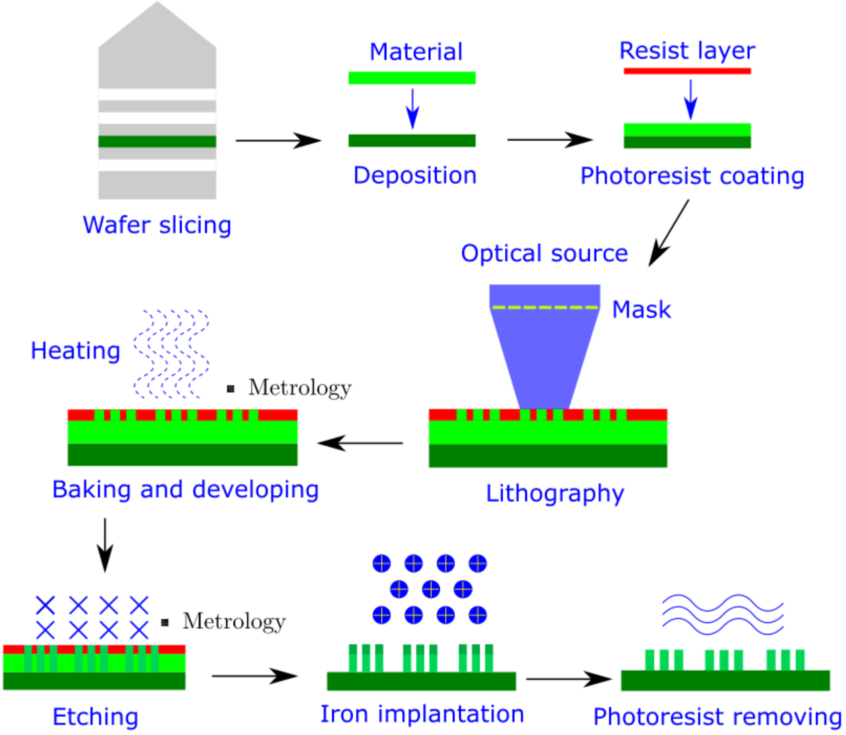


Figure 1.2: Main steps in microchip manufacturing.

One important step during the manufacturing of microchips is metrology. Metrology is the process to measure and characterize the actual features of printed patterns. Small effects during the manufacture can cause catastrophic deviations in the chip's functional behavior. These effects can be reticle defectivity (a tiny imperfection on the wafer such as particle interference, refraction, or other physical or chemical defects) [11, 12], or improper illumination (e.g. the position or angle of the light source was assigned incorrectly) during lithography. Therefore, metrology is an essential step to monitor the quality of the manufacturing process, detect and measure errors, and check if certain minimum quality standards are satisfied. There are several popular wafer-metrology technologies: critical dimension scanning electron microscopy (CD-SEM), critical dimension small-angle X-ray scattering (CD-SAXS), scatterometry, 3D atomic force microscopy (AFM), transmission electron microscopy (TEM), see e.g. [13] a comparison of these technologies. Among all these technologies, optical metrology uses light-based methods to measure and characterize the properties of semiconductor devices [7, 14], and it has been proven a powerful method due to the following features: (1) it can achieve high accuracy, e.g., interferometry is considered as the most accurate measurement technology in modern optical metrology [15], (2) it has a high speed of measurement [16–18], which is important in mass production of semiconductor devices, (3) it is a non-destructive technique and therefore the measuring does not jeopardize the printed structure nor contaminate the material on the wafer [19].

The smallest feature size that can be reliably printed onto a wafer during the lithography process is called the critical dimension. The critical dimension directly determines the size of transistors that can be created on the microchip, which makes it a crucial parameter in the final performance of semiconductor devices. Today's advanced microchips contain billions of transistors [20] or hundreds of layers [21], and the fine details of those chips are in the range of several nanometers. When the components on a microchip or the printed detail have become much smaller than the illumination wavelength or even half the wavelength, it is impossible for conventional camera imaging to directly capture the fine features of the printed patterns in optical metrology, due to Abbe's diffraction limit [22]. To obtain the fine details through such optical images, an extra target-reconstruction process is needed, where the scattering images of a sequence of given targets are computed based on numerically solving Maxwell's equations. This target-reconstruction process is usually performed based on a Maxwell solver. Insights yielded from the Maxwell solver can be used to understand the scatterometric optical signal from the structure, which is also beneficial for developing new metrology hardware and algorithms in optical metrology [23]. Hence, high-performance optical metrology relies significantly on an accurate and efficient Maxwell solver.

In the following sections, we will review the conventional Maxwell-solver approaches in computational electromagnetics and then recall a recently developed Maxwell solver based on the spatial spectral method given in [24], which forms the foundation for the thesis at hand.

1.2 Numerical methods in computational electromagnetics

Since the second half of the 20th century, a variety of numerical techniques in computational electromagnetics have been developed for solving electromagnetic problems. The capabilities of methods developed in computational electromagnetics got strengthened by the evolution of high-performance computing facilities. Computational methods can be classified into different types, based on several features, depending on if there is a time-harmonic assumption in the governing equations or not. Methods with such an assumption are classified as frequency-domain methods and those without are time-domain methods.

If a Fourier transformation is performed from the spatial domain to the spectral domain, these methods are classified as real-space methods or Fourier-space methods and depending on which form of Maxwell's equations is used, these methods are classified as differential-equation methods or integral-equation methods. A numerical method in an electromagnetic field solver may possess one or more of the above-mentioned features and a hybrid method is usually a combination and modification of several of the above methods. Here we briefly review the most widely-used numerical methods in computational electromagnetics.

1.2.1 FDTD

One of the widely used method is the finite difference time domain (FDTD) method. FDTD was proposed by Yee in 1966 [25], and the key idea in FDTD is to use finite differences to approximate the derivatives in the differential form of Maxwell's equations. In his original paper [25], Yee used the leapfrog scheme to approximate the electric field and magnetic field, resulting in local second-order accuracy in both spatial domain and time domain. Later on, an analysis on numerical stability was given in [26, 27] by Taflove. Boundary conditions must be considered carefully when using FDTD-type techniques. When the FDTD method is used to solve open-space scattering and radiation problems, the computational domain must be truncated due to the limitation of computers. An absorbing boundary condition (ABC) is usually applied to truncate the computational domain. In 1994, Berenger proposed a perfectly matched layer (PML) [28, 29] and it has proven to be one of the most robust material ABCs [30–32]. One important property of the FDTD method is its conservation of energy since it ensures that the numerical solution is physically valid [33]. A survey shows that the number of FDTD-related publications exhibits an exponential growth since the 1980s [34], and the FDTD method is one of the most widely used methods in computational electromagnetics today.

There are several reasons that make the FDTD method successful in solving electromagnetic problems. First of all, the explicit derivation in FDTD yields simplicity and elegance, which is useful for understanding and implementation when modeling an electromagnetic problem. Secondly, FDTD method has the intrinsic ability to handle complicated geometries and inhomogeneous materials by specifying material parameters on the entire computational domain. Thirdly, FDTD is a time-domain method and therefore the solution of an FDTD-based system contains not only the electric and magnetic field on the computational domain, but also their evolution with respect to time. One disadvantage of the FDTD method is that it becomes inefficient when the object geometry does not conform with the (commonly used) rectangular grid in the spatial domain. [1] Another drawback is that the FDTD method is expensive in terms of computational resources when the computational domain is large, due to the following two reasons. On one hand, the FDTD algorithm has an inherent numerical-dispersion error, and the spatial discretization length of the grid scheme must be fine enough to keep the numerical-dispersion error small [35, Chapter 12]. On the other hand, for explicit FDTD schemes, the time step and the space step must together obey the Courant–Friedrichs–Lewy (CFL) condition to

[1] The finite element method is more suitable for such problems.

guarantee the stability [26, 33]. These two factors may result in a significant increase in the cost of obtaining the solution in terms of e.g. memory and computation time.

Since invented, more than a half-century ago, FDTD-type techniques have been successfully used to solve electromagnetic problems such as in wafer metrology [36–38], optical fibers [39], radar-cross-section calculation [40, 41], various antennas [42–44], waveguides [45, 46], microstrips [42], etc. The body of applications of the FDTD method is so large, and it is still expanding into other non-electromagnetic areas such as acoustics [47–49], biology [50, 51] and machine learning [52, 53].

1.2.2 FIT

The finite integration technique (FIT), which was proposed by Weiland [54, 55], is another popular time-domain method used in computational electromagnetics. Apart from the fact that FDTD is based on the differential form of Maxwell's equations, the FIT discretizes the integral form of Maxwell's equations.

In the spatial domain, FIT performs the discretization by a staggered grid, where the original grid is coupled with another dual grid. In a case of dual orthogonal grid, FIT results in a numerical scheme that is equivalent to the FDTD system [35]. A non-orthogonal grid can also be used in FIT, which yields higher accuracy and extended flexibility on complicated geometries [56]. FIT uses the leapfrog scheme or other second-order implicit methods in the time domain. Together, the FIT method transforms Maxwell's equations into a dual grid cell complex, and the algebraic properties of the derived linear equations based on FIT also lead to the conservation of energy and charge [57].

FIT shares many numerical characteristics with FDTD, such as simplicity in understanding and implementation. FIT also encounters similar difficulties that occur in FDTD. Much like Yee's algorithm in the FDTD method, FIT becomes inflexible when a Cartesian grid is used to discretize a complex geometry, and also the CFL condition must be satisfied in the standard explicit FIT method to generate a stable discretization scheme. Proper absorbing boundary conditions (ABCs) and perfectly matched layers (PMLs) should also be applied in FIT to model an unbounded domain.

Today, FIT serves as a basis for CST Studio Suite [58], a commercial simulation tool in computational electromagnetics. Interesting applications of FIT in computational lithography can be found in [59].

1.2.3 FEM

The initial work on the finite element method (FEM) can be traced back to A. Hrennikoff when solving differential equations in elasticity [60] and R. Courant when solving variational problems in potential theory [61] in the earlier 1940s. Other pioneers of FEM include J. Argyris in Germany, L. Oganesyan in USSR and K. Feng in China. The key idea of FEM is to represent the solution with given basis functions and unknown coefficients on small elements, thus converting a partial differential equation (PDE) problem into a set of algebraic problems [35, 62].

In electrical engineering, FEM was first applied to solve a classical waveguide-mode problem in 1969 [63]. Later on, this approach attracted a lot of attention and a significant amount of effort was given on improving this approach. Nowadays, FEM has become a very powerful numerical method in computational electromagnetics, with many applications such as antenna radiation [64, 65], microwave engineering [66], optical waveguides [67, 68], etc. Even though it is possible to use FEM to solve Maxwell's equations in the time domain [69, 70], FEM is more widely used as a frequency-domain method to solve electromagnetic problems.

One of the reasons that makes FEM widely used is its flexibility on problems with a complex structure or with inhomogeneous media. Various geometrical elements have been developed to discretize the computational domain, e.g. in 3D problems the right-angled bar, skewed bar, tetrahedron, cylindrical shell, etc. are used [62]. There are two main types of elements in FEM analysis: node-based elements and edge-based elements [71, 72]. The traditional node-based elements are usually easier to implement, while the edge-based elements model the electric and magnetic fields more accurately and result in sparser system matrices [62]. Analogous to other PDE approach like FDTD, truncating the mesh correctly on the boundary of the computational domain is a crucial part to improve the accuracy. ABCs and PMLs are used to suppress wave reflections back into the computational domain.

A major difference between FDTD and FEM is that the former approximates the differential operators while the latter approximates the functions or fields themselves. FEM usually yields a sparse linear system, which can be solved efficiently with direct methods (e.g., using H-matrices [73]) or iterative techniques (e.g., conjugate gradient method [74]). Another difference is that, unlike the rectangular grids commonly used in the FDTD, more flexible meshes can be used in FEM to conform with the object surface. Furthermore, the adaptive mesh can be generated to refine the discretization locally and increase the accuracy.

1.2.4 MoM

The method of moments (MoM), sometimes called the boundary element method (BEM), is another widely used frequency domain approach in computational electromagnetics. Unlike differential-form-based methods like FEM or FDTD, MoM is based on an integral-representation form of Maxwell's equations. The idea of MoM can be traced back to Galerkin's work in the 1920s and this method has arisen much attention after its formulation was presented by Harrington in 1967 [75]. The MoM transforms the boundary value problem (BVP) of Maxwell's equations into a dense matrix equation, by handling the integral representation of Maxwell's equations with Green's functions. MoM can be used to solve both volume integral equations (VIEs) and surface integral equations (SIEs) [35]. In case of a SIE for a perfectly conducting object, the integral equation can be the electric field integral equation (EFIE), the magnetic field integral equation (MFIE), or the combined field integral equation (CFIE). For homogeneous penetrable objects, various formulations exist, among which the Müller [76] and Poggio-Miller-Chang-Harrington-Wu-Tsai (PMCHWT) [77–79] formulations are the most popular ones.

At an abstract level, MoM can be summarized into the following steps. Firstly, the unknown equation is represented by a summation of a finite set of weighted basis functions. Commonly used basis functions include pulse functions, piecewise-linear or hat functions, piecewise sinusoidal functions [35,80,81] for 1D cases, and the famous Rao–Wilton–Glisson (RWG) basis function [82] for 2D surfaces. Secondly, the substitution of these approximations for the unknown function is performed in the integral equation, which incorporates a Green function that is specific for the background medium and the boundary conditions. After a testing procedure, we arrive at a set of linear equations or a matrix equation. Thirdly, various numerical methods can be used to solve the matrix equation, including direct methods such as LU factorization and Krylov-subspace iterative methods (e.g., conjugate gradient method [74], BI-CG-based methods [83–85] and the IDR(s) method [86]).

Since the 1960s, MoM has been developed further and has applied to wave scattering and antenna radiation problems [35,87,88], and it is still one of the predominant numerical methods in computational electromagnetics today. There is no ABCs or PMLs needed in MoM, which makes it efficient to solve open-region electromagnetic problems. Furthermore, the method of moments is much more efficient when it deals with an SIE for problems involving piecewise homogeneous regions [35], since the number of unknowns from a surface discretization is much less than its counterpart in a VIE. The capabilities of MoM have been significantly extended by fast algorithms such as the fast multipole method (FMM) [89,90] and the multilevel fast multipole algorithm (MLFMA) [91–93]. By using low-rank approximations for the MoM matrix, FMM speeds up the computation of the matrix-vector product (MVP), and MLFMA further reduces the computational complexity to $O(N \log N)$. Another important development in MoM is the characteristic basis function method (CBFM) [94], which can be highly parallelized and combined with MLFMA [95,96].

1.2.5 RCWA

The rigorous coupled-wave analysis (RCWA) is a Fourier space method for periodic scatterers. RCWA is semi-analytical method which is based upon a mode expansion method [97,98]. The key idea of RCWA is to compute the mode propagation along the longitudinal direction and expand the periodic relative permittivity function as well as the electric and magnetic fields into a summation of harmonic waves in the transverse plane. RCWA seeks a solution that satisfies Maxwell's equations in the computational domain and matches the boundary conditions at the interfaces in the transverse direction.

In [97], RCWA is shown to be inherently stable, via a criterion of energy conservation and achieved convergence of the solution, with an accuracy that depends only on the number of terms in the Fourier expansion. RCWA usually requires solving eigenvalue problems along the vertical direction, and results in a linear system for each horizontal interface and which can be solved directly, i.e., without using iterative methods.

RCWA was first invented to solve a planar-grating diffraction problem by Moharam and Gaylord in 1981 [99], and it has been extended to analyze various transmission and reflection grating diffraction problems, see e.g. [100–102]. RCWA has been shown very suitable for solving lamellar metallic and dielectric gratings problems. However, RCWA handles the gratings whose boundaries are parallel to the longitudinal direction and the transverse plane more efficiently than gratings with an arbitrary profile. For instance, when the grating has a sinusoidal profile, RCWA with staircase approximation exhibits a slower convergence as compared to the differential method [103, 104].

1.3 Research challenges

In [24], a spatial spectral method was proposed to solve electromagnetic scattering problems in a dielectric layered medium. The spatial spectral method is a frequency-domain method based on a domain-integral equation. Furthermore, the spatial spectral method is a hybrid method, where the Gabor transformation is used to perform discretization in both the spatial domain and spectral domain.

The spatial spectral method has the following features:

- A Gabor-frame-based discretization is used to represent all involved functions in the domain-integral equation, in the two directions perpendicular to the stratification of the background medium, also known as the transverse plane. Due to the inherited property of Gabor frame functions, a fast and exact transformation is established for these functions between the spatial and spectral domains. In the direction of the stratification, also known as the longitudinal direction, a set of piecewise linear (PWL) functions is used as basis functions.

- A normal vector field (NVF) formulation is used in the field-material operator to improve the system's accuracy. Proposed by [105] for periodic problems formulated in the spectral domain, the normal vector field formulation defines a multiplication satisfying the Li factorization rules [106], which yields better accuracy. The spatial spectral method also employs an NVF formulation [107], and improved accuracy has been observed [108].

- An analytical expression for the Green function for a stratified medium exists only in the spectral domain, which is the main reason to go to the spectral domain in this spatial spectral solver, since the tedious Sommerfeld integrals to obtain the spatial-domain Green function can be avoided. A special path in the spectral domain is carefully chosen to avoid poles and branch cuts of the Green function along the integration path [108–110].

- To improve computation speed, an alternative set of basis functions and a set of Dirac-delta function are proposed for 3D problems. These functions yield an equidistant list-based representation in both spatial and spectral domain. With this, the multiplication operation becomes pointwise and an FFT-based Fourier transformation is developed [111].

The spatial spectral method has been successfully applied on multiple electromagnetic problems. For 2D cases, this spatial spectral method was used to solve a problem with two bar-shaped scatterers in a layered medium and a grating coupler problem, in both TE [109] and TM polarization [108]. For 3D cases, this spatial spectral method was applied on a single scatterer with different shapes [110], and a large grating consisting of multiple scatterers [110, 111]. Numerical experiments show that this spatial spectral method yields good accuracy when compared with other numerical references, and it reaches an $O(N \log N)$ complexity for the MVP in the sense of computation time. Specifically, the large 3D grating problem in [110, 111] can be effectively solved using the spatial spectral method on a regular computer, whereas the FEM-based JCMWave struggles to tackle such a large grating problem with the same computational resources. Additionally, a recent development [112] on the spatial spectral method in combination with analytical geometry parametrization shows that accurate parameter reconstruction in inverse-scattering problems can be reached, owing to the use of continuous Gabor frames instead of local basis functions on a mesh.

In modern optical scatterometry for wafer metrology, an accurate and efficient Maxwell solver is expected to solve the following challenging scenarios in modern wafer metrology: when a scatterer has a complex geometry, when a scatterer has high optical contrast, when scatterers are spread out across the layered medium in the z direction, when scatterers are distributed over a large-scale transverse domain, and when coupled to sensor modeling. The previously mentioned spatial spectral method has shown strengths on solving scattering problems of finite scatterers with a competitive computational complexity. Naturally, it is beneficial to explore the limits of this spatial spectral method on more difficult problems and test its performance on real industry-based problems. Therefore, the main goal of this research is to extend the capability of the previously developed spatial spectral method to a higher level. To be specific, this study focuses on solving the following problems.

1. Develop analytical methods with reduced computational cost to calculate Gabor coefficients of a complex geometry accurately.

2. Expand the current spatial spectral solver's capability to solve cases with higher optical contrast and larger computational domain.

3. Test this method on real industry-driven applications.

Regarding Problem 1: since the spatial spectral solver is a Gabor-frame-based method, Gabor coefficients of electric fields, contrast current density and contrast function must be computed accurately. When the scatterer has a complex structure, heavy numerical integration is involved since usually a special function is contained in the integrand that is expensive to evaluate. Key difficulty in this problem is the development of alternative analytical methods to compute Gabor coefficients. A qualified method should outperform those conventional numerical-integration-based methods in both accuracy and computation time.

Regarding Problem 2: in general, to make the spatial spectral method work on more difficult cases, such as higher contrast scatterers and larger computational domain, improvements on two aspects should be considered. On one hand, the formulation of the original spatial spectral system, such as the field-material interaction operator, should be tested and improved. On the other hand, more powerful iterative methods and preconditioning techniques should be considered to accelerate the convergence rate when solving the system iteratively.

Problem 3 concerns a set of application-driven problems. Potential industry-based applications of the spatial spectral Maxwell solver include a wave-propagation problem in an integrated computer generated waveguide hologram, a large-scale scattering problem of a metalens, and a diffraction-based overlay (DBO) problem in optical metrology. Challenges in this problem include the implementation of an inner-layered incident Gaussian beam source, validation of the simulation results with numerical references. Insights obtained by analyzing the solver's performances on these real cases can help to understand this spatial spectral method better and provide guidance for future improvements.

1.4 Outline of the thesis

This thesis is organized as follows.

In Chapter 2, we start the formulation from Maxwell's equations and then state the electromagnetic scattering problem in a layered medium. We then derive the electric field (domain) integral equation based on the transmission-line equations and the Green function in a layered medium. A summary of the spatial spectral method follows, where we focus on two important aspects of this method: the NVF formulation and the Gabor frames. Last but not least, we discuss how to solve the overall linear system efficiently through advanced iterative methods such as BiCGstab(ℓ) and IDR(s), and preconditioning techniques.

In Chapter 3, we develop a method to numerically calculate the integrals involved in computing Gabor coefficients of a characteristic function supported on a polygonal domain, based on a Taylor series expansion and recurrence relations. We transform the double integral into a sequence of line integrals containing the complex error function as part of its integrand. Then we apply the Taylor series expansion of the complex error function and derive a sequence of fundamental integrals. A second-order inhomogeneous difference equation is derived, and we solve all the fundamental integrals by Olver's algorithm. As an improvement to this method, we introduce three requirements for this method to guarantee accuracy.

Key disadvantages of the method from Chapter 3 are addressed and overcome in Chapter 4. A rational expansion for the Faddeeva function given in [113] is employed, owing to its global fast convergence on the whole complex plane. We re-formulate the fundamental integrals such that they contain the Faddeeva function in their integrands and apply the rational expansions. Another set of second-order linear difference equations is then derived and we solve the system again with Olver's algorithm. This rational-expansion-based method inherits the fast convergence property in [113] and therefore it significantly outperforms direct numerical quadrature in terms of computation time while maintaining accuracy.

In Chapter 5, a modified field-material interaction operator is introduced. Instead of a direct multiplication in the spatial domain, the updated operator contains a pair of

Fourier transformations (FTs) and a restriction operation in the spectral domain. We explain why this pair of FTs is necessary for scattering problems with high contrast and why the restriction operation in the spectral domain is crucial. Numerical experiments on a high-contrast problem show that the 3D spatial spectral solver based on this modified field-material interaction operator yields accurate results when compared with a FEM-based reference.

To reduce the number of iterations used by the iterative method used to solve the linear system of the spatial spectral method, we propose a normal-vector-field block-diagonal(NVF-BD) preconditioner in Chapter 6, for both 2D TM and 3D cases. This preconditioner is directly related to the normal vector field and it has a block-diagonal structure. We show the preconditioned system has a clustered eigenvalue distribution and therefore the preconditioned system has the potential to yield a faster convergence rate of the iterative method. Three experiments with high-contrast scatterers, a negative-permittivity scatterer, and a larger geometrical dimension are studied. Numerical evidence reveals that the number of iterations can be significantly reduced by applying the NVF-BD preconditioner.

Three applications of the improved spatial spectral solver are studied in Chapter 7. In the first one, we simulate the scattering by a finite single-pad resist-only metrology target. The second one is for an integrated computed generated waveguide hologram on a layered medium with an inner-layer Gaussian-beam source. The third application is a metalens scattering problem with plane-wave excitation.

Finally, we conclude this thesis in Chapter 8 by drawing overall conclusions concerning the main results of this thesis and by discussing future research topics in relation to the spatial spectral method.

Chapter 2

Formulation

2.1 Maxwell's equations

We start the formulation from Maxwell's equations, which govern the behavior of the electric and magnetic fields. Let $\mathbf{x}$ be the spatial position vector, t the time variable, $\boldsymbol{\mathcal{E}}(\mathbf{x}, t)$ be the electric field, $\boldsymbol{\mathcal{H}}(\mathbf{x}, t)$ be the magnetic field, $\boldsymbol{\mathcal{J}}(\mathbf{x}, t)$ be the electric current density, $\boldsymbol{\mathcal{D}}(\mathbf{x}, t)$ be the electric flux density, and $\boldsymbol{\mathcal{B}}(\mathbf{x}, t)$ be the magnetic flux density (or magnetic induction). We assume the vectors $\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{B}}$ are continuous and have continuous derivatives, and vanish when $t < 0$. Then the classical differential form of Maxwell's equations is given by:

$$\nabla \times \boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{J}} + \frac{\partial}{\partial t}\boldsymbol{\mathcal{D}}, \tag{2.1}$$

$$\nabla \times \boldsymbol{\mathcal{E}} = -\frac{\partial}{\partial t}\boldsymbol{\mathcal{B}}, \tag{2.2}$$

$$\nabla \cdot \boldsymbol{\mathcal{B}} = 0, \tag{2.3}$$

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

where ϱ is the electric charge density. The above equations are often named after their discoverers: Eq. (2.1) is called Ampère-Maxwell's law, Eq. (2.2) is called Faraday's law of induction, Eq. (2.3) is called Gauss's law for magnetic fields, and (2.4) is called Gauss's law for electric fields.

Although the divergence operator and the curl operator in Maxwell's equations do not rely on the choice of the coordinate system, here we focus on a right-handed Cartesian coordinate system with $\mathbf{x} = (x, y, z)$ in 3D. By taking the divergence on both sides of Eq. (2.1), we get the following equation of continuity:

$$\nabla \cdot \boldsymbol{\mathcal{J}} + \frac{\partial}{\partial t}\varrho = 0, \tag{2.5}$$

which expresses the conservation of charge, i.e. the amount of electric current crossing the boundary enclosing a finite and bounded volume is equal to the rate of decrease of the electric charge inside the volume.

Throughout the thesis, we always use j (instead of i) to denote the imaginary unit, and we define the following Fourier transformations between time domain and frequency domain:

$$\mathbf{E}(\omega) = \mathcal{F}[\boldsymbol{\mathcal{E}}(t)](\omega) = \int_{\mathbb{R}} \boldsymbol{\mathcal{E}}(t)e^{-j\omega t}dt, \tag{2.6a}$$

$$\boldsymbol{\mathcal{E}}(t) = \mathcal{F}^{-1}[\mathbf{E}(\omega)](t) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathbf{E}(\omega)e^{jt\omega}d\omega, \tag{2.6b}$$

where $\omega \in \mathbb{R}$ is the angular frequency. Note that Eq. (2.6a) denotes the 1D (forward) Fourier transformation, Eq. (2.6b) the 1D inverse Fourier transformation, and the integral is performed per component of the vector field $\boldsymbol{\mathcal{E}}$ and $\mathbf{E}$.

By applying the forward Fourier transformation (2.6a) to Eqs (2.1) and (2.2) we have Maxwell's equations in the frequency domain:

$$\nabla \times \mathbf{H}(\mathbf{x},\omega) = \mathbf{J}(\mathbf{x},\omega) + j\omega\mathbf{D}(\mathbf{x},\omega), \tag{2.7}$$

$$\nabla \times \mathbf{E}(\mathbf{x},\omega) = -j\omega\mathbf{B}(\mathbf{x},\omega), \tag{2.8}$$

where ω becomes a constant in the case of harmonic time dependence.

To make the system (2.1)-(2.4) uniquely solvable, we must supplement Maxwell's equations by boundary conditions and constitutive relations concerning the properties of the material media [114]. The constitutive relations are generally given by [115]

$$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}(\mathbf{E}), \tag{2.9a}$$

$$\mathbf{B} = \mu_0\left[\mathbf{H} + \mathbf{M}(\mathbf{H})\right], \tag{2.9b}$$

$$\mathbf{J} = \mathbf{J}(\mathbf{E}), \tag{2.9c}$$

where $\mathbf{P}$ is the polarization of the medium and $\mathbf{M}$ is the magnetization of the medium. In general, the vector fields $\mathbf{P}$ and $\mathbf{M}$ depend on the electric field $\mathbf{E}$ and the magnetic field $\mathbf{H}$, and the dependency can be nonlinear. When the relations are linear, it makes sense to define them in the frequency domain, but then the relations depend on the frequency in case of dispersive media. The constants ε_0 and μ_0 are the permittivity and the permeability of free space, respectively. The values of ε_0 and μ_0 are given by

$$\mu_0 = 1.25663706212(19)\cdot 10^{-6}\,H/m \tag{2.10}$$

$$\varepsilon_0 = \frac{1}{c_0^2\mu_0} \approx 8.8541878128(13)\cdot 10^{-12}\,F/m, \tag{2.11}$$

where c_0 is the speed of the light in vacuum.

For most materials under weak-field conditions, the linear material responses of Eq. (2.9) are used in electromagnetics:

$$\mathbf{D} = \varepsilon\mathbf{E} = \varepsilon_0\varepsilon_r\mathbf{E}, \tag{2.12a}$$

$$\mathbf{B} = \mu\mathbf{H} = \mu_0\mu_r\mathbf{H}, \tag{2.12b}$$

$$\mathbf{J} = \sigma\mathbf{E}, \tag{2.12c}$$

where ε, μ, and σ are the permittivity, the permeability, and the electric conductivity of the medium, and they can depend on both on the position $\mathbf{x}$ and the angular frequency ω. The ε_r is called the relative permittivity, and μ_r is called the relative permeability, of the medium. In case of an anisotropic material, ε, μ, and σ are tensors that can be represented by 3×3 matrix functions [114]. In case of isotropic material, ε, μ, and σ become scalar functions. Furthermore, in the case of a single homogeneous isotropic material, ε, μ, σ reduce to constants. In free space, the constitutive relations in Eq. (2.12) become

$$\mathbf{D} = \varepsilon_0 \mathbf{E}, \tag{2.13a}$$
$$\mathbf{B} = \mu_0 \mathbf{H}, \tag{2.13b}$$
$$\mathbf{J} = \mathbf{0}. \tag{2.13c}$$

Concerning boundary conditions, the electromagnetic field quantities must satisfy radiation conditions at spatial infinity and interface conditions between different media where a material parameter shows discontinuity. The radiation conditions are often called Sommerfeld's radiation conditions [114], if they are stated based on electromagnetic potentials, or Silver-Müller radiation conditions [116–118], if they are stated based on the electromagnetic fields. Regarding the boundary conditions at an interface, let ∂D be a smooth interface between domain 1 and domain 2, and $\mathbf{n}$ be a normal vector pointing from domain 1 to domain 2. Then the following boundary conditions hold for the electromagnetic field quantities when crossing the interface between two materials:

$$\mathbf{n} \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_s, \tag{2.14a}$$
$$\mathbf{n} \times (\mathbf{E}_1 - \mathbf{E}_2) = \mathbf{0}, \tag{2.14b}$$
$$\mathbf{n} \cdot (\mathbf{B}_1 - \mathbf{B}_2) = 0, \tag{2.14c}$$
$$\mathbf{n} \cdot (\mathbf{D}_1 - \mathbf{D}_2) = \rho_s, \tag{2.14d}$$

where the subscripts 1 and 2 indicate the field in the corresponding domain, and $\mathbf{J}_s$ and ρ_s denote the electric surface current and surface charge densities, respectively, that are supported on the interface only. The following Figure 2.1 shows two domains with different material parameters and the corresponding normal vector $\mathbf{n}$ defined on the interface.

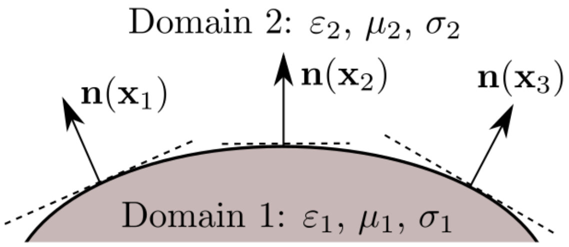

Domain 2: ε_2, μ_2, σ_2
$\mathbf{n}(\mathbf{x}_1)$
$\mathbf{n}(\mathbf{x}_2)$
$\mathbf{n}(\mathbf{x}_3)$
Domain 1: ε_1, μ_1, σ_1

Figure 2.1: An example of two media separated by a curved interface.

2.2 Domain integral equation for a layered medium

Based on Maxwell's equations that we stated in Section 2.1, we now derive a domain integral equation for the electromagnetic scattering problem in a layered medium.

2.2.1 Electromagnetic scattering problem in a layered medium

We describe the geometry of the dielectric layered medium and state the electromagnetic scattering problem first. We set up a standard right-handed Cartesian coordinate system and make the z-direction point downward[1], as shown in Figure. 2.2. Consider a dielectric layered medium with L layers, and the layered medium is enclosed between two half-spaces. All layers are stacked along the z direction and separated by the interfaces at $z = z_0, z_1, \ldots, z_{L-1}, z_L$, with $z_{l-1} > z_l$ for $l = 1, \ldots, L$. Let $V_l = \mathbb{R} \times \mathbb{R} \times (z_{l-1}, z_l)$ be the space of the l-th layer for $1 \leq l \leq L$, $V_0 = \mathbb{R}\times\mathbb{R}\times(-\infty, z_0)$, $V_{L+1} = \mathbb{R}\times\mathbb{R}\times(z_L, \infty)$ be the two half-space, then clearly the closure of $V_0 \bigcup V_1 \ldots V_L \bigcup V_{L+1}$ is equal to $\mathbb{R}^3$. We assume that the background medium for each layer is homogeneous, isotropic, and nonmagnetic, i.e. $\mu_r = 1$ throughout. Denote the relative permittivity of the background medium in layer l by $\varepsilon_{rb,l}$, then we can define the following background relative permittivity function as

$$\varepsilon_{rb}(\mathbf{x}) = \varepsilon_{rb}(z) = \begin{cases} \varepsilon_{rb,0}, & \text{if } z \in (-\infty, z_0), \\ \varepsilon_{rb,l}, & \text{if } z \in (z_{l-1}, z_l), \quad l = 1, \ldots, L, \, , \\ \varepsilon_{rb,L+1}, & \text{if } z \in (z_l, \infty). \end{cases} \tag{2.15}$$

Note that $\varepsilon_{rb,n}$ is a real constant for a lossless medium and becomes a complex constant, with a negative imaginary part, for a lossy material. In Figure 2.2 we show an example of a stratified background medium with $L = 4$, with scattering objects, in yellow, embedded in layer 1.

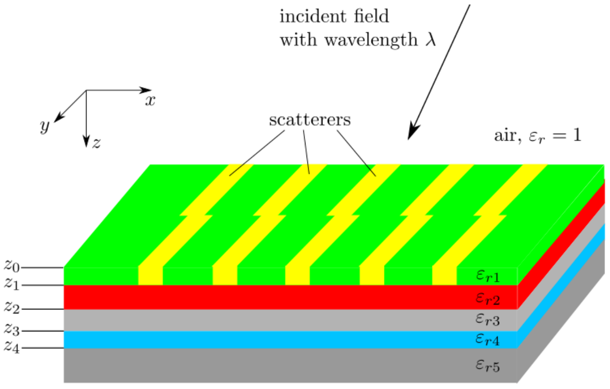


Figure 2.2: An example of a layered medium with dielectric scatterers embedded in the top layer.

[1] This is the coordinate system we use throughout this thesis unless stated otherwise.

Scattering objects, which are made of a different material than the surrounding background medium, are assumed to occupy a finite domain D in one of the background layers, i.e., $D \subset V_l \subset \mathbb{R}^3$ for some $1 \leq l \leq L$. The relative permittivity of the scatterers is denoted by $\varepsilon_{rs} \in \mathbb{C}$ and with this we can define a global relative permittivity function in the presence of scatterers in a layered medium:

$$\varepsilon_r(\mathbf{x}) = \begin{cases} \varepsilon_{rb}(z), & \text{if } \mathbf{x} \in \mathbb{R}^3 \backslash D, \\ \varepsilon_{rs}, & \text{if } \mathbf{x} \in D. \end{cases} \tag{2.16}$$

Clearly, $\varepsilon_r(\mathbf{x})$ distinguishes from $\varepsilon_{rb}(\mathbf{x})$ only within the scatterer domain D. To make the following formulation convenient, we define a contrast function

$$\chi(\mathbf{x}) = \frac{\varepsilon_r(\mathbf{x}) - \varepsilon_{rb}(z)}{\varepsilon_{rb}(z)} = \frac{\varepsilon_r(\mathbf{x})}{\varepsilon_{rb}(z)} - 1, \quad \mathbf{x} \in \mathbb{R}^3. \tag{2.17}$$

As a consequence, the contrast function $\chi(\mathbf{x})$ is only supported on the domain D occupied by the scatters.

The incident electric field $\mathbf{E}^i(\mathbf{x})$ plays an important role in electromagnetic scattering problems. The incident field is an electromagnetic field that satisfies Maxwell's equations in the absence of the scatterers, i.e. for the situation where the permittivity is that of the layered background medium everywhere. The incident electric field $\mathbf{E}^i(\mathbf{x})$ within the layered medium can be calculated as in [10]. Various applications with plane waves propagating along the z direction can be found in Chapter 6. In Section 7.3, we discuss a case with a plane wave propagating from the bottom and traveling along the $-z$ direction. In Section 7.2, we simulate a scattering problem with a beam-type incident field traveling along the x direction and within the layered medium. More cases with beam-type incident fields are analyzed in Section 7.1.

Now we can state the electromagnetic scattering problem due to scatterers embedded in the layered medium: given (1) the coordinate information, (2) the material properties of the layered background medium and scatterers, and (3) the information of the incident field, e.g. wavelength and incident angle, we should determine the total electric field, both in near-field and far-field regions, accurately and efficiently.

2.2.2 Transmission-line equations

The linearity of Maxwell's equations allows us to separate the electric and magnetic fields according to their sources. Let the incident field be the field originating from a source in the upper or lower half-space of the background medium (in the case of plane waves, we consider the sources are at infinity), and the scattered fields be the fields caused by the presence of the scattering objects from which the incident fields scatter, then we can write the total fields as the sum of the incident fields and the scattered fields, i.e.

$$\mathbf{E} = \mathbf{E}^i + \mathbf{E}^s, \tag{2.18a}$$

$$\mathbf{H} = \mathbf{H}^i + \mathbf{H}^s, \tag{2.18b}$$

where we use the superscript ^{i} to represent the incident fields and the superscript ^{s} to represent the scattered fields. We also call $\mathbf{E}$ and $\mathbf{H}$ the total electric field and total magnetic field, respectively.

Now we derive a domain integral equation for the electromagnetic scattering problem in a layered medium. From Eq. (2.7), (2.8) and (2.12), the total electric field satisfies

$$\nabla \times \mathbf{H}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) + j\omega\varepsilon_0\varepsilon_r(\mathbf{x})\mathbf{E}(\mathbf{x}), \tag{2.19a}$$

$$\nabla \times \mathbf{E}(\mathbf{x}) = -j\omega\mu_0\mathbf{H}(\mathbf{x}), \tag{2.19b}$$

where we assume $\mu_r \equiv 1$, i.e., the layered medium and the scatterers are nonmagnetic.

We now consider two cases of the layered medium. First, we consider the case with just a layered medium and the absence of dielectric scatterers, clearly $\mathbf{J}(\mathbf{x}) \equiv 0$ and $\varepsilon_r(\mathbf{x}) = \varepsilon_{rb}(z)$ for all $\mathbf{x} \in \mathbb{R}^3$. Given the incident electric field $\mathbf{E}^i(\mathbf{x})$, then Eq. (2.19) becomes

$$\nabla \times \mathbf{H}^i(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb}(z)\mathbf{E}^i(\mathbf{x}), \tag{2.20a}$$

$$\nabla \times \mathbf{E}^i(\mathbf{x}) = -j\omega\mu_0\mathbf{H}^i(\mathbf{x}). \tag{2.20b}$$

Second, we consider a layered medium with dielectric scatterers embedded in layer ℓ. In this case, the total electric field $\mathbf{E}(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^3$, satisfies

$$\nabla \times \mathbf{H}(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb}(z)\mathbf{E}(\mathbf{x}) + j\omega\varepsilon_0\varepsilon_{rb}(z)\chi(\mathbf{x})\mathbf{E}(\mathbf{x}), \tag{2.21a}$$

$$\nabla \times \mathbf{E}(\mathbf{x}) = -j\omega\mu_0\mathbf{H}(\mathbf{x}). \tag{2.21b}$$

By substituting Eq. (2.20) in Eq. (2.21) and employing Eq. (2.18), we arrive at the differential equations that govern the scattered electric field $\mathbf{E}^s$:

$$\nabla \times \mathbf{H}^s(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb}(z)\mathbf{E}^s(\mathbf{x}) + \mathbf{J}^c(\mathbf{x}), \tag{2.22a}$$

$$\nabla \times \mathbf{E}^s(\mathbf{x}) = -j\omega\mu_0\mathbf{H}^s(\mathbf{x}), \tag{2.22b}$$

where

$$\mathbf{J}^c(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb,\ell}\chi(\mathbf{x})\mathbf{E}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^3. \tag{2.23}$$

Note that $\mathbf{J}^c(\mathbf{x})$ in Eq. (2.23) has the same units as the electric current density $\mathbf{J}(\mathbf{x})$, and it vanishes outside D, owing to the contrast function $\chi(\mathbf{x})$. Therefore, we call $\mathbf{J}^c(\mathbf{x})$ the contrast current density. Since there are no other current density sources, we drop the superscript ^{c} from now on, to ease the notation. Eq. (2.23) is known as the field-material interaction, which is essentially a constitutive relation between the scattered field and the contrast current density. Eq. (2.22) can be rewritten in the following form:

$$\nabla \times \nabla \times \mathbf{E}^s(\mathbf{x}) - k_0^2\varepsilon_{rb}(z)\mathbf{E}^s(\mathbf{x}) = -j\omega\mu_0\mathbf{J}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^3. \tag{2.24}$$

A key step to solve the above system with a general inhomogeneous term $\mathbf{J}(\mathbf{x})$ is to find a specific solution of Eq. (2.24) with the inhomogeneous term being the Dirac delta function. This solution is called the Green function. In free space, the Green function has a well-known analytical expression. For a layered medium, the Green function is only available in the spectral domain via an analytical expression [119]. Even though it is possible to obtain a spatial-domain Green function via Sommerfeld integrals or Fourier transformations [120, 121], the required calculations are usually tedious and sophisticated due to the Green function's singularities and oscillatory behavior in the spatial domain. Therefore, it is advantageous to use the Green function in the spectral domain directly.

We introduce a pair of 2D Fourier transformations on the transverse plane denoted by $\mathcal{F}_T[\cdot]$ and its inverse $\mathcal{F}_T^{-1}[\cdot]$:

$$\mathcal{F}_T[f(\mathbf{x}_T)] \equiv \hat{f}(\mathbf{k}_T) = \int_{\mathbb{R}^2} f(\mathbf{x}_T) e^{j\mathbf{k}_T \cdot \mathbf{x}_T} d\mathbf{x}_T, \tag{2.25a}$$

$$\mathcal{F}_T^{-1}[\hat{f}(\mathbf{k}_T)] \equiv f(\mathbf{x}_T) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \hat{f}(\mathbf{k}_T) e^{-j\mathbf{x}_T \cdot \mathbf{k}_T} d\mathbf{k}_T. \tag{2.25b}$$

where we use $\mathbf{x}_T = (x, y)$ and $\mathbf{k}_T = (k_x, k_y)$ to represent the transverse vectors. Subsequently, the scattered electric field $\mathbf{E}^s(x, y, z)$, scattered magnetic field $\mathbf{H}^s(x, y, z)$ and contrast current density $\mathbf{J}(x, y, z)$ can be represented by the following spectral representation:

$$\mathbf{E}^s(\mathbf{x}_T, z) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \mathbf{e}^s(\mathbf{k}_T, z) \exp\{-j\mathbf{x}_T \cdot \mathbf{k}_T\} d\mathbf{k}_T, \tag{2.26}$$

$$\mathbf{H}^s(\mathbf{x}_T, z) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \mathbf{h}^s(\mathbf{k}_T, z) \exp\{-j\mathbf{x}_T \cdot \mathbf{k}_T\} d\mathbf{k}_T, \tag{2.27}$$

$$\mathbf{J}(\mathbf{x}_T, z) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \mathbf{j}(\mathbf{k}_T, z) \exp\{-j\mathbf{x}_T \cdot \mathbf{k}_T\} d\mathbf{k}_T, \tag{2.28}$$

where $\mathbf{k}_T$ is the transverse part of the wave vector of the incident field $\mathbf{k}$, and the integral is performed per Cartesian component. The spectral fields $\mathbf{e}^s(\mathbf{k}_T, z)$, $\mathbf{h}^s(\mathbf{k}_T, z)$, $\mathbf{j}(\mathbf{k}_T, z)$ are essentially Fourier coefficients according to Eq. (2.25a), e.g., $\mathbf{e}^s(\mathbf{k}_T, z) = \mathcal{F}_T[\mathbf{E}^s(\mathbf{x}_T, z)](\mathbf{k}_T, z)$.

Substituting the electric field $\mathbf{E}^s(\mathbf{x})$ and the magnetic field $\mathbf{H}^s(\mathbf{x})$ in Eq. (2.22b) in their spectral representations in Eq. (2.26) and (2.27), we have

$$\begin{aligned}
\nabla \times \mathbf{E}^s(\mathbf{x}) =& \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \nabla \times \Big\{ \mathbf{e}^s(\mathbf{k}_T, z) \exp\{-j\mathbf{x}_T \cdot \mathbf{k}_T\} \Big\} d\mathbf{k}_T \\
=& \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \Big\{ \exp\{-j\mathbf{x}_T \cdot \mathbf{k}_T\} \nabla \times \mathbf{e}^s(\mathbf{k}_T, z) \\
& \qquad + \Big(\nabla \exp\{-j\mathbf{x}_T \cdot \mathbf{k}_T\} \Big) \times \mathbf{e}^s(\mathbf{k}_T, z) \Big\} d\mathbf{k}_T \\
=& - \frac{j\omega\mu_0}{(2\pi)^2} \int_{\mathbb{R}^2} \mathbf{h}^s(\mathbf{k}_T, z) \exp\{-j\mathbf{x}_T \cdot \mathbf{k}_T\} d\mathbf{k}_T \\
=& - j\omega\mu_0 \mathbf{H}^s(\mathbf{x}),
\end{aligned} \tag{2.29}$$

therefore the spectral-domain fields $\mathbf{e}^s(\mathbf{k}_T, z)$ and $\mathbf{h}^s(\mathbf{k}_T, z)$ satisfy

$$-j\mathbf{k}_T \times \mathbf{e}^s(\mathbf{k}_T, z) + \mathbf{u}_z \times \frac{\partial}{\partial z} \mathbf{e}^s(\mathbf{k}_T, z) = -j\omega\mu_0 \mathbf{h}^s(\mathbf{k}_T, z), \tag{2.30}$$

where the differential operator $\frac{\partial}{\partial z}$ is performed per component.

Analogously, Eq. (2.22a) is reduced to:

$$-j\mathbf{k}_T \times \mathbf{h}^s(\mathbf{k}_T, z) + \mathbf{u}_z \times \frac{\partial}{\partial z}\mathbf{h}^s(\mathbf{k}_T, z) = \mathbf{j}(\mathbf{k}_T, z) + j\omega\varepsilon_0\varepsilon_{rb}(z)\mathbf{e}^s(\mathbf{k}_T, z). \tag{2.31}$$

To ease the subsequent derivation, we follow the analysis in [23, 119, 122] and introduce a rotated coordinate system in the spectral domain. Let $\mathbf{u}_k = \mathbf{k}_T/k_T$, where $k_T = \|\mathbf{k}_T\|$, then the vectors $\mathbf{u}_k$, $\mathbf{u}_z \times \mathbf{u}_k$ and $\mathbf{u}_z$ form another basis in the spectral domain $\mathbb{R}^3$.

Since each layer of the background medium is isotropic, the spectral representations $\mathbf{e}^s(\mathbf{k}_T, z)$, $\mathbf{h}^s(\mathbf{k}_T, z)$ and $\mathbf{j}(\mathbf{k}_T, z)$ can be decomposed into their transverse parts and longitudinal parts, based on the polarization states:

$$\mathbf{e}^s(\mathbf{k}_T, z) = e_{\|}(\mathbf{k}_T, z)\mathbf{u}_k + e_{\perp}(\mathbf{k}_T, z)(\mathbf{u}_z \times \mathbf{u}_k) + e_z(\mathbf{k}_T, z)\mathbf{u}_z, \tag{2.32}$$

$$\mathbf{h}^s(\mathbf{k}_T, z) = h_{\perp}(\mathbf{k}_T, z)\mathbf{u}_k + h_{\|}(\mathbf{k}_T, z)(\mathbf{u}_z \times \mathbf{u}_k) + h_z(\mathbf{k}_T, z)\mathbf{u}_z, \tag{2.33}$$

$$\mathbf{j}(\mathbf{k}_T, z) = j_{\|}(\mathbf{k}_T, z)\mathbf{u}_k + j_{\perp}(\mathbf{k}_T, z)(\mathbf{u}_z \times \mathbf{u}_k) + j_z(\mathbf{k}_T, z)\mathbf{u}_z, \tag{2.34}$$

where we employ the same notations used in [23]: the subscripts $_{\|}$, $_{\perp}$ are used to denote the parallel polarization and perpendicular polarization, respectively (or equivalently the e-polarization and h-polarization in [122]). Note that the spectral-domain magnetic field $\mathbf{h}^s$ along the $\mathbf{u}_k$ direction is $h_{\perp}$ instead of $h_{\|}$. The subscript $_z$ denotes the coefficient along the longitudinal direction $\mathbf{u}_z$.

By substituting Eqs. (2.32)-(2.34) in Eqs. (2.30) and (2.31), we can derive the following two sets of coupled ordinary differential equations (ODEs):

$$\frac{d}{dz}e_{\|}(\mathbf{k}_T, z) = \frac{j\gamma(\mathbf{k}_T, z)^2}{\omega\varepsilon_0\varepsilon_{rb}(z)}h_{\|}(\mathbf{k}_T, z) + \frac{k_T}{\omega\varepsilon_0\varepsilon_{rb}(z)}j_z(\mathbf{k}_T, z), \tag{2.35}$$

$$\frac{d}{dz}h_{\|}(\mathbf{k}_T, z) = -j\omega\varepsilon_0\varepsilon_{rb}(z)e_{\|}(\mathbf{k}_T, z) - j_{\|}(\mathbf{k}_T, z), \tag{2.36}$$

and

$$\frac{d}{dz}e_{\perp}(\mathbf{k}_T, z) = j\omega\mu_0 h_{\perp}(\mathbf{k}_T, z), \tag{2.37}$$

$$\frac{d}{dz}h_{\perp}(\mathbf{k}_T, z) = -\frac{j\gamma(\mathbf{k}_T, z)^2}{\omega\mu_0}e_{\perp}(\mathbf{k}_T, z) + j_{\perp}(\mathbf{k}_T, z), \tag{2.38}$$

where $\gamma(\mathbf{k}_T, z)^2 = k_T^2 - \omega^2\mu_0\varepsilon_0\varepsilon_{rb}(z)$, and $\gamma(\mathbf{k}_T, z)$ is determined according to the standard branch-cut definition, i.e., the condition $-\frac{\pi}{2} < \arg\{\gamma(\mathbf{k}_T, z)\} \leq \frac{\pi}{2}$ is satisfied. Eqs (2.35)-(2.36) and Eqs (2.37)-(2.38) are two first-order linear ordinary differential systems, and they are usually called transmission-line equations. Furthermore, the transmission-line equations are supplemented with the longitudinal components of the electric and magnetic fields:

$$e_z = \frac{-k_T h_{\|}(\mathbf{k}_T, z) + j[j_z(\mathbf{k}_T, z)]}{\omega\varepsilon_0\varepsilon_{rb}(z)}, \tag{2.39}$$

$$h_z = \frac{k_T e_{\perp}(\mathbf{k}_T, z)}{\omega\mu_0}. \tag{2.40}$$

Hence the original vector problem in Eq. (2.30) and (2.31) is reduced to the two scalar transmission-line problems in Eqs (2.35)-(2.38). In other words, for a given spectral contrast current density $\mathbf{j}(\mathbf{k}_T, z)$ in Eq. (2.34), we can get the solution $\mathbf{e}^s(\mathbf{k}_T, z)$ and $\mathbf{h}^s(\mathbf{k}_T, z)$ by solving the transmission-line equations in Eqs (2.35)-(2.38) together with Eqs (2.39) and (2.40).

2.2.3 Green function in a homogeneous medium

Following the analysis in [119, 122], now we define the dyadic Green function denoted by $G^h(\mathbf{k}_T, z, z') \in \mathbb{C}^{3\times3}$. Consider the following three contrast-current density functions

$$\mathbf{j}_p(\mathbf{k}_T, z, z') = \delta(z - z')\mathbf{u}_p, \tag{2.41}$$

where $\mathbf{u}_p \in \{\mathbf{u}_k, \mathbf{u}_z \times \mathbf{u}_k, \mathbf{u}_z\}$, and $\delta(z)$ is the Dirac delta function. Note that $\mathbf{j}_p$ represents an oriented impulsive source at $z = z'$ with unit amplitude along the $\mathbf{u}_p$ direction. We then define $G(\mathbf{k}_T, z, z') \cdot \mathbf{u}_p$ as the solution $\mathbf{e}^s(\mathbf{k}_T, z)$ of Eqs. (2.35)-(2.38), due to the specified source term $\mathbf{j}_p(\mathbf{k}_T, z, z')$. Therefore, for a general contrast current density $\mathbf{j}(\mathbf{k}_T, z)$ in a homogeneous medium, the solution of the system (2.30) and (2.31) can be written as

$$\mathbf{e}^s(\mathbf{k}_T, z) = \mathbf{e}(\mathbf{k}_T, z) - \mathbf{e}^i(\mathbf{k}_T, z) = \int_{\mathbb{R}} G^h(\mathbf{k}_T, z, z') \cdot \mathbf{j}(\mathbf{k}_T, z')dz'. \tag{2.42}$$

The dyadic Green function $G^h(\mathbf{k}_T, z, z')$ in a homogeneous medium with permittivity $\varepsilon_{rb,\ell}$ is given by [122]:

$$G^h(\mathbf{k}_T, z, z') = \left(-\mathbf{k}_T\mathbf{k}_T - j\mathbf{k}_T\mathbf{u}_z\frac{d}{dz} - j\mathbf{u}_z\mathbf{k}_T\frac{d}{dz} + \mathbf{u}_z\mathbf{u}_z\frac{d^2}{dz^2}\right)\frac{\exp(-\gamma|z - z'|)}{2j\gamma\omega\varepsilon_0\varepsilon_{rb,\ell}}, \tag{2.43}$$

where the factor $\exp(-\gamma|z - z'|)$ is used to propagate the field over a distance $|z - z'|$, i.e. from the z coordinate of the source (z') to the z coordinate of the observation point (z). Note that the Green function given in Eq. (2.43) also satisfies the radiation conditions when $|z| \to \infty$. Furthermore, the full expansion of Eq. (2.43) yields 9 scalar terms in the form of a 3×3 matrix function, and the full expression of the 9 scalar terms of $G^h(\mathbf{k}_T, z, z')$ can be found in [24] or [119].

We conclude this section by giving the solution of Eq. (2.24) for a homogeneous background medium:

$$\begin{aligned}\mathbf{E}^s(\mathbf{x}) =&\mathbf{E}(\mathbf{x}) - \mathbf{E}^i(\mathbf{x}) \\ =&\mathcal{F}_T^{-1}\left\{\int_{\mathbb{R}} G^h(\mathbf{k}_T, z, z') \cdot \mathcal{F}_T\{\mathbf{J}(\mathbf{x}_T, z')\}dz'\right\}.\end{aligned} \tag{2.44}$$

2.2.4 Green function in a layered medium

We now consider the Green function in a layered medium and derive the corresponding domain integral equation. In each layer of a layered medium, the overall scattered electric field contains not only the contribution from the homogeneous medium of the pertaining layer (which has been studied in the previous subsection), but also the reflections from the material interfaces above and below that layer. The complete spectral scattered electric field $\mathbf{e}^s(\mathbf{k}_T, z)$ in a particular layer of the layered medium can be decomposed as

$$\mathbf{e}^s(\mathbf{k}_T, z) = \mathbf{e}^{s,h}(\mathbf{k}_T, z) + \mathbf{e}^{s,u}(\mathbf{k}_T, z) + \mathbf{e}^{s,d}(\mathbf{k}_T, z), \tag{2.45}$$

where $\mathbf{e}^{s,h}(\mathbf{k}_T, z)$ represents the homogeneous scattered electric field and can be written as

$$\mathbf{e}^{s,h}(\mathbf{k}_T, z) = \int_{\mathbb{R}} G^h(\mathbf{k}_T, z, z') \cdot \mathbf{j}(\mathbf{k}_T, z') dz', \tag{2.46}$$

where $G^h(\mathbf{k}_T, z, z')$ is a 3×3 matrix function pertaining to the Green function in a homogeneous medium equal to that of the layer where both z and z' are confined to. Furthermore, $\mathbf{e}^{s,u}(\mathbf{k}_T, z)$ in Eq. (2.45) represents the reflected electric field propagating upward from the bottom interface, and $\mathbf{e}^{s,d}(\mathbf{k}_T, z)$ in Eq. (2.45) represents the reflected electric field propagating downward from the upper interface[2].

The reflections $\mathbf{e}^{s,u}(\mathbf{k}_T, z)$ and $\mathbf{e}^{s,d}(\mathbf{k}_T, z)$ rely on the propagation of homogeneous scattered electric field $\mathbf{e}^{s,h}(\mathbf{k}_T, z)$ and corresponding effective reflection coefficients [110], denoted by $R^{u,u}$, $R^{u,d}$, $R^{d,u}$, $R^{d,d}$. These effective reflection coefficients can be computed based on the effective coefficients for e-polarization [108] and h-polarization [109]. More details about effective reflection coefficients of layer interfaces are given in [115]. If we assume all scatterers are bounded to the interval $[z_a, z_b]$ in the longitudinal direction, then the reflected fields $\mathbf{e}^{s,u}$ and $\mathbf{e}^{s,d}$ can be written as

$$\begin{aligned}
\mathbf{e}^{s,d}(\mathbf{k}_T, z) =& R^{d,u}(\mathbf{k}_T) \cdot \mathbf{e}^{s,h}(\mathbf{k}_T, z_a) e^{-\gamma(z_b - z)} + R^{d,d}(\mathbf{k}_T) \cdot \mathbf{e}^{s,h}(\mathbf{k}_T, z_b) e^{-\gamma(z_b - z)} \\
=& \int_{\mathbb{R}} \Big\{ R^{d,u}(\mathbf{k}_T) G^h(\mathbf{k}_T, z_a, z') e^{-\gamma(z_b - z)} \\
& \quad + R^{d,d}(\mathbf{k}_T) G^h(\mathbf{k}_T, z_b, z') e^{-\gamma(z_b - z)} \Big\} \cdot \mathbf{j}(\mathbf{k}_T, z') dz' \\
\triangleq & \int_{\mathbb{R}} G^d(\mathbf{k}_T, z, z') \cdot \mathbf{j}(\mathbf{k}_T, z') dz',
\end{aligned} \tag{2.47}$$

and

$$\begin{aligned}
\mathbf{e}^{s,u}(\mathbf{k}_T, z) =& R^{u,u}(\mathbf{k}_T) \cdot \mathbf{e}^{s,h}(\mathbf{k}_T, z_a) e^{-\gamma(z - z_a)} + R^{u,d}(\mathbf{k}_T) \cdot \mathbf{e}^{s,h}(\mathbf{k}_T, z_b) e^{-\gamma(z - z_a)} \\
=& \int_{\mathbb{R}} \Big\{ R^{u,u}(\mathbf{k}_T) G^h(\mathbf{k}_T, z_a, z') e^{-\gamma(z - z_a)} \\
& \quad + R^{u,d}(\mathbf{k}_T) G^h(\mathbf{k}_T, z_b, z') e^{-\gamma(z - z_a)} \Big\} \cdot \mathbf{j}(\mathbf{k}_T, z') dz' \\
\triangleq & \int_{\mathbb{R}} G^u(\mathbf{k}_T, z, z') \cdot \mathbf{j}(\mathbf{k}_T, z') dz',
\end{aligned} \tag{2.48}$$

[2] The bottom interface means the interface closer to the bottom of Fig. 2.2 and the upper interface means the interface closer to the top of Fig. 2.2, with respect to the orientation on the page and the reading direction. Similarly, upward means the direction toward negative z, and downward means the direction toward positive z.

where we introduce two 3 × 3 matrix functions $G^u(\mathbf{k}_T, z, z')$ and $G^d(\mathbf{k}_T, z, z')$ to ease following formulation.

After combining Eqs (2.46), (2.47) and (2.48), we can rewrite Eq. (2.45) as

$$\begin{aligned} \mathbf{e}^s(\mathbf{k}_T, z) &= \int_{\mathbb{R}} \left\{ G^h(\mathbf{k}_T, z, z') + G^u(\mathbf{k}_T, z, z') + G^d(\mathbf{k}_T, z, z') \right\} \cdot \mathbf{j}(\mathbf{k}_T, z') dz' \\ &\triangleq \int_{\mathbb{R}} G(\mathbf{k}_T, z, z') \cdot \mathbf{j}(\mathbf{k}_T, z') dz', \end{aligned} \tag{2.49}$$

where the Green function $G(\mathbf{k}_T, z, z')$ for a layered medium is defined in the last step.

The Green function $G(\mathbf{k}_T, z, z')$ is a superposition of the matrix functions G^h, G^u and G^d, and it computes the complete scattered electric field within a layered medium from the contrast current source $\mathbf{j}(\mathbf{k}_T, z)$. Note that the discretization of the Green function $G(\mathbf{k}_T, z, z')$ along the longitudinal direction can be done in the spatial domain completely. In practice, the convolution regarding z and z' in Eq. (2.49) can be computed efficiently in a recursive manner. Furthermore, the multiplication of the Green function and the contrast current density in the spectral domain can be done pointwise [111].

We conclude this section by giving the solution of Eq. (2.24) for a layered background medium:

$$\begin{aligned} \mathbf{E}^s(\mathbf{x}) =& \mathbf{E}(\mathbf{x}) - \mathbf{E}^i(\mathbf{x}) \\ =& \mathcal{F}_T^{-1} \left\{ \int_{\mathbb{R}} G(\mathbf{k}_T, z, z') \cdot \mathcal{F}_T \big\{ \mathbf{J}(\mathbf{x}_T, z') \big\} dz' \right\}. \end{aligned} \tag{2.50}$$

Eq. (2.50) is the domain integral equation in a layered medium.

2.3 The spatial spectral method

We arrive at a position to recall the integral equation in the spatial spectral method developed in [24]. The overall domain integral equation for the scattering problem in a layered medium can be written as

$$\mathbf{E}^i(\mathbf{x}) = \mathbf{E}(\mathbf{x}) - \mathcal{F}_T^{-1} \left\{ \int_{\mathbb{R}} G(\mathbf{k}_T, z, z') \cdot \mathcal{F}_T \big\{ \mathbf{J}(\mathbf{x}_T, z') \big\} dz' \right\} \tag{2.51}$$

$$\mathbf{J}(\mathbf{x}) = j\omega\varepsilon_0\varepsilon_{rb}(z)\chi(\mathbf{x})\mathbf{E}(\mathbf{x}), \tag{2.52}$$

where the Green operator G is defined in (2.49), $\mathcal{F}_T$ and $\mathcal{F}_T^{-1}$ form a pair of 2D Fourier transformations with respect to the definition in (2.25a) and (2.25b), the background relative permittivity function $\varepsilon_{rb}(z)$ and the contrast function $\chi(\mathbf{x})$ are given in (2.15) and (2.17), respectively.

There are several aspects that make the spatial spectral method unique. First of all, the total electric field $\mathbf{E}$ and the contrast current $\mathbf{J}$ are represented as linear transformations on an auxiliary field $\mathbf{F}$. Based on a local normal vector formulation, $\mathbf{F}(\mathbf{x})$ is continuous almost everywhere and it is constructed to improve the accuracy not only near the field-material interface but on the whole computational domain. We will discuss the normal vector field formulation and the auxiliary field **F** in Section 2.3.1.

Secondly, a Gabor frame is used as the discretization in the transverse plane, which results in an efficient transformation between the spatial domain and the spectral domain. Furthermore, a set of basis functions associated with Gabor frames is developed for 3D problems, which yields a fast multiplication operation and an FFT-based Fourier transformation [111]. The Gabor-frame-based discretization of the spatial spectral method yields an $O(N \log N)$ computational efficiency with N being the number of unknowns, see various applications in [108–110]. In Section 2.3.2 we recall some fundamentals of Gabor analysis.

Thirdly, a complex integration path in the spectral domain is chosen carefully to avoid poles and branch cuts in the spectral complex plane of the Green operator. Here we do not recall the sophisticated discretization of the Green operator but refer interested readers to [109, 110].

2.3.1 Normal vector field formulation

For spectral expansion methods [97, 98, 123] pertaining to scattering by periodic structures, a Fourier series expansion is used to represent the electric field, electric flux density, and current density. A spatial product of two functions (e.g., in the field-material interaction Eq. (2.52)) is commonly required to be performed in the spectral domain. In other words, we need to be able to compute the Fourier coefficients of the product function from the Fourier coefficients of the given functions that form the product. The framework to compute these coefficients is referred to as the Fourier factorization rules. When the given functions do not have concurrent discontinuities, Laurent's rule states that the Fourier coefficients of the product function are well approximated by a discrete convolution. However, when the two functions have concurrent discontinuities (e.g., both $\chi(\mathbf{x})$ and **E** in Eq. (2.52) possess a discontinuity when crossing the interfaces of the scatterer), Laurent's rule leads to poor convergence of the resulting Fourier expansion of the product function [124, 125].

A remedy for this poor convergence is the inverse rule of factorization, which was first empirically discovered by Granet, and independently by Lalanne and Morris for metallic grating problems in TM polarization for scattering by periodic objects [125, 126]. Later on, Li systematically developed and formulated the factorization rules for the products of discontinuous periodic functions in a more general case [106] and those rules are known as the Li factorization rules today, which serve as a framework to compute the Fourier coefficients of **E** and **J** when a discontinuous material property is being multiplied.

From the Fourier factorization rules given above, the normal vector field is then introduced by assembling the continuous components of **E** and **D** into an auxiliary field **F** [105], such that the auxiliary field **F** is continuous when crossing the interface in the transverse plane. This formulation has been modified and improved for several computational frameworks [103, 107, 123, 127]. Following the definitions of projection operators in [107], for any

$\mathbf{v} \in \mathbb{C}^3$ there are projection operators $\mathcal{P}_n$ and $\mathcal{P}_T$ constructed as follows

$$\mathcal{P}_n \mathbf{v} = (\mathbf{N} \cdot \mathbf{v})\mathbf{N}, \tag{2.53}$$

$$\mathcal{P}_T \mathbf{v} = (\mathbf{T}_1 \cdot \mathbf{v})\mathbf{T}_1 + (\mathbf{T}_2 \cdot \mathbf{v})\mathbf{T}_2, \tag{2.54}$$

where the normal vector field $\mathbf{N}(\mathbf{x})$ is normal to every material interface, and the two tangential vector fields $\mathbf{T}_1(\mathbf{x})$, $\mathbf{T}_2(\mathbf{x})$ are tangential to every material interface [105], almost everywhere in the computational domain. All vector fields $\mathbf{N}, \mathbf{T}_1, \mathbf{T}_2$ are defined on the whole computational domain according to the geometry of the configuration. Furthermore, $\mathcal{P}_n + \mathcal{P}_T = \mathcal{I}$, where $\mathcal{I}$ is the identity operator in $\mathbb{C}^3$, and $\mathcal{P}_T\mathcal{P}_n = \mathcal{P}_n\mathcal{P}_T = 0$. Note that $\mathcal{P}_T$ can be rewritten as $\mathcal{I} - \mathcal{P}_n$, which avoids the need to construct the vector fields $\mathbf{T}_1(\mathbf{x})$ and $\mathbf{T}_2(\mathbf{x})$.

Now an auxiliary field $\mathbf{F}$ is constructed based on the projection operators defined in (2.53),

$$\mathbf{F} = \mathcal{P}_T \cdot \mathbf{E} + \frac{1}{\varepsilon_0 \varepsilon_{rb}(\mathbf{x})} \mathcal{P}_n \cdot \mathbf{D}, \tag{2.55}$$

where $1/(\varepsilon_0\varepsilon_{rb})$ is one of the possible scaling factors given in [107]. The auxiliary field $\mathbf{F}$ is essentially a mix of the continuous parts of $\mathbf{E}$ and $\mathbf{D}$, therefore it is continuous throughout.

Following the analysis given in [24, 107], the electric field $\mathbf{E}$ and the contrast current density $\mathbf{J}$ can be represented by the auxiliary field $\mathbf{F}$ together with the operators $\mathcal{C}_\varepsilon$ and $\chi\mathcal{C}_\varepsilon$ through

$$\mathbf{E} = \mathcal{C}_\varepsilon \cdot \mathbf{F} = \left(\mathcal{P}_T + \frac{1}{\varepsilon_0 \varepsilon_{rb}(1+\chi)} \mathcal{P}_n \right) \cdot \mathbf{F}, \tag{2.56}$$

$$\mathbf{J} = \chi\mathcal{C}_\varepsilon \cdot \mathbf{F} = j\omega\varepsilon_0\varepsilon_{rb} \left(\chi\mathcal{P}_T + \frac{\chi}{\varepsilon_0 \varepsilon_{rb}(1+\chi)} \mathcal{P}_n \right) \cdot \mathbf{F}, \tag{2.57}$$

where the operators $\mathcal{C}_\varepsilon$ and $\chi\mathcal{C}_\varepsilon$ are essentially 3×3 matrix functions.

By substituting $\mathbf{E}$ and $\mathbf{J}$ in Eq. (2.51) and (2.52) in (2.56) and (2.57), we obtain the electric field integral equation (EFIE) of the spatial spectral method with the normal vector field formulation:

$$\begin{aligned} \mathbf{E}^i(\mathbf{x}) = & \mathcal{C}_\varepsilon(\mathbf{x}) \cdot \mathbf{F}(\mathbf{x}) \\ & - \mathcal{F}_T^{-1} \left\{ \int_{\mathbb{R}} G(\mathbf{k}_T, z, z') \cdot \mathcal{F}_T \big\{ \chi\mathcal{C}_\varepsilon(\mathbf{x}_T, z') \cdot \mathbf{F}(\mathbf{x}_T, z') \big\} dz' \right\}. \end{aligned} \tag{2.58}$$

Note that the auxiliary field $\mathbf{F}(\mathbf{x})$ is transformed to the contrast current density $\mathbf{J}(\mathbf{x})$ by the operator $\chi\mathcal{C}_\varepsilon$. Therefore, the integration interval along the longitudinal direction is reduced to $[z_a, z_b]$ as defined just above Eq. (2.47), where the scatterers are bounded in z direction.

2.3.2 Discretization based on Gabor frames

To make a model in an analytical form applicable in practice, a proper discretization scheme is usually required to take advantage of computers. Successful examples of such discretization schemes in computational electromagnetics have been discussed in Section 1.2. Here we continue on the discretization of the EFIE (2.58) in the spatial spectral method.

From Eq. (2.58), it is clear that only the spatial domain is involved along the z direction, while in the x and y directions all fields and the Green function need to be represented in both the spatial domain and the spectral domain. Piecewise-linear (PWL) functions have been shown to be effective along the z direction in many cases [123, 128], and they are also used as basis functions along the z direction in the spatial spectral method [108–110]. Here we do not go into the details of the PWL functions, but refer the interested readers to [24]. The discretization in the transverse directions in the spatial spectral method is based on Gabor frames. Following [129, 130], we introduce the general notations of Gabor analysis in $L^2(\mathbb{R}^d)$, i.e. the space of Lebesgue square-integrable functions on $\mathbb{R}^d$, see e.g. [131, Chapter 4], and briefly explain how Gabor frames are used to discretize the EFIE in Eq. (2.58).

Given $\mathbf{x}, \boldsymbol{\tau}, \boldsymbol{\xi} \in \mathbb{R}^d$, we define the translation operator $T_{\boldsymbol{\tau}}$ and the modulation operator $M_{\boldsymbol{\xi}}$ for a function $f(\mathbf{x})$ by

$$\mathcal{T}_{\boldsymbol{\tau}} f(\mathbf{x}) = f(\mathbf{x} - \boldsymbol{\tau}), \tag{2.59}$$

$$\mathcal{M}_{\boldsymbol{\xi}} f(\mathbf{x}) = e^{2\pi j \boldsymbol{\xi} \cdot \mathbf{x}} f(\mathbf{x}), \tag{2.60}$$

respectively. Translation and modulation operators play a key role in Gabor analysis. Let Λ be a separable lattice in $\mathbb{R}^{2d}$ that takes the form

$$\Lambda = \alpha \mathbb{Z}^d \times \beta \mathbb{Z}^d, \tag{2.61}$$

where $\alpha, \beta > 0$ are called lattice parameters. A Gabor system G_s of a window function $g \in L^2(\mathbb{R}^d) \backslash \{0\}$ and lattice parameters $\alpha, \beta > 0$ is defined as

$$G_s(g, \alpha, \beta) = \{\mathcal{M}_{\beta \mathbf{n}} \mathcal{T}_{\alpha \mathbf{m}} g\}, \quad \mathbf{m}, \mathbf{n} \in \mathbb{Z}^d. \tag{2.62}$$

Note that the integer-valued $\mathbf{m}$ and $\mathbf{n}$ represent the spatial shift index and the frequency modulation index, respectively. The Gabor system $G_s(g, \alpha, \beta)$ is called a Gabor frame or a Weyl-Heisenberg frame, if there exist positive constants A and B such that for all $f \in L^2(\mathbb{R}^d)$,

$$A\|f\|^2 \leq \sum_{\mathbf{m},\mathbf{n} \in \mathbb{Z}^d} |\langle f, \mathcal{M}_{\beta \mathbf{n}} \mathcal{T}_{\alpha \mathbf{m}} g \rangle_{L^2}|^2 \leq B\|f\|^2. \tag{2.63}$$

The associated Gabor frame operator $\mathcal{S}$ mapping from $L^2(\mathbb{R}^d)$ to $L^2(\mathbb{R}^d)$ is defined as

$$\mathcal{S} f = \sum_{\mathbf{m},\mathbf{n} \in \mathbb{Z}^d} \langle f, \mathcal{M}_{\beta \mathbf{n}} \mathcal{T}_{\alpha \mathbf{m}} g \rangle_{L^2} \mathcal{M}_{\beta \mathbf{n}} \mathcal{T}_{\alpha \mathbf{m}} g, \tag{2.64}$$

and $\mathcal{S}$ is bounded and invertible. When the window function g is chosen as a normalized Gaussian function, the Gabor system constitutes a frame iff $\alpha\beta < 1$, and this case is referred to as oversampling. In practice, the parameters α and β are often chosen equally and their product as the ratio of two integers, for computational reasons. If the oversampling parameters $\alpha = \beta = \sqrt{q/p}$, for some integers $q < p$, are too small, i.e, α, β are close to 0, then a very large number of Gabor coefficients is required, which is inefficient. If the oversampling parameters $\alpha = \beta$ are too large, i.e., α, β are close to 1, then a situation called critical sampling is approached, where $\alpha\beta = 1$. The Balian-Low Theorem [129] states that the Gabor system $G_s(g, \alpha, \beta)$ with critical sampling cannot even constitute a frame.

In the spatial spectral method, the following two-dimensional version of a normalized Gaussian function is used as the Gabor frame window function for the discretization in x and y directions

$$g(x, y) = 2^{\frac{1}{2}} \exp\left(-\pi\frac{x^2}{T_x^2} - \pi\frac{y^2}{T_y^2}\right), \tag{2.65}$$

where T_x and T_y denote the Gabor window lengths in the spatial domain. Further, we define $K_x = 2\pi/T_x$ and $K_y = 2\pi/T_y$ as the spectral-domain window lengths. The associated Gabor frame is defined based on the above translation and modulation operations, i.e.

$$g_{\mathbf{m},\mathbf{n}}(x, y) = g(x - \alpha_x m_x T_x, y - \alpha_y m_y T_y) \exp\left(j\beta_x n_x K_x x + j\beta_y n_y K_y y\right), \tag{2.66}$$

where the integer-valued $\mathbf{m} = (m_x, m_y)$ and $\mathbf{n} = (n_x, n_y)$ represent the spatial shift index and the frequency modulation index in the x and y directions, respectively.

One of the fundamental results in Gabor analysis, see e.g. [129, 132], states that for a given Gabor frame $G_s(g, \alpha, \beta)$ a dual window function $\eta \in L^2(\mathbb{R}^d)$ exists and an induced dual Gabor frame $G_s(\eta, \alpha, \beta)$. Furthermore, every $f(\mathbf{x}) \in L^2(\mathbb{R}^d)$ has the following Gabor-frame expansion

$$\begin{aligned} f(\mathbf{x}) &= \sum_{\mathbf{m},\mathbf{n}\in\mathbb{Z}^d} \langle f, \mathcal{M}_{\beta\mathbf{n}}\mathcal{T}_{\alpha\mathbf{m}}\eta(\mathbf{x})\rangle_{L^2} \mathcal{M}_{\beta\mathbf{n}}\mathcal{T}_{\alpha\mathbf{m}} g(\mathbf{x}) \\ &= \sum_{\mathbf{m},\mathbf{n}\in\mathbb{Z}^d} f_{\mathbf{m},\mathbf{n}} g_{\mathbf{m},\mathbf{n}}(\mathbf{x}), \end{aligned} \tag{2.67}$$

where $g_{\mathbf{m},\mathbf{n}}(\mathbf{x})$ is a Gabor frame function and $f_{\mathbf{m},\mathbf{n}}$ is a corresponding Gabor coefficient. Note that the convergence of (2.67) is in the norm $\|\cdot\|_2$ of $L^2(\mathbb{R}^d)$ and the convergence is unconditional [129, 132]. The Gabor coefficient $f_{\mathbf{m},\mathbf{n}}$ is computed based on the following Gabor transformation

$$f_{\mathbf{m},\mathbf{n}} = \langle f, \mathcal{M}_{\beta\mathbf{n}}\mathcal{T}_{\alpha\mathbf{m}}\eta(\mathbf{x})\rangle_{L^2} = \int_{\mathbb{R}^d} f(\mathbf{x})\eta^*_{\mathbf{m},\mathbf{n}}(\mathbf{x})d\mathbf{x}, \tag{2.68}$$

where $\eta_{\mathbf{m},\mathbf{n}}(\mathbf{x})$ is called the dual frame function. The dual window function $\eta(\mathbf{x})$ in Eq. (2.65) and (2.66) is not unique and all dual window functions $\eta(\mathbf{x})$ belong to an affine subspace of $L^2(\mathbb{R}^d)$, based on a canonical dual window function [129]. The canonical dual window function is defined as $\eta^\circ = \mathcal{S}^{-1}g$ and it yields a canonical dual frame $G_s(\mathcal{S}^{-1}g, \alpha, \beta)$. Several methods of approximating the inverse Gabor frame operator $\mathcal{S}^{-1}$, which is the key to computing the canonical dual window function, are reviewed in [133].

In the spatial spectral method, the dual window function $\eta(x, y)$ is computed based on the canonical dual window function that involves the generalized pseudo inverse of the frame operator, see [132] or [134], where the Zak transform is used as the main tool. The dual frame function is then defined based on the dual window function by

$$\eta_{\mathbf{m},\mathbf{n}}(x, y) = \eta(x - \alpha_x m_x T_x, y - \alpha_y m_y T_y) \exp\left(j\beta_x n_x K_x x + j\beta_y n_y K_y y\right). \tag{2.69}$$

The corresponding Gabor-frame expansion and Gabor transformation used in the spatial spectral method become

$$f(x, y) = \sum_{\mathbf{m},\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{m},\mathbf{n}} g_{\mathbf{m},\mathbf{n}}(x, y), \tag{2.70}$$

$$f_{\mathbf{m},\mathbf{n}} = \int_{\mathbb{R}^2} f(x, y)\eta^*_{\mathbf{m},\mathbf{n}}(x, y) dx dy. \tag{2.71}$$

Now we consider the Fourier transform of the spatial Gabor frame window function $g(x, y)$ in Eq. (2.65) according to Eq. (2.25a). We get

$$\hat{g}(k_x, k_y) = 2^{\frac{1}{2}} T_x T_y \exp\left(-\pi\frac{k_x^2}{K_x^2} - \pi\frac{k_y^2}{K_y^2}\right). \tag{2.72}$$

We then form the following Gabor system by translation and modulation operations and arrive at

$$\hat{g}_{\mathbf{m},\mathbf{n}}(k_x, k_y) = \hat{g}(k_x - \beta_x m_x K_x, k_y - \beta_y m_y K_y) \exp\left(-j\alpha_x n_x T_x k_x - j\alpha_y n_y T_y k_y\right). \tag{2.73}$$

One can readily verify that $\hat{g}_{\mathbf{m},\mathbf{n}}$ forms a Gabor frame as well, as long as the condition in Eq. (2.63) is fulfilled. We call Eq. (2.73) the Gabor frame in the spectral domain. Therefore, every $\hat{f}(k_x, k_y) \in L^2(\mathbb{R}^2)$ has the following Gabor frame expansion in the spectral domain

$$\hat{f}(k_x, k_y) = \sum_{\mathbf{m},\mathbf{n}\in\mathbb{Z}^2} \hat{f}_{\mathbf{m},\mathbf{n}} \hat{g}_{\mathbf{m},\mathbf{n}}(k_x, k_y), \tag{2.74}$$

$$\hat{f}_{\mathbf{m},\mathbf{n}} = \int_{\mathbb{R}^2} \hat{f}(k_x, k_y)\hat{\eta}^*_{\mathbf{m},\mathbf{n}}(k_x, k_y) dk_x dk_y, \tag{2.75}$$

where $\hat{\eta}_{\mathbf{m},\mathbf{n}}(k_x, k_y)$ is the spectral dual frame.

An important property of the Gabor-frame expansion is that, given a Gabor-frame expansion of a function $f(x, y)$ through Eq. (2.70), i.e. all Gabor coefficients $f_{\mathbf{m},\mathbf{n}}$ are known, we can easily obtain the spectral Gabor-frame expansion for $\hat{f}(k_x, k_y)$, through the following algebraic operations on the Gabor coefficients:

$$\hat{f}_{\mathbf{m},\mathbf{n}} = f_{\mathbf{n},\mathbf{m}} \exp\big\{-2\pi j(\alpha_x\beta_x m_x n_x + \alpha_y\beta_y m_y n_y)\big\}. \tag{2.76}$$

Eq. (2.76) yields significant advantages in practice, since it essentially establishes an efficient transformation between the spatial domain and the spectral domain.

The Gabor-frame expansion and Gabor transformation in Eq. (2.70) and (2.71) play a key role in the spatial spectral method [24]. When the expansion in the z direction is fixed to the PWL discretization, all constants, vector fields, and matrix functions in the EFIE (2.58) are represented in the transverse plane by Gabor frames according to Eq. (2.70). Their counterparts in the spectral domain are computed efficiently based on the relation of the spatial Gabor coefficients and spectral Gabor coefficients given in Eq. (2.76). Furthermore, a commonly used choice for the oversampling parameters is $\alpha_x = \alpha_y = \beta_x = \beta_y = \sqrt{2/3}$, since it has been shown to be efficient as a trade-off between oversampling and the width of the dual window function $\eta(x, y)$ in the spatial spectral electromagnetic solver [24].

2.4 Iterative methods and preconditioning

For practical implementation, the discrete Gabor system in Eq. (2.70) and (2.71) must be restricted to finite dimensions. Assume the translation index $\mathbf{m} = (m_x, m_y)$ and the modulation index $\mathbf{n} = (n_x, n_y)$ are restricted to

$$\begin{aligned} -M_x \le m_x \le M_x, &\quad -M_y \le m_y \le M_y, \\ -N_x \le n_x \le N_x, &\quad -N_y \le n_y \le N_y. \end{aligned}$$

for some $M_x, M_y, N_x, N_y \in \mathbb{N}^+$, and we also assume that the number of PWL functions used in z-direction discretization is $N_z \in \mathbb{N}^+$. Then the main EFIE with normal vector field formulation in Eq. (2.23) can be represented by the following linear system:

$$(C - G \cdot M) \cdot \mathbf{u} = \mathbf{f}, \tag{2.77}$$

where $C, M \in \mathbb{C}^{N \times N}$ are the matrix representations of the operators $\mathcal{C}_\varepsilon$ and $\chi \mathcal{C}_\varepsilon$ in Eq. (2.56) and (2.57), respectively. The matrix $G \in \mathbb{C}^{N \times N}$ is the matrix representation of the Green operator in combination with a pair of Fourier transformations in Eq. (2.58), and $\cdot$ is the standard operation for matrix-vector-product (MVP). The inhomogeneous term $\mathbf{f} \in \mathbb{C}^N$ represents the discretized incident field $\mathbf{E}^i$, and $\mathbf{u} \in \mathbb{C}^N$ contains the unknown Gabor coefficients of the auxiliary field $\mathbf{F}$. Note that N represents the number of unknowns and it can be calculated via the formula

$$N = 3 \cdot N_z \cdot (2M_x + 1) \cdot (2M_y + 1) \cdot (2N_x + 1) \cdot (2N_y + 1), \tag{2.78}$$

where 3 denotes the three vector-field components along the x, y, z directions.

There is a unique solution $\mathbf{u}$ of the linear system (2.77) if system matrix $A = C - GM$ is nonsingular, and

$$\mathbf{u} = A^{-1} \cdot \mathbf{f} = (C - GM)^{-1} \cdot \mathbf{f}. \tag{2.79}$$

The electric field $\mathbf{E}(\mathbf{x})$ can be reconstructed from the solution $\mathbf{u}$ or the auxiliary field $\mathbf{F}$ by applying the operator $\mathcal{C}_\varepsilon$. Theoretically, the solution of the linear system Eq. (2.77) can be computed by direct methods such as Gaussian elimination. However, for a large linear system (e.g., the total number of unknowns in the applications in Chapter 7 can reach a level of 10^8), storing and processing the system matrix is then computationally expensive, if not impossible. An iterative method is a widely used alternative strategy to direct methods to solve a large linear system, by taking an initial vector and generating a sequence of subsequent approximating vectors that approach the true solution. During the procedure, only $O(N)$ coefficients need to be stored instead of the full system matrix. Additionally, for the spatial spectral method each MVP can be performed with an operational complexity of $O(N \log N)$.

2.4.1 Iterative methods

Consider a general linear system $A\mathbf{u} = \mathbf{f}$. An iterative method starts from an initial guess $\mathbf{u}_0$ and computes a sequence of approximations $\mathbf{u}_1, \mathbf{u}_2, \ldots$ that are intended to approach the true solution $\mathbf{u}$, and terminates when the residual $\mathbf{r}_n = A \cdot \mathbf{u}_n - \mathbf{f}$ satisfies $\|\mathbf{r}_n\|/\|\mathbf{f}\| \leq \varepsilon_{\text{tol}}$, for some given relative error tolerance ε_{tol}. Among all iterative methods to solve large linear systems, Krylov subspace methods are used extensively [135]. A Krylov subspace method iteratively computes $\mathbf{u}_n$, such that $\mathbf{u}_n - \mathbf{u}_0$ belongs to the following Krylov subspace

$$\mathcal{K}_n(A, \mathbf{r}_0) = \text{span}(\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, \ldots, A^{n-1}\mathbf{r}_0), \tag{2.80}$$

for all $n \in \mathbb{N}^+$, and $\mathbf{r}_0 = \mathbf{f} - A \cdot \mathbf{u}_0$ is the initial residual.

When the system matrix A is real-symmetric and positive definite, a very popular Krylov subspace method is the conjugate gradient (CG) method [74]. By using a short recurrence, the CG method minimizes the matrix norm of the residual error over the Krylov subspace. The CG method is extremely simple and efficient, since usually only a small number of vectors is required to be stored, owing to the short recurrence. Unfortunately, it is impossible to extend the same efficiency of the CG method to a general system matrix A [136]. If the system matrix A is real-symmetric but indefinite, the minimum residual method (MINRES) has been shown to be effective since for each iteration only one MVP of A and seven vector operations are required [137].

When the system matrix A is nonsymmetric, there are three dominant types of Krylov methods. The first one is the generalized minimum residual method (GMRES) [138], which is a generalization of CG and has become a popular method. The GMRES requires an increasing amount of computational resources since a new orthogonal basis vector for the Krylov subspace has to be computed and stored at each iteration. This property implies that GMRES cannot be practically useful in case of poor convergence, i.e. when a large number of iterations yields only little improvement on the residual. A restarting technique can reduce the computational burden of GMRES, but this also significantly slows the convergence [139].

The Bi-CG method [140] is another generalization of the CG method for a general matrix A. The Bi-CG method is equivalent to CG in the symmetric case, but for a nonsymmetric A it requires the MVP with A^H, which makes it almost twice as expensive as CG. Furthermore, in practice the matrix A^H may not even be available, especially if the system A is implemented implicitly. The Bi-CG method is based on the nonsymmetric Lanczos method [141]. It was succeeded by its generalizations: Bi-CGSTAB [83], BiCGstab2 [84] and BiCGstab(ℓ) [85].

Another dominant type of Krylov method for nonsymmetric matrix A is the induced dimension reduction (IDR) method [142], followed by an improved and generalized algorithm IDR(s) [86] with a parameter s. IDR(1) is mathematically equivalent to BI-CGSTAB, and outperforms BI-CGSTAB when $s > 1$. IDR(s) has also been shown competitive with or superior to most BI-CG-based methods [86].

Back to the linear system (2.77) of the spatial spectral solver, the matrices C, G, M are implemented implicitly, and the system matrix $A = C - GM$ is nonsymmetric and indefinite. In [24], BiCGstab(ℓ) is used widely. Throughout this thesis, we take IDR(s) as the main iterative method.

2.4.2 Preconditioning

The efficiency of a computational method is usually measured by its accuracy and computation time. For iterative methods, the total computation time is determined by the total number of iterations required to reach a certain residual-error level and the computational cost per iteration. We hope for a small number of iterations, or in other words, a fast rate of convergence, with a relatively low cost per iteration. For a symmetric system matrix A, the rate of convergence for CG and MINRES can be guaranteed, since there exist descriptive convergence bounds, which are only relying on the distribution of the eigenvalues of A, see e.g. [135]. Therefore, given a residual tolerance, the number of iterations can be estimated and bounded. However, for a general matrix A, the convergence theory is very limited. For GMRES, Bi-CG-type methods, and IDR(s), there is not even a descriptive way to guarantee the convergence rate or to bound the number of iterations needed a priori. Estimating the rate of convergence of these methods for a general system matrix A is still an open theoretical problem.

Preconditioning is usually a crucial component in reducing the number of iterations and it is also widely used with Krylov subspace iterative methods. The essence of preconditioning can be understood as transforming the following original linear system

$$A\mathbf{u} = \mathbf{f}, \tag{2.81}$$

into a preconditioned system

$$PA\mathbf{u} = P\mathbf{f}, \tag{2.82}$$

for some matrix P, called a preconditioner. Note that the preconditioned system (2.82) shares the same solution with the original system (2.81) when P is nonsingular, and the system matrix in (2.82) becomes PA.

A good preconditioner P should always satisfy the following conditions:

- The number of iterations of the preconditioned system (2.82) is significantly reduced.
- It should not be expensive to construct P and execute the related MVP with P.

We consider two limiting cases. If $P = I$, i.e. P is the identity matrix, then the second condition above is satisfied, but the number of iterations will not be reduced at all. If

$P = A^{-1}$, then above the first condition above is satisfied, since only one iteration will be required, i.e. $\mathbf{u} = P\mathbf{f} = A^{-1}\mathbf{f}$, but computing this preconditioner itself is as expensive as solving the original system. Therefore, finding a good preconditioner is a trade-off and its complexity is always somewhere between I and A^{-1}.

In electromagnetic scattering problems, a large number of iterations is often observed in cases with high-contrast, negative-valued permittivities, or large scatterers. In Chapter 6 we propose a preconditioner for the spatial spectral method and show how this preconditioner reduces the number of iterations.

Chapter 3

Gabor coefficients computation of 2D indicator functions supported on polygonal domain based on the Taylor expansion of the complex error function.[1]

3.1 Computation of Gabor coefficients for objects with polygonal cross section

The Gabor transformation connects the spectral and spatial domain in a recently developed spatial spectral Maxwell solver. A key step involves computing the Gabor coefficients of the characteristic function of dielectric scattering objects. Therefore, computing Gabor coefficients accurately and efficiently is significant in the further development of this Maxwell solver. We discuss a method to numerically calculate the integrals involved in computing the Gabor coefficients, based on Gauss's theorem and recurrence relations.

3.1.1 Introduction

In optical scatterometry and metrology for the production of integrated circuits, a fast and reliable Maxwell solver is key to reconstruct the geometry parameters of metrology targets on the wafer. This is because the key geometrical details of metrology targets that are relevant for process control are much smaller than the wavelength of the light that illuminates the target. The tendency in the semiconductor industry is to shrink the area occupied by the targets, which implies that the finiteness of the target becomes observable and the periodicity assumption of the target no longer suffices. This has led us to develop a Maxwell solver based on a volume integral equation that is capable of simulating finitely

[1] Section 3.1 was published as [143] and Section 3.2 was published as [144]

sized dielectric targets in a layered medium [24]. In this Maxwell solver, the fields are expressed in terms of Gabor frames, that allow for efficient transformations between the (continuous) spatial and (continuous) spectral domains. The spatial domain is employed for the field-material interactions and the spectral domain for an efficient convolution with the layered-medium Green function. The Gabor frame discretization is an effective tool in this Maxwell solver to transform between spatial and spectral domain. However, as a consequence, the efficient and accurate calculation of Gabor coefficients to describe scattering objects in this representation plays a key role in this process [24].

The 3D full-wave Maxwell solver employs a Gabor-frame discretization in the two directions parallel to the layer interfaces. A two-dimensional (2D) cross-section of a scattering object embedded in a multi-layered medium can be approximated adequately by a collection of polygons. To compute Gabor coefficients of geometrical objects that represent dielectric scatterers embedded in the layered medium, Gabor transformations of 2D characteristic functions, which represent the 2D cross-sections of scattering objects, are required. These functions are equal to one on the support of a polygon and zero outside. One 2D integral is required for each Gabor coefficient in the geometrical representation of each polygon that represents part of a scattering object. Often, a large number of polygons is required, which leads to a vast number of two-dimensional integrals to be computed. In most cases the analytical solution of these integrals is hard to get due to the complicated structure of integrand and the support of the integration domain. Various numerical quadrature methods can be applied to acquire numerical approximations, such as quadrature rules based on interpolating functions. In real applications, these double integrals can occur billions of times, which brings a heavy computational burden to the preprocessing step of this type of Maxwell solver. Therefore, even a small improvement in this preprocessing step can bring a lot of benefits.

Inspired by the idea exploited in the local normal-vector field formulation in [107], we then transform a double integral into a sequence of line integrals by solving two ordinary differential equations and applying Gauss's theorem. Rather than computing the double integral, these line integrals can either be computed numerically via various quadrature rules or be developed further analytically in the form of recurrence relations. We discuss several examples and perform benchmarking to show how these two methods compare in terms of accuracy, numerical stability, and computation time.

In Section II, we give the statement of the problem and define the pertaining integrals needed to compute the Gabor coefficients of characteristic functions. Section III discusses a way to calculate the 1D line integrals based on a recurrence relation. In Section IV, we give two numerical examples to show how the methods work on rectangular and triangular cross sections. Finally, we draw conclusions in Section V.

3.1.2 Statement of the problem

3.1.2.1 2D Gabor transform

For any $f(x, y) \in \mathrm{L}^2(\mathbb{R}^2)$, we have its Gabor frame expansion

$$f(x, y) = \sum_{\mathbf{m},\mathbf{n}} f_{\mathbf{m},\mathbf{n}} g_{\mathbf{m},\mathbf{n}}(x, y), \tag{3.1}$$

in which $g_{\mathbf{m},\mathbf{n}}(x, y)$ is a Gabor frame function and $f_{\mathbf{m},\mathbf{n}}$ is a Gabor coefficient, and $\mathbf{m} = (m_x, m_y)$ and $\mathbf{n} = (n_x, n_y)$.

The Gabor frame functions (2D) are defined as:

$$g_{\mathbf{m},\mathbf{n}}(x, y) = g_{m_x,n_x}(x, \alpha_x X, \beta_x K_x) g_{m_y,n_y}(y, \alpha_y Y, \beta_y K_y), \tag{3.2}$$

where

$$g_{m,n}(x, X, K) = g_X(x - mX)e^{jnKx}, \tag{3.3}$$

where $g_X(x) = 2^{1/4} \exp[-\pi(x/X)^2]$ is Gaussian window function. Further, $m_x, m_y, n_x, n_y \in \mathbb{Z}$, α_x, α_y, β_x, β_y are oversampling parameters ($\alpha_x\beta_x = \alpha_y\beta_y = q/p$), and X, Y are spacing parameters of the window functions of the Gabor frame that satisfy $X = 2\pi/K_x$ and $Y = 2\pi/K_y$.

Gabor coefficients are defined via the 2D Gabor transformation:

$$f_{\mathbf{m},\mathbf{n}} = \iint_{\mathbb{R}^2} f(x, y)\eta^*_{\mathbf{m},\mathbf{n}}(x, y)dxdy, \tag{3.4}$$

where $\eta_{\mathbf{m},\mathbf{n}}(x, y)$ is the 2D dual window function, obtained by multiplying its counterparts in one-dimensional (1D). From [134], the 1D dual window can be represented again by the original Gabor frame expansion:

$$\begin{aligned} \eta_{m,n}(x) &= \eta(x - m\alpha X)e^{jn\beta Kx} \\ &= \sum_{l,k} \left(\gamma_{l,k} g_X(x - (m + l)\alpha X)e^{-jkm\alpha\beta KX + j(k+n)\beta Kx}\right), \end{aligned} \tag{3.5}$$

where $\gamma_{l,k}$ are Gabor coefficients. With this representation, the integral in (3.4) can be rewritten as a linear combination of integrals of the form:

$$\begin{aligned} I_{\mathbf{m+l},\mathbf{n+k}} = \iint_{\mathbb{R}^2} & f(x, y) \exp\left\{-\pi\left[\frac{x}{X} - (m_x + l_x)\alpha_x\right]^2\right\} \\ & \cdot \exp\left\{-j(k_x + n_x)\beta_x K_x x\right\} \\ & \cdot \exp\left\{-\pi\left[\frac{y}{Y} - (m_y + l_y)\alpha_y\right]^2\right\} \\ & \cdot \exp\left\{-j(n_y + k_y)\beta_y K_y y\right\} dxdy. \end{aligned} \tag{3.6}$$

This type of integral has to be computed many times.

We are interested in the Gabor coefficients for the characteristic function of a scattering object that coincides with a 2D domain D. The characteristic function $s(x, y)$ is then given by:

$$s(x,y) = \begin{cases} 1, & (x,y) \in D \\ 0, & (x,y) \in \mathbb{R}^2 \backslash D \end{cases}, \tag{3.7}$$

and the Gabor coefficients of this characteristic function can then be expressed in terms of the integrals defined in Eq. (3.6), where the domain of integration can now be restricted to the support D of $s(x, y)$.

3.1.2.2 A special case where analytical solution is available

In the case of a rectangular object that is aligned with the x and y direction, i.e., when integral domain D in (3.7) is an aligned rectangle, the double integral $I_{\mathbf{m+l,n+k}}$ defined in (3.12) is analytically integrable because the integrand is separable in its x and y arguments.

$$\begin{aligned} I_{\mathbf{m+l,n+k}} = & \int_{x_1}^{x_2} e^{-\pi[\frac{x}{X} - \alpha_x(m_x+l_x)]^2} e^{-j(n_x+k_x)\beta_x K_x x} dx \\ & \cdot \int_{y_1}^{y_2} e^{-\pi[\frac{y}{Y} - \alpha_y(m_y+l_y)]^2} e^{-j(n_y+k_y)\beta_y K_y y} dy \\ = & P_1 \cdot P_2 \cdot P_3, \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} P_1 = & \frac{XY}{4} e^{-\pi\beta_x(n_x+k_x)[\beta_x(n_x+k_x)+2i\alpha_x(m_x+l_x)]} \\ & e^{-\pi\beta_y(n_y+k_y)[\beta_y(n_y+k_y)+2i\alpha_y(m_y+l_y)]}, \end{aligned} \tag{3.9}$$

$$\begin{aligned} P_2 = & \operatorname{erf}\left[-\frac{\sqrt{\pi}}{X}(\alpha_x X(m_x+l_x) - x_1) + i\beta_x\sqrt{\pi}(n_x+k_x)\right] \\ & - \operatorname{erf}\left[-\frac{\sqrt{\pi}}{X}(\alpha_x X(m_x+l_x) - x_2) + i\beta_x\sqrt{\pi}(n_x+k_x)\right], \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} P_3 = & \operatorname{erf}\left[-\frac{\sqrt{\pi}}{Y}(\alpha_y Y(m_y+l_y) - y_1) + i\beta_y\sqrt{\pi}(n_y+k_y)\right] \\ & - \operatorname{erf}\left[-\frac{\sqrt{\pi}}{Y}(\alpha_y Y(m_y+l_y) - y_2) + i\beta_y\sqrt{\pi}(n_y+k_y)\right]. \end{aligned} \tag{3.11}$$

When the integration domain D is not a rectangle, the double integral $I_{\mathbf{m+l,n+k}}$ in (3.12) cannot be decomposed into a product of two integrals to obtain an analytical solution. Therefore, numerical integration is required to evaluate $I_{\mathbf{m+l,n+k}}$.

3.1.2.3 Gauss's theorem

Based on Gauss's theorem we can rewrite the integral type $I_{\mathbf{m+l,n+k}}$ as a contour integral over functions $P(x,y)$ and $Q(x,y)$, owing to the finite support of $s(x,y)$, i.e.

$$\begin{aligned} I_{\mathbf{m+l,n+k}} &= \iint_D \frac{\partial}{\partial x}P(x,y) + \frac{\partial}{\partial y}Q(x,y)dxdy \\ &= \oint_{\partial D} -Q(x,y)dx + P(x,y)dy. \end{aligned} \tag{3.12}$$

Functions $P(x,y)$ and $Q(x,y)$ are given by

$$\begin{aligned} P(x,y) =& h_1(y,\mathbf{p}_2)\cdot \mathrm{erf}[h_2(x,\mathbf{p}_3)], \\ Q(x,y) =& h_1(x,\mathbf{p}_1)\cdot \mathrm{erf}[h_2(y,\mathbf{p}_4)], \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} h_1(x,\mathbf{p}_1) =& \frac{Y}{4}\mathrm{exp}\bigg\{ -\frac{\pi}{X^2}(x-\alpha_x(m_x+l_1)X)^2 \\ & -\pi\beta_y^2(n_y+k_y)^2 - j\Big(\beta_x K_x(n_x+k_x)x \\ & +2\pi\alpha_y\beta_y(m_y+l_y)(n_y+k_y)\Big)\bigg\}, \end{aligned} \tag{3.14}$$

$$\begin{aligned} h_1(y,\mathbf{p}_2) =& \frac{X}{4}\mathrm{exp}\bigg\{ -\frac{\pi}{Y^2}(y-\alpha_y(m_y+l_y)Y)^2 \\ & -\pi\beta_x^2(n_x+k_x)^2 - j\Big(\beta_y K_y(n_y+k_y)y \\ & +2\pi\alpha_x\beta_x(m_x+l_x)(n_x+k_x)\Big)\bigg\}, \end{aligned} \tag{3.15}$$

$$\begin{aligned} h_2(x,\mathbf{p}_3) =& \sqrt{\pi}\left(\frac{x}{X}-\alpha_x(m_x+l_x)\right) \\ & + j\sqrt{\pi}\beta_x(n_x+k_x), \end{aligned} \tag{3.16}$$

$$\begin{aligned} h_2(y,\mathbf{p}_4) =& \sqrt{\pi}\left(\frac{y}{Y}-\alpha_y(m_y+l_y)\right) \\ & + j\sqrt{\pi}\beta_y(n_y+k_y), \end{aligned} \tag{3.17}$$

and vectors

$$\begin{aligned} \mathbf{p}_1 &= (X,Y,\alpha_x,\alpha_y,m_x+l_x,n_x+k_x,m_y+l_y,n_y+k_y), \\ \mathbf{p}_2 &= (Y,X,\alpha_y,\alpha_x,m_y+l_y,n_y+k_y,m_x+l_x,n_x+k_x), \\ \mathbf{p}_3 &= (X,\alpha_x,m_x+l_x,n_x+k_x), \\ \mathbf{p}_4 &= (Y,\alpha_y,m_y+l_y,n_y+k_y), \end{aligned}$$

contain independent parameters. Essentially, $P(x,y)$ and $Q(x,y)$ are the same function with different variable and parameter notations.

3.1.2.4 Parametrization for polygonal object

Suppose the scattering object is an M-sided polygon, and along the i-th edge of the polygon (for $i = 1, 2, \dots, M$ counterclockwise), we apply the following parametrization:

$$\begin{cases} \frac{x_i(t)}{X} = c_{i1}t + d_{i1}, \\ \frac{y_i(t)}{Y} = c_{i2}t + d_{i2}, \end{cases} \tag{3.18}$$

where $t \in [0, 1)$. After substituting (3.18) in (3.12) we get

$$I_{\mathbf{m+l,n+k}} = \sum_{i=1}^{M} [I(\mathbf{c}_i) + I(\mathbf{d}_i)], \tag{3.19}$$

and the representations of $I(\mathbf{c}_i)$ and $I(\mathbf{d})_i$ are:

$$I(\mathbf{c}_i) = c_0^i \int_0^1 \exp\left(-c_1^i t^2 + c_2^i t\right) \operatorname{erf}(c_3^i t + c_4^i) dt, \tag{3.20}$$

$$I(\mathbf{d}_i) = d_0^i \int_0^1 \exp\left(-d_1^i t^2 + d_2^i t\right) \operatorname{erf}(d_3^i t + d_4^i) dt, \tag{3.21}$$

where vectors

$$\mathbf{c}_i = (c_0^i, c_1^i, c_2^i, c_3^i, c_3^i),$$
$$\mathbf{d}_i = (d_0^i, d_1^i, d_2^i, d_3^i, d_3^i),$$

denote the boundary-related parameters, which satisfy $c_0^i, d_0^i, c_2^i, d_2^i, c_4^i, d_4^i \in \mathbb{C}$, $c_1^i, d_1^i, c_3^i, d_3^i \in \mathbb{R}$, $c_1^i, d_1^i \geq 0$.

3.1.3 Recurrence relation and Olver's algorithm

We are not aware of any method to find a general solution for (3.20) analytically, due to the product of the Gaussian function and the complex error function $\operatorname{erf}(z)$ in the integrand. Numerical integration is possible but slow. To avoid a numerical sampling of the integral, we use a Taylor series expansion to replace the complex error function $\operatorname{erf}(z)$.

3.1.3.1 Taylor series expansion

Notice that the product of a Gaussian function and polynomial is analytically integrable, so we consider the Taylor series of a complex error function first.

We have

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)n!} z^{2n+1}, \tag{3.22}$$

and for $\forall z \in \mathbb{C}$ this Taylor series converges to $\operatorname{erf}(z)$ since it is an entire function [145].

In practice we need to truncate this infinite series after a finite number of terms to approximate $\mathrm{erf}(z)$. Hence for a given $N \in \mathbb{N}$, we can represent $\mathrm{erf}(z)$ as:

$$\begin{aligned}\mathrm{erf}(z) =&\frac{2}{\sqrt{\pi}}\left[\sum_{n=0}^{N-1}\frac{(-1)^n z^{2n+1}}{(2n+1)n!}+\sum_{n=N}^{\infty}\frac{(-1)^n z^{2n+1}}{(2n+1)n!}\right] \\ \triangleq& P_{2N}(z)+R_{2N}(z).\end{aligned} \tag{3.23}$$

Notice that $P_{2N}(z)$ is actually equivalent to the first $2N$ items in the original Taylor series of $\mathrm{erf}(z)$ in (3.22). For a given $\varepsilon > 0$, this truncation number N can be determined correspondingly.

3.1.3.2 Recurrence relation

Recalling (3.20) and omitting the boundary notation i, we have:

$$I_\alpha = c_0 \int_0^1 e^{-c_1 t^2 + c_2 t}\mathrm{erf}(c_3 t + c_4)dt, \tag{3.24}$$

where $c_0, c_2, c_4 \in \mathbb{C}$, $c_1, c_3 \in \mathbb{R}$, $c_1 \geq 0$. Together with (3.23) we have

$$\begin{aligned}I_\beta =& c_0 \int_0^1 e^{-c_1 x^2 + c_2 x} P_{2N}(c_3 x + c_4)dx \\ =&\frac{2c_0}{\sqrt{\pi}}\sum_{n=0}^{N-1}\frac{(-1)^n}{(2n+1)n!}\int_0^1 e^{-c_1 x^2 + c_2 x}(c_3 x + c_4)^{2n+1}dx \\ =&\frac{2c_0 e^{c_7}}{\sqrt{\pi}c_3}\sum_{n=0}^{N-1}\frac{(-1)^n}{(2n+1)n!}\int_{c_4}^{c_3+c_4} e^{-c_5 y^2 + c_6 y}y^{2n+1}dy,\end{aligned} \tag{3.25}$$

where $y = c_3 x + c_4$ and when $c_3 \neq 0$ we have

$$c_5 = \frac{c_1}{c_3^2}, \tag{3.26}$$

$$c_6 = \frac{2c_1 c_4}{c_3^2} + \frac{c_2}{c_3}, \tag{3.27}$$

$$c_7 = -\frac{c_1 c_4^2}{c_3^2} - \frac{c_2 c_4}{c_3}. \tag{3.28}$$

We note that if $c_3 = 0$, the integral in (3.24) has a closed-form expression, see Section 3.1.4.1. For fixed c_1, c_2, the difference between I_α and I_β is also bounded, and therefore I_β is an approximation of original integral I_α.

To ease the following analysis, we define the sequence

$$p_m = \frac{1-(-1)^m}{2} = \begin{cases}1, & m \text{ is odd}, \\ 0, & m \text{ is even}.\end{cases} \tag{3.29}$$

Therefore, if we let $m = 2n+1$, then $n = \frac{1}{2}(m-1)$, $0 \leq n \leq N-1$, and we have:

$$\begin{aligned}
I_\beta &= \frac{2c_0 e^{c_7}}{\sqrt{\pi}c_3} \sum_{m=0}^{2N-1} \frac{(-1)^{\frac{m-1}{2}}}{m(\frac{m-1}{2})!} p_m \int_{c_4}^{c_3+c_4} e^{-c_5 y^2 + c_6 y} y^m dy \\
&= \frac{2c_0 e^{c_7}}{\sqrt{\pi}c_3} \sum_{m=0}^{2N-1} \frac{(-1)^{\frac{m-1}{2}}[1-(-1)^m]}{2m(\frac{m-1}{2})!} \\
&\qquad \cdot \int_{c_4}^{c_3+c_4} e^{-c_5 y^2 + c_6 y} y^m dy \\
&\triangleq \frac{2c_0 e^{c_7}}{\sqrt{\pi}c_3} \sum_{m=0}^{2N-1} d_m I_m,
\end{aligned} \tag{3.30}$$

where

$$d_m = \frac{(-1)^{\frac{m-1}{2}}}{m} \cdot \frac{[1-(-1)^m]}{2}, \tag{3.31}$$

$$I_m = \frac{1}{(\frac{m-1}{2})!} \int_{c_4}^{c_3+c_4} e^{-c_5 y^2 + c_6 y} y^m dy. \tag{3.32}$$

With integration by parts applied to (3.32), we find that the sequence I_m satisfies the following second-order inhomogeneous linear recurrence relation:

$$I_{m-1} + \frac{c_6(\frac{m-1}{2})!}{2(\frac{m}{2})!} I_m - c_5 I_{m+1} = \frac{1}{2(\frac{m}{2})!}(\gamma_m - \xi_m), \tag{3.33}$$

where

$$\gamma_m = c_0 e^{-c_5(c_3+c_4)^2 + c_6(c_3+c_4)}(c_3+c_4)^m, \tag{3.34}$$

$$\xi_m = c_0 e^{-c_5 c_4^2 + c_6 c_4} c_4^m. \tag{3.35}$$

Once (3.33) is solved with two initial conditions, we obtain an approximation of I_α without numerical quadrature.

3.1.3.3 Olver's algorithm

It is well known that computing the minimal solution of a second-order difference equation, or three-term recurrence relation, directly from given initial values y_0, y_1 is an unstable procedure, see e.g. [146, 147]. F. W. J. Olver proposed a classic and stable algorithm for second-order inhomogeneous linear difference equations [148] in this form:

$$\begin{aligned}
&a_r y_{r-1} - b_r y_r + c_r y_{r+1} = d_r, \\
&y_0 = k, \quad y_1 = m,
\end{aligned} \tag{3.36}$$

where $a_r, b_r, c_r, d_r \in \mathbb{C}$ are given sequences, $r \in \mathbb{N}$ and k, $m \in \mathbb{C}$ are given constants.

The main idea of Olver's algorithm is to transfer the initial value problem to an equivalent boundary value problem (BVP). To do this, Olver abandoned the initial condition $y_1 = m$ and replaced it by another boundary condition $y_N = p$, which is obtained based on the asymptotic property of the minimal solution y_r, for sufficiently large N. In most cases (also in our case later on), $y_N = 0$. The solution of the well-conditioned BVP is $y_r^{(N)}$, Olver also proved that $y_r^{(N)}$ convergences to the required solution y_r of the original initial value problem as $N \to \infty$. This linear BVP can be solved efficiently by either Gaussian elimination or LU decomposition [149]. Therefore, for sufficient large N, $y_r^{(N)}$ will be an approximate solution of the difference equation (3.36). Olver's algorithm can also prevent unstable error propagation when solving higher-order inhomogeneous difference equations [150].

Back to problem (3.33) and comparing with (3.36), we have

$$a_r I_{r-1} - b_r I_r + c_r I_{r+1} = d_r, \tag{3.37}$$

where

$$a_r = 1,\, b_r = -\frac{c_6(\frac{m-1}{2})!}{2(\frac{m}{2})!}, \tag{3.38}$$

$$c_r = -c_5,\, d_r = \frac{1}{2(\frac{m}{2})!}(\gamma_m - \xi_m). \tag{3.39}$$

Hence the system (3.33) can be solved stably via Olver's algorithm.

3.1.4 Numerical experiments

Generally, every M-sided polygon can be decomposed into a combination of rectangles and right-angled triangles, so in this section we try to reconstruct these two fundamental objects with the method proposed in the preceding sections.

3.1.4.1 Rectangular object

In the first experiment we consider a support function $s_1(x, y)$ defined on a rectangular domain with coordinates $A(-\frac{1}{2}, -\frac{1}{2}), B(\frac{1}{2}, -\frac{1}{2}), C(\frac{1}{2}, \frac{1}{2}), D(-\frac{1}{2}, \frac{1}{2})$.

The Gabor parameters are given by: $-5 \leq m_i, n_i, k_i, l_i \leq 5$ and $m_i, n_i, k_i, l_i \in \mathbb{Z}$ where $i = x, y$.

$$X = Y = 1, K_x = K_y = 2\pi. \tag{3.40}$$

$$\alpha_x = \alpha_y = \beta_x = \beta_y = \sqrt{\frac{2}{3}}. \tag{3.41}$$

From Section 3.1.2.2 we recall that an analytical solution exists for (3.12) that can be used as a reference to check the method of the recurrence relation for the calculation of the Gabor coefficients.

Based on the parametrization step (3.18) of this rectangular object's boundary, it is easy to notice that along each side we have either $c_0 = 0$ or $c_3 = 0$ in (3.20). I_α vanishes when $c_0 = 0$, and when $c_3 = 0$, I_α is integrable:

$$\begin{aligned} I_\alpha =& c_0 \int_0^1 e^{-c_1 t^2 + c_2 t} \mathrm{erf}(c_4) dt \\ =& \frac{c_0\sqrt{\pi}}{2\sqrt{c_1}} \mathrm{erf}(c_4) e^{\frac{c_2^2}{4c_1}} \left[\mathrm{erf}(\frac{2c_1 - c_2}{2\sqrt{c_1}}) + \mathrm{erf}(\frac{c_2}{2\sqrt{c_1}}) \right]. \end{aligned} \tag{3.42}$$

Therefore, in the rectangular case the Gabor coefficients calculated based on the line integration are the same as the analytical solutions for the double integral. Figure 3.1 shows the reconstructed characteristic function defined on the rectangular domain.

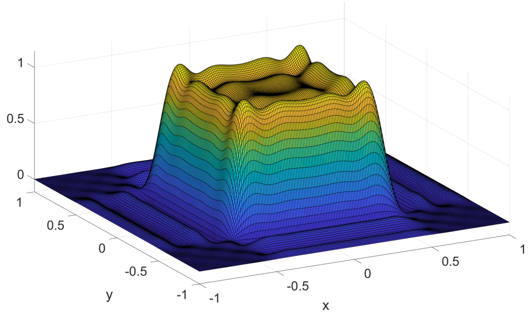


Figure 3.1: Reconstructed characteristic function for a rectangular object, via computed Gabor coefficients. The original characteristic function $f_1(x, y)$ was supported on a rectangle with vertex coordinates $A(-\frac{1}{2}, -\frac{1}{2}), B(\frac{1}{2}, -\frac{1}{2}), C(\frac{1}{2}, \frac{1}{2}), D(-\frac{1}{2}, \frac{1}{2})$. Oversampling parameter $q = 2$, $p = 3$, spacing of the window function $X = Y = 1$, spacial shift $-5 \leq m_x, m_y \leq 5$, spectral shift $-5 \leq n_x, n_y \leq 5$.

3.1.4.2 Triangular object

In the second experiment we consider a characteristic function $f_2(x, y)$ defined on a triangular domain with vertex coordinates $A(-\frac{1}{2}, -\frac{1}{2}), B(\frac{1}{2}, -\frac{1}{2}), C(\frac{1}{2}, \frac{1}{2})$. We keep the same Gabor parameter setting as in the previous experiment.

In this triangular case the double integral in (3.6) is not analytically integrable and the line integral (3.12) along the sloped boundary is non-trivial. Therefore, this is a good benchmark to test the proposed method in a non-trivial case. In this experiment we use direct numerical quadrature on (3.12) as a reference. Figure 3.2 shows the reconstructed characteristic function for the triangular object.

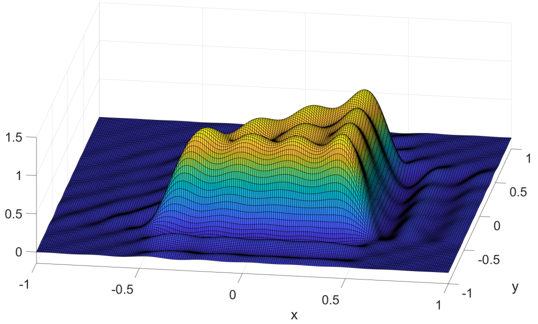


Figure 3.2: Reconstructed characteristic function for a triangular object based on recurrence relation and Olver's algorithm. The original characteristic function $f_2(x, y)$ was supported on a triangle with vertex coordinates $A(-\frac{1}{2}, -\frac{1}{2}), B(\frac{1}{2}, -\frac{1}{2}), C(\frac{1}{2}, \frac{1}{2})$. Same parameters were used as in Fig. 3.1.

The original support function defined on a triangular domain can be reconstructed successfully with the proposed method based on the recurrence relation and Olver's algorithm. However, 4.66% of the total calculated Gabor coefficients have an absolute error larger than 10^{-4} compared to the numerical reference.

There are two reasons why the proposed approximate integral I_β has large errors at some points.

Firstly, the real part and imaginary part of $\text{erf}(z)$ itself can get very large when z is far from the origin and complex-valued. Convergence of the Taylor series around the origin is poor compared to the increase of $\text{erf}(z)$. Therefore, thousands of terms are required to reach the qualified approximation, which results in difficulties when solving the recurrence relation. However, to make sure there exists a recurrence relation in the truncated series, one must use the Taylor series around the origin to approximate $\text{erf}(z)$, since there is no known explicit representation for n-th derivative of $\text{erf}(z)$. Even so, at some points the error can still be large along some directions, for example in the current case for the triangle it can be larger than 10^{-4}.

Secondly, the absolute value of the integrals I_m can be very large before vanishing eventually, which means a high working precision is required to represent them accurately in the propagation through the backward substitution process in the recurrence relation. Furthermore, the coefficient sequence d_m in (3.31) is alternating. Hence, one must set a high working precision to avoid loss of significant figures and due to cancellation errors. A high working precision results in an increase in computation time.

Figure 3.3 displays the absolute value of I_m with parameter setting: $m_x + l_x = 1, m_y + l_y = 21, n_x + k_x = 18, n_y + k_y = 19$, and the maximum of $|I_m|$ reaches a level of 10^{360}.

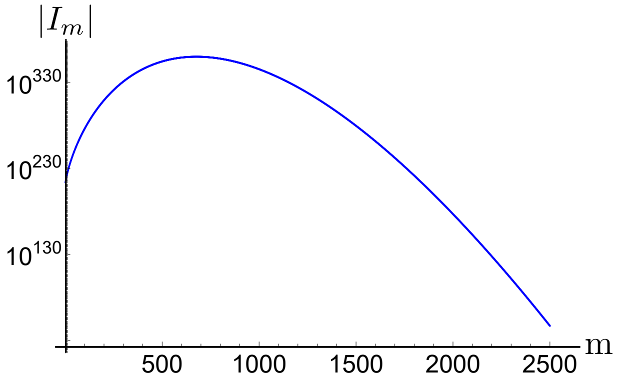


Figure 3.3: In this specific example, we have $N = 1250$, i.e., 2500 integrals I_m were calculated based on Olver's algorithm to approximate $\mathrm{erf}(z)$. The error in the approximation is 1.10652×10^{-129}.

Based on this sequence I_m, we obtained an approximate integral I_β and therefore one Gabor coefficient. Note that a total of 1166886 integrals were computed to reconstruct Figure 3.2 and we obtain a difference $I_\alpha - I_\beta = -1.09719 \times 10^{-129} - 1.4339 \times 10^{-130} j$.

3.1.5 Conclusion and outlook

The complex error function $\mathrm{erf}(z)$ and the Faddeeva function (also known as the plasma dispersion function) $w(z) = e^{-z^2}\mathrm{erfc}(-iz)$ are connected by elementary relations [151], and both of them can be evaluated in different ways such as a Padé rational approximation [152], a rational Chebyshev approximation by Schonfelder [153], continued fractions [154] and also what we used here Taylor series expansion [145]. In this paper we selected a Taylor-series expansion to approximate the complex error function $\mathrm{erf}(z)$, and then approximated the target integral I_α that occurs in the calculation of Gabor coefficients. The main reason for choosing a Taylor-series expansion is that it can yield a second-order inhomogeneous recurrence relation, which is beneficial in computing Gabor coefficients because calculating abundant integrals is not necessary anymore, as long as this recurrence relation can be solved stably with two initial conditions. Olver's algorithm was used to complete this means for computing Gabor coefficients. As applications for this proposed method, two characteristic functions defined on rectangular domain and triangular domain were reconstructed. In the triangular case, we observed that 4.66% of the Gabor coefficients have an error larger than the set threshold of 10^{-4} and two reasons were given for this outcome.

In future research, other expansions are expected to better approximate the complex error function within the integral I_α, with which the new basis function sets either yield much faster convergence (and therefore coefficients would vanish dramatically and the approximated integral can be evaluated even term by term), or relatively faster convergence but at meantime still generate a recurrence relation.

Acknowledgment

This work was funded by NWO-TTW as part of the HTSM program under project number 16184 and as part of the NWO-TTW perspective program under project number P16-08. L. Sun is grateful to B. P. de Hon and J. M. L. Maubach from TU Eindhoven and S. Chevillard from the French Institute for Research in Computer Science and Automation for their suggestions and comments.

3.2 A note on Gabor coefficient computing with Taylor series expansion

We present an improvement on a previously proposed method for computing Gabor coefficients of characteristic functions with polygonal cross sections, based on a Taylor series expansion and Olver's algorithm. Several requirements are proposed to make the method more robust. Numerical evidence is given to show a convergent solution can be obtained based on a sufficiently high truncation number and working precision.

3.2.1 Introduction

In [143] a numerical method to compute Gabor coefficients for objects with polygonal cross sections was proposed. The key components of this method are: (1) a 1D-integral formulation derived from a double integral based on Gauss's theorem, (2) a Taylor series expansion of the complex error function, (3) derivation of a second-order inhomogeneous difference equation and (4) solution with Olver's algorithm. The main benefit of this method is that it transforms an integration problem into an evaluation problem, where the former can be computationally expensive when the complex error function is contained in the integrand. However, as observed in the second numerical experiment in [143], this method failed on some points.

We explain why this method failed on those points previously and we remedy the problem by introducing several requirements for this method. Furthermore, we show how this method can yield a convergent solution with these requirements.

3.2.2 Requirements of the Taylor-Olver method

Gabor coefficients for characteristic functions supported on a polygonal domain can be computed using the following fundamental integrals [143]:

$$I = \int_0^1 e^{-c_1x^2+c_2x}\mathrm{erf}(c_3x+c_4)dx, \tag{3.43}$$

where c_1, c_2, c_3, c_4 are given constants. Then we use a truncated Taylor series to approximate the complex error function:

$$\mathrm{erf}(z) \approx \frac{2}{\sqrt{\pi}}\sum_{n=0}^{N-1}\frac{(-1)^n z^{2n+1}}{(2n+1)n!}, \tag{3.44}$$

which thereafter yields an approximated integral $\tilde{I}$:

$$\tilde{I} = \frac{2e^{c_7}}{\sqrt{\pi}c_3} \sum_{m=0}^{2N-1} s_m I_m, \tag{3.45}$$

where

$$s_m = \begin{cases} \frac{(-1)^{\frac{m-1}{2}}}{m}, & m \text{ is odd}, \\ 0, & m \text{ is even}, \end{cases}$$
$$I_m = \frac{1}{(\frac{m-1}{2})!} \int_{c_4}^{c_3+c_4} e^{-c_5 y^2 + c_6 y} y^m dy,$$

and c_5, c_6, c_7 are given constants. Note that only the odd-indexed I_m contribute to the final result. By applying integration by parts, one obtains the following second-order difference equation:

$$I_{m-1} - b_m I_m - c_5 I_{m+1} = d_m, \forall m \geq 1 \tag{3.46}$$

where

$$b_m = -\frac{c_6(\frac{m-1}{2})!}{2(\frac{m}{2})!},$$
$$d_m = p(m, c_3 + c_4) - p(m, c_4),$$

and $p(m, y) = \frac{1}{2(m/2)!} e^{-c_5 y^2 + c_6 y} y^m$. The half-integer factorials in b_m are calculated with the Γ function. We use Olver's algorithm to solve this equation and assemble I_m together to get Gabor coefficients [143].

The following requirements emphasize three crucial points to obtain correct Gabor coefficients when using the proposed Taylor-Olver method.

Requirement 1: The truncation number N of the Taylor series in Eq. (3.44) must be sufficient. An accurate approximation of the partial sum in Eq. (3.44) to the complex error function $\text{erf}(z)$ is a necessary condition to obtain an accurate approximated integral $\tilde{I}$ in Eq. (3.45). The truncation number N can be determined based on either desired accuracy [145] or numerical evidence.

Requirement 2: The working precision w, which indicates how many significant digits should be maintained in internal computations, must be high enough to guarantee the final accuracy. This is because:

- the integrals I_m can reach an extremely high value before vanishing eventually, e.g., the I_m in Fig. 3 of [143] reaches a level of 10^{360}, where the Gabor coefficient itself is a small number.

- the coefficient sequence s_m in (3.45) is alternating, therefore large cancellation errors occur if the working precision is not high enough.

- the truncated tridiagonal system is sensitive to d_m, which means the first term I_0 of the difference equation must be calculated with high accuracy.

Numerical evidence shows that a working precision $w = N$ yields stable results.
Requirement 3: Truncation number N' in Olver's algorithm should be large enough. Olver's algorithm transforms an semi-infinite matrix system, which is corresponding to the difference equation, into a truncated tridiagonal system. Olver provided a way to automatically determine the truncation number based on the desired accuracy [148]. Numerical evidence shows that a truncation number of the tridiagonal system $N' = 1.5N$ yields a stable result.

3.2.3 Numerical Results

To demonstrate the importance of above requirements, we recalculated one of the failed integrals in the second experiment in [143] for $c_1 = 3.14$, $c_2 = 49.3 - 5.1i$, $c_3 = -1.8$ and $c_4 = 15.4 - 14.5i$ in Eq. (3.43). One can observe the range of $\text{erf}(c_3 x + c_4)$ for $x \in [0, 1]$ in Fig. 3.4. A large truncation number N is needed for the Taylor series to converge due to a relatively large distance from the origin, which therefore makes this I one of the most difficult ones in the triangle example to compute with the Taylor-Olver method.

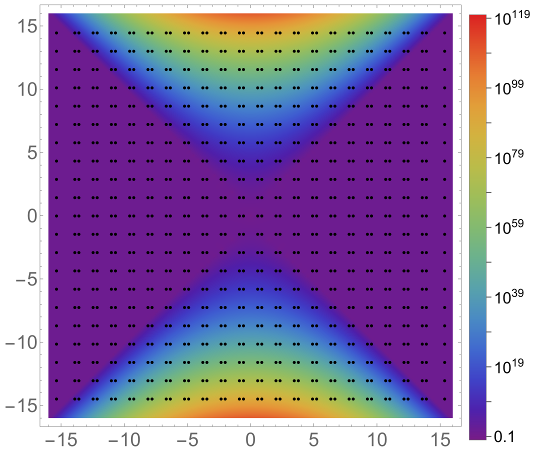


Figure 3.4: Magnitude of $\text{erf}(z)$ on a log scale. The black dots represent all c_4 occurring in Simulation 2 in [143]

Following above three requirements, we obtained the result $\tilde{I} = -1.87 \times 10^{27} + 1.72 \times 10^{26} i$. Compared with a high accuracy numerical reference, this solution has absolute error 1.52×10^{-95} and relative error 8.09×10^{-123}. The truncation number of the Taylor series is $N = 2800$, the working precision used is $w = 2800$, the dimension of the truncated tridiagonal system in Olver's algorithm is $N' = 4200$. Fig. 3.5 shows the computed integral sequence I_m from Eq. (3.45), compared with the numerical reference.

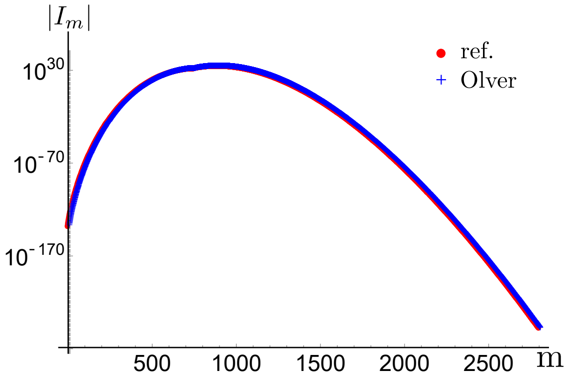


Figure 3.5: Solution of Eq. (3.46) based on Olver's algorithm.

Fig. 3.6 shows a convergent solution obtained by increasing the truncation number N of the Taylor series, as along as the proposed requirements are satisfied. This result also implies that an insufficient truncation number of the Taylor series can be catastrophic.

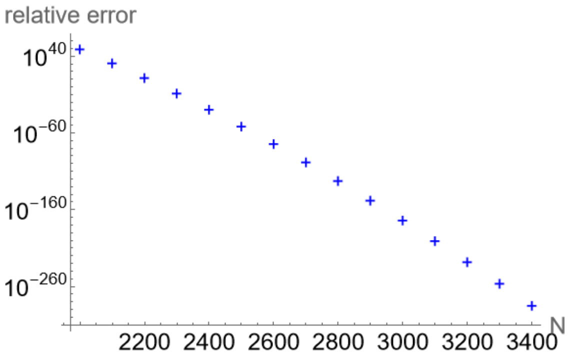


Figure 3.6: Convergence obtained by increasing the truncation number of the Taylor series in Eq. (3.44).

Overall, we proposed three requirements to make the previous Taylor-Olver method more robust. In the future, optimization of the working precision should be considered to reduce the computation time.

Acknowledgment

This work was funded by NWO-TTW as part of the HTSM program under project number 16184.

Chapter 4

Gabor coefficients computation of 2D indicator functions supported on polygonal domain based on a rational expansion of the Faddeeva function[1]

We propose a method to compute Gabor coefficients of a two-dimensional (2D) indicator function supported on a polygonal domain by means of rational expansion of the Faddeeva function and by solving second-order linear difference equations. This method has the following three attractive features: (1) the problem of computing Gabor coefficients is formulated as the calculation of a sequence of integrals with a uniform structure, (2) a rational expansion based on fast Fourier transform (FFT) is used to approximate the Faddeeva function on the entire complex plane, (3) second-order inhomogeneous linear difference equations are derived for previous integrals and they are solved stably with Olver's algorithm. Numerical quadrature to compute Gabor coefficients is avoided. Numerical examples show this rational-expansion-based method significantly outperforms numerical quadrature in terms of computation time while maintaining accuracy.

4.1 Introduction

Gabor analysis has become an active research area with its foundation widely developed [129, 132, 133], since Dennis Gabor suggested to use translated and modulated elementary functions to analyze a signal in 1946 [156]. Later developed Gabor frames play an important role in time-frequency analysis owing to their convenient series expansions to represent a function in $L^2(\mathbb{R}^d)$, based on a decomposition into translated and modulated versions of a window function. As applications, Gabor frames have been successfully used in e.g. image processing [130, 157, 158], signal and wireless communications [159–161], Gaussian beams [162, 163], antenna analysis and design [164–166]. The related Wilson basis, which

[1] This chapter was published as [155]

can be constructed from a tight Gabor frame with redundancy 2, has also been used to solve electromagnetic scattering problems, reflection-transmission problems, and optical fiber connection problems [167–169].

Gabor frame theory has been applied in computational electromagnetics as well. This work is intended for the Gabor-frame-based spatial spectral Maxwell solver for layered media of [24, 108–110], where a Gabor frame discretization is used to constitute efficient transformations between the (continuous) spatial and (continuous) spectral domains. To model an electromagnetic scattering problem accurately, the contrast function, the contrast current density and the electric fields must be represented well by Gabor frames. On the transverse plane, the contrast function is described by a 2D indicator function. When the scatterer has a polygonal cross section, efficient and accurate calculation of Gabor coefficients of a 2D indicator function supported on this polygonal domain is therefore important. A Gabor coefficient of a 2D indicator function supported on a polygonal domain is defined as a 2D integral, but it can be transferred into a summation of 1D integrals that contain an exponential function and the complex error function in the integrand, see [143, 170]. It is well known that evaluation of integrals with complicated integrands or weakly singular kernels, for instance poorly behaved special functions, is an important and difficult computational problem [171, 172]. For the integrals containing the complex error function inside, there is no closed form available in most cases and numerical integration techniques are therefore required, e.g., Gauss–Legendre quadrature is shown as an efficient numerical method in [170]. Computing integrals of this specific type accurately and efficiently is crucial to reduce the associated computational burden. Nevertheless, we intend to further develop methods to analytically simplify and ease the computational procedure in the calculation of Gabor coefficients.

In [143], the authors derived second-order difference equations for these integrals based on the Taylor-series expansion of the complex error function as shown in [173]. Most numerical integration can be avoided once the difference equations are solved by Olver's algorithm. On the other hand, there is a remaining issue that the Taylor series expansion requires thousands of polynomial terms to reach the desired accuracy in certain regions of the complex plane, which takes this approach far away from practical usefulness. Later on in [144], three requirements were proposed to improve this Taylor-Olver method and numerical evidence showed that a convergent solution can be obtained based on a sufficiently high truncation number and working precision. However, this method still requires too much computation time due to the slow convergence of the Taylor-series expansion of the complex error function. Hence, we are seeking a better expansion that can yield not only a difference equation, but also exhibits a fast convergence to the complex error function, the Faddeeva function, or other members of the same family of special functions.

The elementary approximation of the complex error function or the Faddeeva function is a well-studied research area. When the argument of these functions becomes complex-valued, there exist several classes of algorithms to evaluate these functions to high precision, such as a Padé approximation [152, 174], methods based on the repeated trapezoidal rule [175], a Chebyshev approximation [153], continued fractions [151, 154], and rational approximations [113, 176]. Among these approximations, the three rational expansions proposed in [113] provide high-accuracy approximations to the Faddeeva function on the entire first quadrant of the complex plane, with only a small number of terms. Furthermore, the coefficients of these rational expansions can be computed once and for all by a single FFT. The rapid convergence of these rational expansions and their implementation advantages are especially interesting to us and also motivate the current work on Gabor coefficient computation for 2D indicator functions of polygonal support.

Based on one of the rational expansions to the Faddeeva function in [113], we propose a method to compute Gabor coefficients of a 2D indicator function supported on a polygonal domain. First of all, the problem of computing Gabor coefficients is transformed into a formulation that relies on a sequence of integrals with a uniform structure. Secondly, the rational expansion is used to approximate the Faddeeva function in the integrand on the entire complex plane. Then we derive second-order inhomogeneous linear difference equations and solve the sequence of integrals stably with Olver's algorithm [148, 177]. With this rational-expansion-based method, a fixed small number of function evaluations is required to compute a Gabor coefficient. A large number of function evaluations, as occurs in most numerical integration methods, is avoided. We then test and analyse this method by computing Gabor coefficients for two indicator functions supported on a triangle and a five-pointed star-shaped polygon.

The structure of this chapter is as follows. In Section 4.2, we briefly review some fundamentals in Gabor analysis, state the problem to be studied by deriving the fundamental integral and apply a rational expansion of the Faddeeva function. Subsequently, we get a sequence of integrals and expand this formulation to the entire complex plane. In Section 4.3 we derive the second-order difference equations and solve the sequence of integrals with Olver's algorithm. Section 4.4 contains two numerical examples and we test this method's accuracy and computation time. Conclusions are drawn in Section 4.5.

4.2 Formulation

We start the formulation by recalling one of the fundamental results in Gabor analysis [129, 132]. For a given Gabor frame $\mathcal{G}(g, \alpha, \beta)$ with window function $g \in L^2(\mathbb{R}^d)\backslash\{0\}$, $\alpha, \beta > 0$ and $\alpha\beta < 1$, there exists a dual window function $\eta \in L^2(\mathbb{R}^d)$ and an induced dual Gabor frame. Every $f(\mathbf{x}) \in L^2(\mathbb{R}^d)$ has the following Gabor-frame expansion:

$$f(\mathbf{x}) = \sum_{\mathbf{k},\mathbf{l}\in\mathbb{Z}^d} f_{\mathbf{k},\mathbf{l}} g_{\mathbf{k},\mathbf{l}}(\mathbf{x}), \tag{4.1}$$

where $\mathbf{k}, \mathbf{l} \in \mathbb{Z}^d$ represent the translation index and modulation index. Note here we use $g_{\mathbf{k},\mathbf{l}}(\mathbf{x})$ to represent the Gabor frame function associated with the oversampling parameters α, β, and $f_{\mathbf{k},\mathbf{l}}$ is a corresponding Gabor coefficient. When the window function g is chosen as a normalized Gaussian function, i.e, $g(\mathbf{x}) = 2^{\frac{d}{4}} \exp\{-\pi\mathbf{x}\cdot\mathbf{x}\}$, the Balian-Low theorem [129] states that $\mathcal{G}(g, \alpha, \beta)$ constitutes a frame iff $\alpha\beta < 1$, and this case is referred as oversampling. In case of a Gaussian window function g, the oversampling parameters α, β determine the rate of decay of both the Gabor window function g and the dual window function η [178, 179]. The convergence of (4.1) is in the norm of $(L^2(\mathbb{R}^d), \|\cdot\|_2)$ and it is unconditional [129, 132].

The Gabor coefficient $f_{\mathbf{k},\mathbf{l}}$ is computed based on the following Gabor transformation

$$f_{\mathbf{k},\mathbf{l}} = \int_{\mathbb{R}^d} f(\mathbf{x}) \eta^*_{\mathbf{k},\mathbf{l}}(\mathbf{x}) d\mathbf{x}, \tag{4.2}$$

where $\eta_{\mathbf{k},\mathbf{l}}(\mathbf{x})$ is the dual frame function based on the translation and modulation operations on the dual window function $\eta(\mathbf{x})$. The choice of dual window function $\eta(\mathbf{x})$ is not unique, and all dual window functions $\eta(\mathbf{x})$ belong to an affine subspace of $L^2(\mathbb{R}^d)$ based on a so-called canonical dual window function $\eta^\circ(\mathbf{x})$ [129]. Several methods exist to approximately compute the canonical dual window function [133]. In the case of oversampling, one important method to compute the canonical dual window functions involves the generalized pseudo inverse of the frame operator, see [132] or [134], where the Zak-transform is used as a main tool. Another method for approximating the dual Gabor window is given in [180], which requires that the Gabor window function g is within the Feichtinger space $S_0 \in L^2$ [181, 182]. In [183], a unified approach to study the invertibility of Gabor frame operators is given for both the continuous case and the discrete case. Different combinations of window and dual window functions can have different decay properties, which allows to make trade-offs between e.g. efficiency and accuracy in practice.

4.2.1 Statement of the problem

From now on we consider Gabor frames in $L^2(\mathbb{R}^2)$. Let $g_{\mathbf{k},\mathbf{l}}(x, y)$ be a 2D Gabor frame function defined as

$$g_{\mathbf{k},\mathbf{l}}(x, y) = g(x - \alpha_x k_x T_x, y - \alpha_y k_y T_y) \exp\{j\beta_x l_x K_x x + j\beta_y l_y K_y y\}, \tag{4.3}$$

where the window function $g(x, y) = 2^{1/2} \exp\{-\pi[(x/T_x)^2 + (y/T_y)^2\}$ is a 2D Gaussian, the integer-valued $\mathbf{k} = (k_x, k_y), \mathbf{l} = (l_x, l_y)$ represent the spatial shift index and the frequency modulation index in x and y directions, respectively. Further more, $\alpha_x = \alpha_y = \beta_x = \beta_y = \sqrt{2/3}$ are oversampling parameters, T_x and T_y denote the Gabor window lengths in the spatial domain and $K_x = 2\pi/T_x$ and $K_y = 2\pi/T_y$ their spectral-domain counterparts.

We are interested in the Gabor coefficients for a 2D indicator function $f(x, y)$ supported on a K-sided polygon D. The indicator function $f(x, y)$ is given by

$$f(x, y) = \begin{cases} 1, & (x, y) \in D \\ 0, & (x, y) \in \mathbb{R}^2 \backslash D \end{cases}, \tag{4.4}$$

and its Gabor coefficients can be computed based on the 2D Gabor transformation:

$$f_{\mathbf{k},\mathbf{l}} = \iint_{\mathbb{R}^2} f(x, y) \eta^*_{\mathbf{k},\mathbf{l}}(x, y) dxdy = \iint_D \eta^*_{\mathbf{k},\mathbf{l}}(x, y) dxdy, \tag{4.5}$$

where the 2D dual frame function is

$$\eta_{\mathbf{k},\mathbf{l}}(x, y) = \eta(x - \alpha_x k_x T_x, y - \alpha_y k_y T_y) \exp\{j\beta_x l_x K_x x + j\beta_y l_y K_y y\}, \tag{4.6}$$

based on a dual window function $\eta(x,y)$. Here we only compute the discrete form of $\eta(x,y)$ based on the pseudo-inverse method [134], and the sampled dual window $\eta(x,y)$ can be represented again by Gabor frames [184]. In Eq. (4.5), we replace the dual frame function $\eta_{\mathbf{k},\mathbf{l}}(x,y)$ by its Gabor frame expansion with frame functions given in Eq. (4.3), and rearrange the integral and summation. Then we identify that Gabor coefficient $f_{\mathbf{k},\mathbf{l}}$ can be written (see [143, 170]) as a linear combination of $I_{\mathbf{m},\mathbf{n}}$ which are integrals of the following form:

$$\begin{aligned} I_{\mathbf{m},\mathbf{n}} = &\iint_D \exp\left\{-\pi\left[\frac{x}{T_x} - \alpha_x m_x\right]^2\right\} \cdot \exp\left\{-j\beta_x n_x K_x x\right\} \\ &\cdot \exp\left\{-\pi\left[\frac{y}{T_y} - \alpha_y m_y\right]^2\right\} \cdot \exp\left\{-j\beta_y n_y K_y y\right\} dxdy \\ = &\oint_{\partial D} -Q(x,y)dx + P(x,y)dy \\ = &\sum_{k=1}^{K} \int_{\partial D_k} -Q(x,y)dx + P(x,y)dy, \end{aligned} \tag{4.7}$$

where Gauss's theorem is applied and ∂D_k represents the kth edge of the polygon boundary ∂D. Functions $P(x,y)$ and $Q(x,y)$ are given by

$$\begin{aligned} P(x,y) =&\tilde{h}_1(y) \cdot \exp\Big\{h_2^2(x)\Big\} \cdot w\Big\{h_2(x)\Big\}, \\ Q(x,y) =&h_1(x) \cdot \exp\Big\{\tilde{h}_2^2(y)\Big\} \cdot w\Big\{\tilde{h}_2(y)\Big\}. \end{aligned} \tag{4.8}$$

where $w(z)$ is the Faddeeva function or the plasma dispersion function [185, 186]. Note that

$$\begin{aligned} h_1(x) =& -\frac{T_y}{4}\exp\Bigg\{-\pi\Big(\frac{x}{T_x} - \alpha_x m_x\Big)^2 - \pi\beta_y^2 n_y^2 \\ &\qquad - j\Big(\beta_x n_x K_x x + 2\pi\alpha_y\beta_y m_y n_y\Big)\Bigg\} \\ h_2(x) =& -\sqrt{\pi}\beta_x n_x + j\sqrt{\pi}\Big(\frac{x}{T_x} - \alpha_x m_x\Big) \end{aligned} \tag{4.9}$$

and one can easily obtain functions $\tilde{h}_1(\cdot)$ and $\tilde{h}_2(\cdot)$ by swapping the subscripts x and y for functions h_1 and h_2. Note that the choices of functions P and Q in Eq. (4.8) are not unique. In [143, 170], $P(x,y)$ and $Q(x,y)$ are represented via the complex error function $\mathrm{erf}(z)$, while here we represent them by the Faddeeva function $w(z)$.

Let $(P_{k,x}, P_{k,y})$ be coordinates of the kth vertex of the polygon D, then we define

$$\begin{aligned} b_{k,x} = \frac{1}{T_x}\left(P_{k+1,x} - P_{k,x}\right), \quad s_{k,x} = \frac{P_{k,x}}{T_x}, \\ b_{k,y} = \frac{1}{T_y}\left(P_{k+1,y} - P_{k,y}\right), \quad s_{k,y} = \frac{P_{k,y}}{T_y}, \end{aligned} \tag{4.10}$$

for $k \in \{1, \ldots, K\}$. Following the analysis of parametrization in [143] and [170], we scale the coordinate variables x_k and y_k of the kth boundary edge with respect to the Gabor window length T_x and T_y, respectively, and then represent them by a parameter t as

$$\begin{cases} \frac{x_k(t)}{T_x} = b_{k,x} \cdot t + s_{k,x}, \\ \frac{y_k(t)}{T_y} = b_{k,y} \cdot t + s_{k,y}, \end{cases} \tag{4.11}$$

where $t \in [0, 1]$. Substituting Eq. (4.11) in Eq. (4.7) we get

$$I_{\mathbf{m},\mathbf{n}} = \sum_k \sum_l c_{k,l,1} \int_0^1 \exp\left\{-c_{k,l,2}t^2 + c_{k,l,3}t\right\} w\left\{c_{k,l,4}t + c_{k,l,5}\right\} dt, \tag{4.12}$$

where $k \in \{1, \ldots, K\}$, and the additional index $l \in \{x, y\}$ means the derivation is associated with the parameterized x-coordinate or y-coordinate in Eq. (4.11). When $l = x$, expressions of all x-related parameters $c_{k,x,i}$, $i \in 1, \ldots, 5$, are given by

$$\begin{aligned} c_{k,x,1} =& (-1)^{k+1}\frac{T_xT_y}{4}b_{k,x} \cdot \exp\Bigg\{ -\pi(s_{k,x}^2 + s_{k,y}^2 + \alpha_x^2 m_x^2 + \alpha_y^2 m_y^2) \\ & + 2\pi(\alpha_x m_x s_{k,x} + \alpha_y m_y s_{k,y}) - 2\pi j(\beta_x n_x s_{k,x} + \beta_y n_y s_{k,y})\Bigg\}, \\ c_{k,x,2} =& \pi\{b_{k,x}^2 + b_{k,y}^2\}, \\ c_{k,x,3} =& -2\pi\big\{b_{k,x}s_{k,x} + b_{k,y}s_{k,y} - \alpha_x m_x b_{k,x} - \alpha_y m_y b_{k,y} \\ & + j(\beta_x n_x b_{k,x} + \beta_y n_y b_{k,y})\big\}, \\ c_{k,x,4} =& j\sqrt{\pi}b_{k,y}, \\ c_{k,x,5} =& -\sqrt{\pi}\beta_y n_y + j\sqrt{\pi}(s_{k,y} - \alpha_y m_y). \end{aligned} \tag{4.13}$$

Corresponding y-related parameters $c_{k,y,i}$, $i \in \{1, \ldots, 5\}$, can be obtained again by swapping the subscripts x and y in (4.13). Note that this procedure only changes $c_{k,x,1}$, $c_{k,x,4}$ and $c_{k,x,5}$, but keeps $c_{k,x,2}$ and $c_{k,x,3}$ the same. Eq. (4.12) shows each Gabor-coefficient-related integral $I_{\mathbf{m},\mathbf{n}}$ of a 2D indicator function supported on a polygon domain is a linear combination of integrals with the same structure, and all integrals are distinguished by five parameters.

For ease of notation, we drop the indices k and l in Eq. (4.12), and state that we want to solve the following problem: Given $c_2 \in \mathbb{R}$ and $c_1, c_3, c_4, c_5 \in \mathbb{C}$, find an efficient method to compute the following integral I accurately

$$I = c_1 \int_0^1 e^{-c_2x^2+c_3x} w(c_4x + c_5)dx. \tag{4.14}$$

4.2.2 Formulation based on a rational expansion of the Faddeeva function

In general it is difficult to find an analytical expression for I in Eq. (4.14), since the integrand contains a complex exponential function and the complex Faddeeva function $w(z)$. However, in some special cases one can find its closed-form expression.

When the pertaining edge ∂D_k of the polygon boundary is parallel to the x axis or to the y axis, we can readily see from Eq. (4.11) that $b_{k,y} \equiv 0$ or $b_{k,x} \equiv 0$, and then we have $c_4 \equiv 0$ or $c_1 \equiv 0$. Therefore

$$I = 0 \quad \text{if} \quad c_1 = 0, \tag{4.15}$$

or, if $c_4 = 0$

$$I = c_1\sqrt{\pi}e^{\frac{c_3^2}{4c_2}}w(c_5) \cdot \left\{2e^{\frac{c_3^2}{4c_2}} - w\left(\frac{jc_3}{2\sqrt{c_2}}\right) - e^{-c_2+c_3}w\left(\frac{j(2c_2-c_3)}{2\sqrt{c_2}}\right)\right\}. \tag{4.16}$$

Equivalent results are given in [143] and [170], based on the complex error function. When a polygon edge is not parallel to any coordinate axis, but the condition $c_3c_4 = 2c_2c_5$ holds, we find that

$$\begin{aligned} I = {} & \frac{c_1\sqrt{\pi}e^{\xi_2}}{2c_4\sqrt{\xi_1}} \cdot \left\{e^{-c_5^2\xi_1}w\left(jc_5\sqrt{\xi_1}\right) - e^{-(c_4+c_5)^2\xi_1}w\left(j(c_4+c_5)\sqrt{\xi_1}\right)\right\} \\ & - \frac{2c_1\sqrt{\pi}e^{\xi_2}}{c_4\sqrt{\xi_1}} \cdot \left\{\mathrm{T}\left(-c_5\sqrt{2\xi_1}, -\frac{j}{\sqrt{\xi_1}}\right) - \mathrm{T}\left(-(c_4+c_5)\sqrt{2\xi_1}, -\frac{j}{\sqrt{\xi_1}}\right)\right\}, \end{aligned} \tag{4.17}$$

where $\xi_1 = 1 + c_2/c_4^2$, $\xi_2 = -c_3c_5/c_4 - c_2c_5^2/c_4^2$, and $\mathrm{T}(x,a)$ is Owen's T function [187].

In all other cases there is no closed-form expression available for the integral I and therefore various numerical integration strategies would be required. However, since the integrand is a product of the complex exponential function $\exp(-c_2x^2 + c_3x)$ and the Faddeeva function $w(c_4x+c_5)$, which change dramatically in certain regions of the complex plane, poor convergence occurs in numerical integration and this causes long computation times [143, 144]. Recalling the relation between $I_{\mathbf{m},\mathbf{n}}$ and the integral type I in Eq. (4.12), it is crucial to find a better method to compute I and get rid of numerical integration completely. To achieve this, our interest goes to an approximation of the Faddeeva function $w(z)$. An ideal approximation should have a fast convergence to the Faddeeva function $w(z)$ and should yield a difference equation so the numerical integration in Eq. (4.14) can be avoided.

In [113], J. A. C. Weideman proposed three rational expansions of the complex Faddeeva function $w(z)$ and the most practical one for us is given by

$$w(z) \approx \frac{1}{\sqrt{\pi}(L-jz)} + 2\sum_{n=0}^{N-1} a_{n+1}\frac{(L+jz)^n}{(L-jz)^{n+2}}, \quad \mathrm{Im}(z) \geq 0, \tag{4.18}$$

where N is a truncation number of the rational expansion series, a_n is an expansion coefficient, and the parameter $L = 2^{-\frac{1}{4}}N^{\frac{1}{2}}$ is chosen to optimize the convergence. This rational expansion (4.18) has the following attractive features:

- High accuracy is achieved uniformly in the complex plane with only a small truncation number N. As shown in [113], the rational expansion approximation (4.18) with $N = 16$ yields a relative error up to 10^{-6} for all z in the entire first quadrant.

- The expansion coefficients a_n can be computed to a high approximation by an FFT, once and for all.

The above promising properties make the rational expansion (4.18) a much better candidate than the Taylor-based approximation to the complex error function $\mathrm{erf}(z)$, as introduced in [173], since the truncation number N in this article is independent of z and the fast convergence saves a summation over thousands of terms that occur in the Taylor-Olver method [143,144]. Assume for all $x \in [0,1]$ that $\mathrm{Im}(c_4x+c_5) \geq 0$, then substitute Eq. (4.18) in Eq. (4.14) and we get

$$\begin{aligned}
I &= c_1 \int_0^1 e^{-c_2x^2+c_3x} w(c_4x+c_5)dx \\
&\approx \frac{c_1}{\sqrt{\pi}} \int_0^1 \frac{e^{-c_2x^2+c_3x}}{L-j(c_4x+c_5)}dx \\
&\quad + 2c_1 \sum_{n=0}^{N-1} a_{n+1} \cdot \int_0^1 e^{-c_2x^2+c_3x} \frac{[L+j(c_4x+c_5)]^n}{[L-j(c_4x+c_5)]^{n+2}}dx \\
&= \frac{c_1j}{\sqrt{\pi}c_4} \int_{\zeta_1}^{\zeta_2} \frac{e^{c_6y^2+c_7y+c_8}}{y}dy \\
&\quad + \frac{2c_1j}{c_4} \sum_{n=0}^{N-1} a_{n+1} \cdot \left\{ \sum_{m=0}^{n} \binom{n}{m} (2L)^m(-1)^{n-m} \int_{\zeta_1}^{\zeta_2} \frac{e^{c_6y^2+c_7y+c_8}}{y^{m+2}}dy \right\},
\end{aligned} \tag{4.19}$$

where we used the variable substitution $y = L - j(c_4x + c_5)$ and applied the binomial theorem in the last step, and we defined

$$\begin{aligned}
& c_6 = \frac{c_2}{c_4^2}, \quad c_7 = -\frac{2c_2L}{c_4^2} + \frac{c_3j}{c_4} + \frac{2c_2c_5j}{c_4^2}, \\
& c_8 = -\frac{c_3c_5}{c_4} - \frac{c_3Lj}{c_4} - \frac{c_2c_5^2}{c_4^2} - \frac{2c_2c_5Lj}{c_4^2} + \frac{c_2L^2}{c_4^2}, \\
& \zeta_1 = L - jc_5, \quad \zeta_2 = L - j(c_4 + c_5).
\end{aligned}$$

Hence, from Eq. (4.19) we see that the problem of computing the integral I is transferred to calculating the following sequence of integrals accurately and efficiently

$$I_m = \frac{c_1}{c_4} \int_{\zeta_1}^{\zeta_2} e^{c_6y^2+c_7y+c_8} \frac{1}{y^m}dy, \quad m = 1, 2, \ldots, N+1. \tag{4.20}$$

4.2.3 March to the whole complex plane

The rational expansion of the Faddeeva function in (4.18) requires $\mathrm{Im}(z) \geq 0$ to avoid the singularity in $\frac{L+jz}{L-jz}$, and therefore the computation of I in Eq. (4.19) holds only for $z = c_4x + c_5$, $\forall\, x \in [0, 1]$ in the upper half-plane. However, $c_4x + c_5$, $x \in [0, 1]$ in Eq. (4.14) can occur in the lower half-plane completely or partially. Now we consider the other cases with $\mathrm{Im}(z) < 0$.

When $\mathrm{Im}(c_4x + c_5) < 0$, we apply the property $w(z) = 2e^{-z^2} - w(-z)$ and get

$$\begin{aligned} I =& c_1 \int_0^1 e^{-c_2x^2+c_3x} w(c_4x + c_5)dx \\ =& c_1 \int_0^1 e^{-c_2x^2+c_3x} \left\{ 2e^{-(c_4x+c_5)^2} - w(-c_4x - c_5) \right\} dx \\ =& \frac{\sqrt{\pi}c_1}{\sqrt{c_2 + c_4^2}} \cdot e^{c_3-c_2-(c_4+c_5)^2} \cdot \left\{ e^{c_2-c_3+c_4^2+2c_4c_5} \cdot w(z_2) - w(z_1) \right\} \\ &- c_1 \int_0^1 e^{-c_2x^2+c_3x} w(-c_4x - c_5)dx, \end{aligned} \tag{4.21}$$

where

$$z_1 = \frac{(2c_2 + 2c_4^2 + 2c_4c_5 - c_3)j}{2\sqrt{c_2 + c_4^2}}, \quad z_2 = \frac{(2c_4c_5 - c_3)j}{2\sqrt{c_2 + c_4^2}},$$

and the remaining integral in the last step of Eq. (4.21) can therefore be computed with (4.19) since $\mathrm{Im}(-c_4x - c_5) \geq 0$.

We conclude this section by considering all possible cases of $\mathrm{Im}(c_4x+c_5)$ when $x \in [0, 1]$. Suppose $\mathrm{Im}(c_4x_0 + c_5) = 0$, then the fundamental integral I in (4.14) has 4 types based on the relation for x_0 and the interval $[0, 1]$:

- Type 1: $x_0 \notin [0, 1]$ and $\mathrm{Im}(c_4x + c_5) \geq 0$ for all $x \in [0, 1]$. This case is discussed in Section 4.2.2, and I can be computed based upon Eq. (4.19).
- Type 2: $x_0 \notin [0, 1]$ and $\mathrm{Im}(c_4x + c_5) \leq 0$ for all $x \in [0, 1]$. As discussed above, I can be computed following the steps in Eq. (4.21).
- Type 3: $x_0 \in [0, 1]$, $\mathrm{Im}(c_4x + c_5) \geq 0$ when $x \in [0, x_0]$ and $\mathrm{Im}(c_4x + c_5) \leq 0$ when $x \in [x_0, 1]$. In this case, I is a combination of two integrals which should be computed via (4.19) on $[0, x_0]$ and via (4.21) on $[x_0, 1]$, respectively.
- Type 4: $x_0 \in [0, 1]$, $\mathrm{Im}(c_4x + c_5) \leq 0$ when $x \in [0, x_0]$ and $\mathrm{Im}(c_4x + c_5) \geq 0$ when $x \in [x_0, 1]$. Analogously, I should be computed via (4.21) on $[0, x_0]$ and via (4.19) on $[x_0, 1]$.

Hence the problem of computing the integral I in (4.14) is transferred to calculating a sequence of integrals I_m in (4.20) for any $c_4x + c_5$ in the complex plane.

4.3 Computation of integrals based on difference equations

Section 4.2 revealed that the computation of the fundamental integral I relies on a sequence of integrals I_m. We now explore a way to compute these integrals that avoids direct numerical integration.

4.3.1 Derivation of the second-order difference equations

Analogous to the derived difference equation in [143], based on the Taylor-series expansion to the complex error function, we expect to derive a difference equation for I_m in (4.20), so numerical integration can be avoided. We apply integration by parts to Eq. (4.20) and get

$$\begin{aligned} &I_m = \frac{c_1}{c_4}\int_{\zeta_1}^{\zeta_2} e^{c_6y^2+c_7y+c_8}\frac{1}{y^m}dy = \\ &\frac{2c_1c_6}{c_4(m-1)}\int_{\zeta_1}^{\zeta_2} e^{c_6y^2+c_7y+c_8}\frac{1}{y^{m-2}}dy + \frac{c_1c_7}{c_4(m-1)}\int_{\zeta_1}^{\zeta_2} e^{c_6y^2+c_7y+c_8}\frac{1}{y^{m-1}}dy \\ &-\frac{c_1}{c_4(m-1)}e^{c_6y^2+c_7y+c_8}\frac{1}{y^{m-1}}\bigg|_{\zeta_1}^{\zeta_2}, \quad m \geq 2 \end{aligned} \tag{4.22}$$

or after rearranging,

$$2c_6I_{m-1} + c_7I_m - mI_{m+1} = \Gamma_m, \quad m \geq 1. \tag{4.23}$$

where $\Gamma_m = \gamma_m(\zeta_2) - \gamma_m(\zeta_1)$ and $\gamma_m(z) = \exp\{c_6z^2 + c_7z + c_8\} \cdot c_1/(c_4z^m)$ are used to represent the inhomogeneous term. It is clear that Eq. (4.23) is a set of second-order linear difference equations and we will have all required I_m in Eq. (4.20) as long as we can solve these difference equations.

Second-order linear difference equations are of great importance and they can be derived from many problems, such as discretization of differential equations and computation of special functions. It is also well known that computing the minimal solution of a second-order difference equation, or three-term recurrence relation employing the forward recurrence is an unstable procedure, see e.g. [146, 147].

In [148], F. W. J. Olver proposed a stable, and by now classic, algorithm for second-order inhomogeneous linear difference equations of the form

$$\alpha_r y_{r-1} - \beta_r y_r + \sigma_r y_{r+1} = \mu_r, \quad r \geq 1 \tag{4.24}$$

where $\alpha_r, \beta_r, \sigma_r, \mu_r$ are given functions of the integer index r, and y_0, y_1 are given initial conditions. The main idea of Olver's algorithm is to transfer the initial value problem to an equivalent boundary value problem (BVP). To do this, Olver abandoned the initial condition y_1 and introduced a boundary condition $y_M = p$ for some sufficiently large M. The value p is determined based on the asymptotic property of the minimal solution y_r and in most cases it is set as $y_M = 0$. Note that this BVP is corresponding to a finite-dimensional linear system and it is an approximation of its counterpart Eq. (4.24), which is an infinite-dimensional one. It was also proven that the solution of the truncated BVP, denoted by $y_r^{(M)}$, convergences to the true solution y_r of Eq. (4.24) as M goes to infinity [148], i.e.

$$\lim_{M\to\infty} y_r^{(M)} = y_r, \quad \text{for } r = 1, 2, \ldots$$

Furthermore, Olver's method has a built-in error estimation technique, which makes it very efficient in practical computation. Overall, Olver's algorithm is a very safe numerical scheme and difficulties will only occur in the most pathological of situations [147].

Following the procedure in Olver's method, we transform Eq. (4.23) into the following tridiagonal system

$$\begin{pmatrix} c_7 & -1 & & & & & \\ 2c_6 & c_7 & -2 & & & & \\ & 2c_6 & c_7 & -3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & 2c_6 & c_7 & -(M-2) \\ & & & & & 2c_6 & c_7 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ \vdots \\ \vdots \\ I_{M-2} \\ I_{M-1} \end{pmatrix} = \begin{pmatrix} \Gamma_1 - 2c_6 I_0 \\ \Gamma_2 \\ \Gamma_3 \\ \vdots \\ \vdots \\ \Gamma_{M-2} \\ \Gamma_{M-1} \end{pmatrix}. \tag{4.25}$$

Note that one can readily calculate the first boundary condition I_0 as

$$I_0 = \frac{c_1}{c_4} \int_{\zeta_1}^{\zeta_2} e^{c_6 y^2 + c_7 y + c_8} dy = \frac{j c_1 \sqrt{\pi}}{2 c_4 \sqrt{c_6}} e^{-\frac{c_7^2}{4c_6} + c_8} \left\{ e^{z_3^2} w(z_3) - e^{z_4^2} w(z_4) \right\}, \tag{4.26}$$

where $z_3 = (c_7 + 2c_6\zeta_1)/(2\sqrt{c_6})$ and $z_4 = (c_7 + 2c_6\zeta_2)/(2\sqrt{c_6})$. The second boundary condition I_M is set to zero based on the asymptotic property of I_m in (4.20).

The tridiagonal system (4.25) can be solved either by Gaussian-elimination without partial pivoting as proposed in [148, 188], or by LU-decomposition [177]. When c_r vanishes for some r in Eq. (4.24), Olver's algorithm needs to be modified by partitioning the tridiagonal system into two parts and attacking them separately [147, 177]. This particular case will never occur in the difference equations we have, since all corresponding c_r in (4.22) are negative integers. In practice we found that using $M = 2N$ is sufficient to ensure an accurate result, where N is the truncation number of the rational expansion of the Faddeeva function in Eq. (4.18). Following Olver's algorithm we get the solution $I_1, I_2, \ldots, I_N, \ldots, I_{2N-1}$, and then the first $N + 1$ terms are assembled into the integral I according to Eq. (4.14).

4.3.2 Algorithm optimization

We optimize the computational procedure of the fundamental integral I and discuss its advantages for implementation.

Let $\mathbf{y} = (I_1, I_2, \ldots, I_{2N-1})^T$ and A be the system matrix in (4.25), then the tridiagonal system can be rewritten into

$$A\mathbf{y} = \mathbf{d}, \tag{4.27}$$

where the vector $\mathbf{d} = \mathbf{d}_1 - \mathbf{d}_2 - \mathbf{d}_3$ contains all inhomogeneous terms and

$$\begin{aligned} \mathbf{d}_1 &= \gamma_1(\zeta_2) \cdot \left(1, 1/\zeta_2, 1/\zeta_2^2, \ldots, 1/\zeta_2^{2N-3}, 1/\zeta_2^{2N-2}\right)^T, \\ \mathbf{d}_2 &= \gamma_1(\zeta_1) \cdot \left(1, 1/\zeta_1, 1/\zeta_1^2, \ldots, 1/\zeta_1^{2N-3}, 1/\zeta_1^{2N-2}\right)^T, \\ \mathbf{d}_3 &= (2c_6 I_0, 0, \ldots, 0)^T. \end{aligned} \tag{4.28}$$

It is obvious that the components of $\mathbf{d}_1$ and $\mathbf{d}_2$ satisfy a two-term difference equation, so in practice $\mathbf{d}_1$ and $\mathbf{d}_2$ can be computed efficiently based on their initial components $\gamma_1(\zeta_2)$ and $\gamma_1(\zeta_1)$.

Let $y_1 = I_1$, $\mathbf{y}_2 = (I_2, \ldots, I_N, I_{N+1})^T$, then we can introduce projection operators $\mathbf{p}^T$ and P by

$$\begin{aligned}\mathbf{p}^T \cdot \mathbf{y} &= y_1,\\ P \cdot \mathbf{y} &= \mathbf{y}_2,\end{aligned}$$

where $\mathbf{p}^T = (1, 0, \ldots, 0)$ and $P \in \mathbb{C}^{N\times(2N-1)}$. In addition, we define the following lower triangular matrix

$$B = \begin{pmatrix} b_{11} & & & & 0 \\ b_{21} & b_{21} & & & \\ b_{31} & b_{32} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ b_{N1} & b_{N2} & \cdots & b_{NN-1} & b_{NN} \end{pmatrix},$$

where its non-zero components are given as $b_{nm} = \binom{n-1}{m-1}(2L)^{m-1}(-1)^{n-m}$ for all $n, m = \{1, 2 \ldots, N\}$. Moreover, we use $\mathbf{w} = (a_1, a_2, \ldots, a_N)^T$ to hold all rational-expansion coefficients of the Faddeeva function $w(z)$. Then we can rewrite Eq. (4.19) as

$$\begin{aligned} I =& \mu_1 y_1 + \mu_2 \mathbf{w}^T B \mathbf{y}_2 \\ =& (\mu_1 \mathbf{p}^T + \mu_2 \mathbf{w}^T B P) \cdot \mathbf{y} \\ =& (\mu_1 \mathbf{p}^T + \mu_2 \mathbf{w}^T B P) \cdot (A^{-1}\mathbf{d}) \end{aligned} \tag{4.29}$$

where $\mu_1 = j/\sqrt{\pi}$ and $\mu_2 = 2j$ are two constants. Here $A^{-1}\mathbf{d}$ is a symbolic representation and this matrix-vector-product is essentially computed following Olver's algorithm. The operators μ_1, μ_2, $\mathbf{p}$, P, B and $\mathbf{w}$ are independent of c_1, c_2, c_3, c_4, c_5 in I, therefore in practice these operators are only required to be computed once and for all in the pre-processing stage.

We conclude this section by summarizing this rational-expansion-based method to compute the integrals $I_{\mathbf{m},\mathbf{n}}$ and therefore the Gabor coefficients as follows:

1. Setting up the tridiagonal matrix A in Eq. (4.25) based on c_1, c_2, c_3, c_4, c_5 given in (4.13).

2. Setting up the inhomogeneous vector $\mathbf{d}$ with the difference equation discussed in Eq. (4.28).

3. Solve the tridiagonal system with Olver's algorithm.

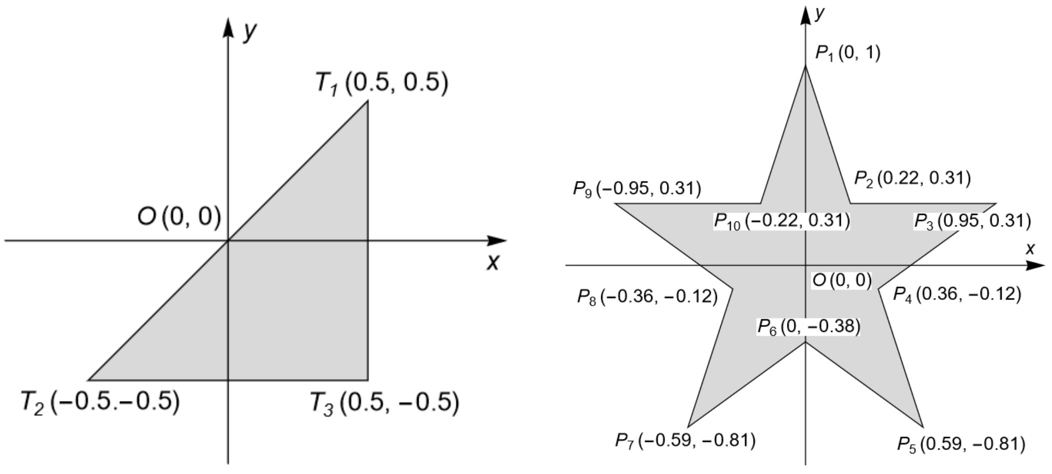
4. Together with pre-processed μ_1, μ_2, $\mathbf{p}$, P, B and $\mathbf{w}$, compute I based on Eq. (4.29).

And we refer to the above steps as the rational-expansion and difference-equation based (RE-DE) algorithm.

4.4 Numerical Examples

In electromagnetic scattering problems with a homogeneous background, the contrast function connects the total electric field and the contrast current density through a field material interaction operation. This operation requires a 2D indicator function that captures the geometry of the scatter. In this section, we compute the Gabor coefficients of two 2D indicator functions supported on different polygonal domains, to test the RE-DE algorithm that we proposed in Sections 4.2 and 4.3.

The first indicator function $f(x, y)$ to be reconstructed is supported on a right-angled triangular domain. We use this example to demonstrate the accuracy and the convergence of the RE-DE algorithm. In the second example the indicator function is supported on a 5-pointed star-shaped polygon domain. Compared to the first example, in this example more fundamental integrals I are required to be computed due to the complexity of the polygon's contour. The supporting domains of these functions are shown in Fig. 4.1 and we use the same Gabor parameters in both examples, see Table 4.1. In both examples we compare the accuracy and computation time of the Gabor coefficients, which are computed by the RE-DE algorithm, with numerical references.



(a) (b)

Figure 4.1: Supporting domains of the indicator function $f(x, y)$: (a) a triangle, (b) a 5-pointed star-shaped polygon, where T_i $(i = 1, 2, 3)$ and P_i $(1 \leq i \leq 10)$ represent the coordinates of the vertices of the triangle and the 5-pointed star-shaped polygon.

Table 4.1: Gabor parameters used to reconstruct the 2D indicator functions $f(x, y)$ supported on the domains in Fig. 4.1.

notation	value	physical meaning
m_x, m_y	$-10:10$	spatial shift index
n_y, n_y	$-10:10$	frequency modulation index
T_x, T_y	1	Gabor window length
K_x, K_y	2π	spectral window length
$\alpha_x, \alpha_y, \beta_x, \beta_y$	$\sqrt{2/3}$	oversampling parameters

Table 4.2 shows the number of integrals evaluated by means of their analytical representations and the integrals calculated by the RE-DE algorithm for both examples. Recalling that the Gabor parameters satisfy $-10 \leq m_x, m_y, n_x, n_y \leq 10$, and that in the case of the triangle example there are three edges and the parametrization is performed with respect to both x and y, we have a total of $21^4 \cdot 3 \cdot 2 =$ 1,166,886 integrals in Example 1. Note that the indicator function $f(x, y)$ is supported on an isosceles right-angled triangle that is aligned such that two of the three edges, i.e. T_2T_3 and T_3T_1, are parallel to one of the coordinate axes. One of $b_{k,x}$ and $b_{k,y}$ in Eq. (4.11) vanishes when the boundary is parallel to either the x or y axis and therefore in total one third of the integrals in Eq. (4.14) is zero. Based on the discussion in Section 4.2.2, another one third of the total integrals are evaluated via Eq. (4.16). A small number of the integrals has an analytical representation in terms of Owen's T function and can be evaluated by Eq. (4.17). The other integrals are calculated based on the RE-DE algorithm. Furthermore, when the fundamental integrals associated with the hypotenuse T_1T_2 are computed by the RE-DE algorithm, one observes that c_4 in Eq. (4.13) is always negative due to the counterclockwise direction of the parametrization in Eq. (4.11). Therefore, only integrals of Types 1, 2 and 3 occur.

Table 4.2: Classification of integrals in Example 1 (triangle) and Example 2 (5-pointed star-shaped polygon). Note that the fundamental integrals in Type i for $i = \{1, 2, 3, 4\}$ are discussed in Section 4.2.3.

Type of integrals	Example 1	Example 2
calculated based on Eq. (4.15)	388,962	388,962
calculated based on Eq. (4.16)	388,962	388,962
calculated based on Eq. (4.17)	882	0
fundamental integral Type 1	184,800	1,528,065
fundamental integral Type 2	184,800	1,528,065
fundamental integral Type 3	18,480	18,522
fundamental integral Type 4	0	37,044
fundamental integrals in all types	388,080	3,111,696
Total	1,166,886	3,889,620

The third column of Table 4.2 shows the classification of the integrals in Example 2. There are 10 edges in this 5-pointed star-shaped polygon example so in total we compute $21^4 \cdot 10 \cdot 2 = 3,889,620$ integrals. Note that in Fig. 4.1 (b) two of the ten edges (boundary segments P_2P_3 and P_9P_{10}) of the 5-pointed star-shaped polygon are parallel to one of the coordinate axes. Analogous to Example 1, one tenth of the integrals vanishes and another one tenth of the integrals can be evaluated via Eq. (4.16). The other integrals are calculated based on the RE-DE algorithm. Compared to the hypotenuse of the triangle in Example 1, the other eight edges of the 5-pointed star-shaped polygon yield both positive and negative values for c_4 in Eq. (4.13), hence all types of integrals discussed in Section 4.2.3 occur.

To compare the accuracy of the fundamental integrals and the Gabor coefficients calculated based on the RE-DE algorithm, we define the following absolute error and relative error. Let N_w be the number of the integrals determined by the discretization parameters in Table 4.1, and $\mathbf{u} \in \mathbb{C}^{N_w}$ contains the integrals calculated based on numerical integration and $\mathbf{v} \in \mathbb{C}^{N_w}$ contains those obtained by the RE-DE algorithm. Then we define the components of the absolute error vector $\mathbf{e}$ as follows

$$e_i = |v_i - u_i|, \quad 1 \leq i \leq N_w \tag{4.30}$$

where u_i, v_i represent the components of $\mathbf{u}$ and $\mathbf{v}$, respectively. The components of the relative error vector $\mathbf{r}$ are defined as

$$r_i = \frac{e_i}{\|\mathbf{u}\|_\infty}, \quad 1 \leq i \leq N_w \tag{4.31}$$

where $\|\mathbf{u}\|_\infty = \max(|u_1|, |u_2|, \ldots, |u_{N_w}|)$ is the ℓ^∞ norm of vector $\mathbf{u}$. We use $\mathbf{r}_{\text{int}}$ and $\mathbf{r}_{\text{gab}}$ to denote the relative error vectors of the fundamental integrals and of the Gabor coefficients, respectively, and we compare the relative errors for different truncation number N in the rational expansion of the Faddeeva function.

We have implemented the RE-DE algorithm in the Wolfram Mathematica language and performed the examples with Mathematica 13.0. All computations were performed on a computer with Intel(R) Core(TM) i7-8850H CPU at 2.60 GHz.

4.4.1 Example 1: a 2D indicator function supported on a triangle

The first example we discuss is an indicator function supported on a triangle. With this example we illustrate the convergence behaviour of the RE-DE algorithm. We compute the fundamental integrals and also the Gabor coefficients of the indicator function based on them. The accuracy of the fundamental integrals and the Gabor coefficients are validated against a numerical reference.

We compute the 388,080 fundamental integrals in Table 4.2 by the RE-DE algorithm with different number of terms in the rational expansions, i.e. $N = 8, 16, 32, 64$. Additionally, we compute all fundamental integrals by numerical integration (NI) with default machine precision in Mathematica, i.e. 16 digits. Then we compare all integrals, which are computed by five different methods, with a reference that is calculated by `NIntegrate` in Mathematica with a 100-digit working precision. Gabor coefficients are computed correspondingly based on the calculated integrals through Eq. (4.12). Absolute error and relative error vectors are computed based on Eq. (4.30) and Eq. (4.31). Table 4.3 shows the ℓ^2 norm and the ℓ^∞ norm of $\mathbf{r}_{\text{int}}$ and $\mathbf{r}_{\text{gab}}$. Note that the ℓ^2 norm of a vector $\mathbf{v}$ is defined as $\|\mathbf{v}\|_2 = (\sum_i |v_i|^2)^{\frac{1}{2}}$ and the inequality $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2$ holds.

Table 4.3: Norm comparison of relative error vectors $\mathbf{r}_{\text{int}}$ and $\mathbf{r}_{\text{gab}}$ in Example 1. Columns $2-5$ correspond to the cases with different truncation numbers N in the rational expansion of the Faddeeva function. The last column corresponds to the numerical integration with machine precision in Mathematica.

Norm	$N = 8$	$N = 16$	$N = 32$	$N = 64$	NI
$\|\mathbf{r}_{\text{int}}\|_\infty$	1.7×10^{-5}	8.3×10^{-8}	1.7×10^{-11}	3.5×10^{-16}	8.1×10^{-12}
$\|\mathbf{r}_{\text{int}}\|_2$	1.4×10^{-4}	6.6×10^{-7}	3.6×10^{-11}	1.6×10^{-15}	2.3×10^{-11}
$\|\mathbf{r}_{\text{gab}}\|_\infty$	2.6×10^{-4}	1.9×10^{-7}	1.9×10^{-13}	8.6×10^{-15}	2.3×10^{-10}
$\|\mathbf{r}_{\text{gab}}\|_2$	8.0×10^{-4}	5.7×10^{-7}	5.7×10^{-13}	2.5×10^{-14}	7.0×10^{-10}

Based on Table 4.3 we make the following observations:

- There is a clear convergent behaviour of the RE-DE algorithm. Both ℓ^2 norm and ℓ^∞ norm of fundamental integral and Gabor coefficient are reduced when increasing the number of terms N in the rational expansion from 8 to 64. This also corresponds to the convergence property of the rational expansions for the Faddeeva function proven in [113].

- The norms of $\mathbf{r}_{\text{gab}}$ are slightly larger than their counterparts for $\mathbf{r}_{\text{int}}$. This is as expected since each Gabor coefficient is a summation of a set of fundamental integrals according to Eq. (4.12).

In order to have a better view on the distribution of the relative errors for the fundamental integrals, we calculate the percentages of the relative errors within specific intervals and show the results in Table 4.4. Here we notice again that the largest relative error of all computed fundamental integrals is pushed to a lower level when increasing the rational-expansion parameter N, as observed in Table 4.3. On the other hand, we expect to a similar convergence behaviour on the computed Gabor coefficients. Analogous to the behavior of a cumulative distribution function, Fig. 4.2 shows the distributions of relative errors of all Gabor coefficients computed in Example 1. Here we see the changes of the percentage of computed Gabor coefficients (along the vertical direction) that have relative error less than a given threshold (along the horizontal direction).

Table 4.4: Distribution of relative errors $\mathbf{r}_{\text{int}}$ of the fundamental integrals in Example 1.

relative error interval	$N = 8$	$N = 16$	NI	$N = 32$	$N = 64$
$(10^{-4}, 10^{-1}]$	0	0	0	0	0
$(10^{-7}, 10^{-4}]$	0.07%	0	0	0	0
$(10^{-10}, 10^{-7}]$	0.15%	0.13%	0	0	0
$(10^{-13}, 10^{-10}]$	0.15%	0.11%	0.031%	0.0062%	0
$(10^{-16}, 10^{-13}]$	0.22%	0.18%	0.15%	0.11%	0.021%
$[0, 10^{-16}]$	99.41%	99.59%	99.82%	99.89%	99.98%

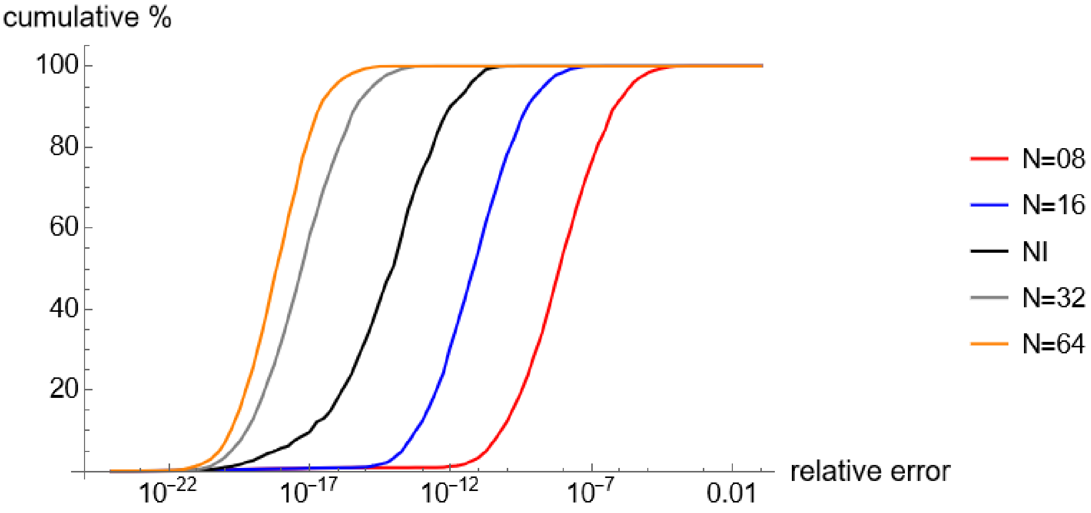


Figure 4.2: Distribution of relative errors $\mathbf{r}_{\text{gab}}$ of the Gabor coefficients in Example 1. Horizontal axis: threshold relative error on a log scale. Vertical axis: percentage of Gabor coefficients whose relative error is less than the threshold.

Both Table 4.4 and Fig. 4.2 suggest a convergent behaviour when increasing the rational-expansion parameter N in the RE-DE algorithm and it is clear that the cases with $N = 32$ and $N = 64$ yield a higher accuracy than the direct numerical integration method.

Next, we investigate the computation time of the RE-DE algorithm. The time spent on initializing the constant operators in Eq. (4.29) is negligible, therefore it is more interesting to compare the time spent on the fundamental integrals only, which is the dominant part during the computing of all Gabor coefficients. Table 4.5 shows the total computation time spent on the calculation of the 388,080 integrals for the different cases.

Table 4.5: Computation time of computing the fundamental integrals in Example 1.

	$N = 8$	$N = 16$	$N = 32$	$N = 64$	NI
Time used [min]	24.9	29.5	50.0	109.2	273.2

Clearly, the RE-DE algorithm significantly outperforms the numerical integration in terms of computation time. Even the case with the largest truncation number $N = 64$ in the rational expansions of the Faddeeva function, the total computation time is less than 40% of that for the numerical integration.

Fig. 4.3 shows the reconstructed indicator function $f(x, y)$ based on the Gabor coefficients computed with the RE-DE algorithm ($N = 16$). Note that there is a clear Gibbs phenomenon [2] close to the boundary of the triangle.

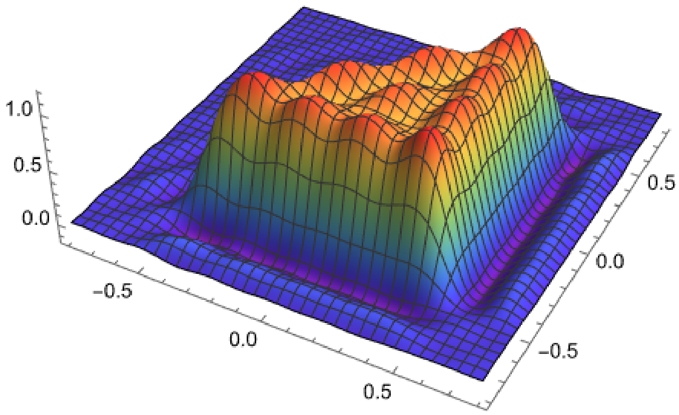


Figure 4.3: Reconstructed indicator function $f(x, y)$ of a triangular domain based on Gabor coefficients computed by the RE-DE algorithm.

4.4.2 Example 2: a 2D indicator function supported on a star-shaped polygon

In the second example we discuss the indicator function supported on a 5-pointed star-shaped polygon. Again we compute the Gabor coefficients, validate their accuracy, and compare the computation time.

In Example 2 there are 3,111,696 fundamental integrals in the form of Eq. (4.14) that are computed by the proposed RE-DE algorithm. Computing a numerical reference for those integrals with 100-digit working precision takes an enormous amount of computation

[2] When approximating a discontinuous non-periodic function with the Fourier series, an overshoot or undershoot effect occurs around the discontinuity. This effect is called Gibbs phenomenon, which was explained by J. W. Gibbs in 1899 [189]. A series of papers about Gibbs phenomenon and reconstruction with exponential accuracy can be found at [190–194]. Due to the close connection between Gabor transformation and Fourier transformation, Gibbs phenomenon occurs here as well.

time (more than 2,600 hours on the current computation hardware), so in this 5-pointed star-shaped polygon example we take the numerical integration with machine precision in Wolfram Mathematica as reference and test the RE-DE algorithm with the rational-expansion parameter $N = 8, 16, 32$. To obtain the global distribution of the relative errors in the Gabor coefficients based on the RE-DE algorithm, we count the number of relative errors located in specific intervals and show the percentage in Table 4.6. Here one observes again the convergent behaviour of the RE-DE algorithm by increasing the truncation number N of the rational expansion of the Faddeeva function.

Table 4.6: Distribution of relative errors $\mathbf{r}_{\text{gab}}$ of the Gabor coefficients in Example 2.

relative error interval	$N = 8$	$N = 16$	$N = 32$
$(10^{-2}, 10^{-1}]$	0.007%	0	0
$(10^{-3}, 10^{-2}]$	0.423%	0	0
$(10^{-4}, 10^{-3}]$	2.82%	0	0
$(10^{-5}, 10^{-4}]$	7.84%	0.05%	0
$(10^{-6}, 10^{-5}]$	15.6%	1.58%	0
$(10^{-10}, 10^{-6}]$	70.6%	61.44%	24.4%
$(10^{-16}, 10^{-10}]$	2.71%	36.93%	75.6%

In Table 4.7 we show the computation time spent on the fundamental integrals with RE-DE algorithm and on the numerical reference. It is clear that the RE-DE algorithm requires much less computation time than the numerical integration method. Even for the case with $N = 32$, which has better accuracy than the reference in Table 4.4 and Fig. 4.2, it takes less than 10% of the time spent for numerical integration. A few comments about the computation time are in order:

- The rational expansion coefficients a_n in Eq. (4.19) are computed via an FFT. This property is inherited from [113].

- The constant operators in Eq. (4.29) are independent of the fundamental integrals, therefore they are computed once during the initialization.

- Numerical integration is avoided in the RE-DE algorithm. The closed form of the boundary condition I_0 is given in Eq. (4.26), and all the other integrals I_m can be computed via Olver's algorithm.

- Solving the difference equations (4.23) with Olver's algorithm is a fast procedure, since Eq. (4.25) is a low-dimensional system, i.e. $(2N - 1) \times (2N - 1)$, and its inhomogeneous term can be generated quickly based on another difference equation as discussed in Eq. (4.28).

Table 4.7: Computation time of computing the fundamental integrals in Example 2.

	$N = 8$	$N = 16$	$N = 32$	reference (NI)
Time [h]	3.30	5.02	7.17	79.87

Finally we show the reconstructed function $f(x, y)$ based on the RE-DE algorithm with $N = 16$ in Fig. 4.4.

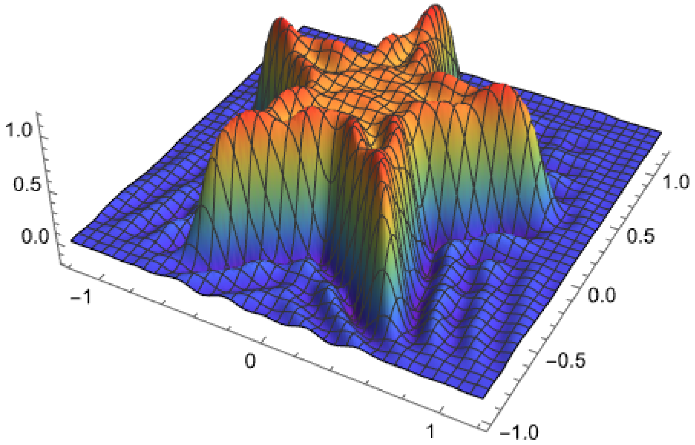


Figure 4.4: Reconstructed indicator function $f(x, y)$ of a 5-pointed star-shaped polygonal domain, based on Gabor coefficients computed by the RE-DE algorithm. The Gibbs phenomenon is also observed around the boundary of the polygon.

In summary, these examples show that the RE-DE algorithm reaches an accuracy that is compatible with or higher than the accuracy of the numerical integration, by choosing a proper truncation number N. Additionally, it significantly outperforms numerical integration in terms of computation time.

4.5 Conclusion

We have proposed a rational-expansion-based method to compute Gabor coefficients of a 2D indicator function supported on a polygon. The method is based on a rational expansion of the Faddeeva function and inherits its fast-convergence property on the whole complex plane. A sequence of integrals with a uniform structure was formulated first and we derived a second-order difference equation after applying the rational expansion of the Faddeeva function. We solved all integrals with Olver's algorithm and optimized this procedure to make it more suitable for implementation. This rational-expansion-based method does not require numerical quadrature.

We tested this method on two numerical examples by computing 3.5 million integrals in total. Numerical results show that this rational-expansion-based method significantly outperforms numerical quadrature methods in terms of computation time, while maintaining accuracy.

Acknowledgments

This work was funded by NWO-TTW as part of the HTSM program under project number 16184. L. Sun is grateful to J. A. C. Weideman from Stellenbosch University for his support. The authors also thank H. G. Feichtinger for his contributions to the contents of this paper.

Chapter 5

Improved Gabor-based multiplication operators for the spatial spectral method

5.1 Introduction

Fourier factorization rules [106, 195, 196] constitute a framework to compute an approximation to the Fourier coefficients of a product function $h(x) = f(x)u(x)$, based on the Fourier coefficients of $f(x)$ and $u(x)$. Laurent's rule of factorization assumes that the truncated Fourier coefficients of $h(x)$ can be computed approximately through a truncated discrete convolution. However, it has been shown that Laurent's rule is valid only when the functions $f(x)$ and $u(x)$ do not have concurrent discontinuities, otherwise convergence issues occur in the computed Fourier coefficients of $h(x)$. When the functions $f(x)$ and $g(x)$ have concurrent but complementary discontinuities, the inverse rule can be used to compute the Fourier coefficients of $h(x)$ for more accuracy. When the functions $f(x)$ and $u(x)$ have concurrent and non-complementary discontinuities, neither Laurent's rule nor the inverse rule should be used. These factorization rules were discovered empirically in [125, 126], and subsequently systemically formulated by Li [106]. Following the Fourier factorization rules given above, the normal vector field (NVF) formulation has been introduced in [105] to overcome the convergence issues in the electric field that result from discontinuities in the permittivity function in the transverse plane. This was further developed in [103, 107, 123, 127] for various computational frameworks.

Gabor analysis (also known as the short-time Fourier transform) is one important branch in time-frequency analysis, which is concerned with localized Fourier transforms [130]. The spatial spectral method developed in [108–111, 197] uses Gabor frames to perform discretization in the transverse directions, such that an efficient transformation between the spatial domain and the spectral domain can be established. Since Gabor frames inherit many of the properties of Fourier series, the problem to accurately multiply two functions represented by Gabor coefficients is similar to that of multiplying two Fourier series. One aspect of this problem was previously addressed by combining the spatial spectral method with the NVF formulation, such that the total electric field **E** and the contrast current density **J** are represented indirectly by a continuous auxiliary field **F**. Hence the auxiliary field and NVF-related operators $\mathcal{C}_\varepsilon$ and $\chi\mathcal{C}_\varepsilon$, which are defined in Eq. (2.56) and (2.57), are represented by their Gabor coefficients once the Gabor frame is set up. Next of the use of the NVF framework, a stable and accurate multiplication between two sets of Gabor coefficients to obtain the Gabor coefficients of the product function is still required to construct an accurate representation of the multiplication operators $\mathcal{C}_\varepsilon$ and $\chi\mathcal{C}_\varepsilon$, such that an iterative solver can be used to obtain the solution to the pertaining scattering problem.

There are no explicit factorization rules known in Gabor analysis. In [197], a multiplication operator for Gabor coefficients is introduced. Even though part of the matrix representation of the multiplication operator is based on Laurent's factorization rule (therefore it can be computed efficiently using FFTs), the overall matrix representation of this operator does not have a Toeplitz structure anymore and therefore the direct connection with the Fourier factorization rules is lost. Nevertheless, reliable results with the spatial spectral method have been widely observed, e.g. the cases presented in [108–111]. However, in recent numerical experiments, we have observed unreliable numerical results for high-contrast scattering problems, where the discontinuities in the operators $\mathcal{C}_\varepsilon$ and $\chi\mathcal{C}_\varepsilon$ become large. Clearly, it is a vital task to construct an accurate and stable multiplication operator for Gabor coefficients such that the spatial spectral Maxwell solver produces reliable results for high-contrast scattering problems. Introducing such an improved multiplication operator for two sets of Gabor coefficients is the main goal of this chapter.

In Section 5.2 we recall the original multiplication operator for two sets of Gabor coefficients, as used in the spatial spectral method [24], and show that the consequent approximation of a product function becomes inaccurate and the corresponding matrix is ill-conditioned for a high-contrast example. In addition, we propose an improved multiplication operator for two sets of Gabor coefficients and show that both the approximation property and the conditioning of the corresponding matrix exhibit better results for the same example. In Section 5.3, the improved multiplication operation is extended to the Gabor-based equidistantly sampled lists, which were invented to reduce computation time in the spatial spectral method. We then show the same behavior for the two multiplication analogs on sampled list, i.e. the advantages of using the improved multiplication under a Gabor-based list representation. In Section 5.4, we further compare the performances of the improved multiplication operator as compared with the original multiplication operator, by testing them on two examples. These two examples correspond to the second type and the third type, according to the Fourier factorization rules [106]. To show the significance of this improved multiplication on the spatial spectral Maxwell solver, we study a 3D problem with a high-contrast scatterer in Section 5.6. Both near-field and far-field solutions are compared against an independent reference and we show that a convergent trend is obtained by refining the discretization parameters. We summarize the outcomes of this chapter in Section 5.6.

5.2 Multiplication of two sets of Gabor coefficients

The Gabor-frame-based spatial spectral Maxwell solver [24] was developed according to the following domain integral representation:

$$\mathbf{E}^i(\mathbf{x}) = \mathcal{C}_\varepsilon(\mathbf{x}) \cdot \mathbf{F}(\mathbf{x}) - \mathcal{F}_T^{-1}\left\{\int_{\mathbb{R}} G(\mathbf{k}_T, z, z') \cdot \mathcal{F}_T\{\chi\mathcal{C}_\varepsilon(\mathbf{x}_T, z') \cdot \mathbf{F}(\mathbf{x}_T, z')\}dz'\right\}, \tag{5.1}$$

where the electric field **E** and the contrast current density **J** are computed through:

$$\mathbf{E} = \mathcal{C}_\varepsilon \cdot \mathbf{F}, \tag{5.2}$$

$$\mathbf{J} = \chi\mathcal{C}_\varepsilon \cdot \mathbf{F}, \tag{5.3}$$

and the expressions of the operators $\mathcal{C}_\varepsilon$ and $\chi\mathcal{C}_\varepsilon$ are given in Section 2.3.1.

Multiplication operations in the transverse plane are involved in Eqs (5.2) and (5.3). Depending on the function representations to be multiplied, there are two multiplication operators defined in the spatial spectral method. The first multiplication operator is defined between two sets of Gabor coefficients and it is used to solve 2D TE or TM scattering problems, see [109,110,197]. The second multiplication operator was introduced in [111] for multiplying two uniformly sampled lists, which yields a faster computation in 3D scattering problems together with Fourier transformations. Next, we recall the definitions of these two multiplication operators.

Given a function $f(x) \in L^2(\mathbb{R})$, we define the following discrete sampled function

$$\overline{\mathbf{f}} = \mathcal{S} \circ f(x) = \{f(l\Delta_x), l = -L, \ldots, L\}, \tag{5.4}$$

where $\mathcal{S}$ represents a sampling operation associated with the spacing $\Delta_x \in \mathbb{R}^+$, and L is an integer. Note that $\overline{\mathbf{f}}$ is in boldface since it is a vector in $\mathbb{C}^{2L+1}$ and we use the overline to indicate that the vector components represent sampled function values.

The fast Gabor transformation described in [134] is used in the spatial spectral method [24], and the spacing Δ_x is defined by the Gabor parameters through

$$\Delta_x = \frac{T_x}{\alpha_x(2N+1)}, \tag{5.5}$$

where T_x is the spatial Gabor window length, α_x is the oversampling parameter, and $\{-N, \ldots, N\}$ is the range for modulation index n_x. When $f(x)$ is sampled with the Δ_x defined in Eq. (5.5), the pertaining sampled function $\overline{\mathbf{f}}$ is equivalent to the equidistant sampled list described in [111]. Furthermore, we can define an approximately double-sampled function as

$$\overline{\mathbf{f}}' = \mathcal{S}_2 \circ f(x) = \{f(l\Delta_x'), l = -L, \ldots, L\} \tag{5.6}$$

where $\mathcal{S}_2$ is a sampling operator associated with the spacing distance $\Delta_x' = T_x/[\alpha_x(4N+1)]$.

The spatial spectral Maxwell solver is closely connected with the Gabor frames. The finite Gabor-frame expansion of a function $f(x) \in L^2(\mathbb{R})$ and the finite Gabor transformation are given as

$$f^{(M,N)}(x) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} f_{m,n} g_{m,n}(x), \tag{5.7}$$

$$f_{m,n} = \int_{\mathbb{R}} f(x) \eta^*_{m,n}(x) dx, \tag{5.8}$$

where $g_{m,n}(x)$ is a Gabor frame function and $f_{m,n}$ is a corresponding Gabor coefficient computed based on the dual frame function $\eta(x)$. The calculation of Gabor coefficients for 2D indication functions has been discussed in Chapters 3 and 4. We use the notation $\underline{\mathbf{f}}$ to indicate a vector whose components are the Gabor coefficients $f_{m,n}$. Following the analysis in [134], one can rewrite Eq. (5.7) and Eq. (5.8) into the following matrix-vector-product form:

$$\underline{\mathbf{f}} = G \cdot \overline{\mathbf{f}}, \tag{5.9}$$

$$\overline{\mathbf{f}} = G^{-1} \cdot \underline{\mathbf{f}}, \tag{5.10}$$

where $G \in \mathbb{C}^{S \times (2L+1)}$ is the matrix representation of the finite Gabor transformation, $S = (2M+1)(2N+1)$, and G^{-1} is its generalized (Moore-Penrose) inverse[1]. Analogously, we can obtain the vector $\underline{\mathbf{f}}' \in \mathbb{C}^{S'}$ which contains the Gabor coefficients of the double-sampled function $\overline{\mathbf{f}}'$, and $S' = (2M+1)(4N+1)$.

To represent the connection between the two sets of Gabor coefficients $\underline{\mathbf{f}}$ and $\underline{\mathbf{f}}'$ efficiently, we assume that the bookkeeping in the vectors $\underline{\mathbf{f}}$ and $\underline{\mathbf{f}}'$ is such that the index n is the fast-running index and m is the slow-running index, i.e. the element numbering follows the pattern $m(2N+1)+n$. We define a restriction operation $\underline{U}$ through the following block diagonal matrix

$$\underline{U} = \begin{pmatrix} U_1 & & & \\ & U_2 & & \\ & & \ddots & \\ & & & U_{2M+1} \end{pmatrix}, \tag{5.11}$$

where all the submatrices have the same structure

$$U_m = [0_{2N+1,N} \; I_{2N+1} \; 0_{2N+1,N}], \tag{5.12}$$

for all $m = 1, 2, \ldots, (2M+1)$. Additionally, the matrix block I_{2N+1} is an identity matrix of dimension $(2N+1) \times (2N+1)$, and $0_{2N+1,N}$ is a zero matrix of dimension $(2N+1) \times N$. Therefore, the dimension of the submatrix U_m is $(2N+1) \times (4N+1)$, the dimension of U is $(2M+1)(2N+1) \times (2M+1)(4N+1)$ or $S \times S'$.

[1] The same transformations are denoted by $\mathcal{B}$ and $\mathcal{B}^{-1}$ in [111].

On the other side, we define an extension operation $\underline{P}$ as the following block diagonal matrix

$$\underline{P} = \begin{pmatrix} P_1 & & & \\ & P_2 & & \\ & & \ddots & \\ & & & P_{2M+1} \end{pmatrix}, \tag{5.13}$$

where the submatrix $P_m = U_m^T$ and the superscript T stands for matrix transpose, for all $m = 1 \dots (2M+1)$. Clearly, $P_m \in \mathbb{C}^{(4N+1)\times(2N+1)}$, and $\underline{P}$ is a zero-padding matrix from $\mathbb{C}^S$ to $\mathbb{C}^{S'}$.

With the restriction and extension operators introduced above, we have $\underline{\mathbf{f}} = \underline{U} \cdot \underline{\mathbf{f}}'$. Unfortunately in the other direction, $\underline{\mathbf{f}}' \neq \underline{P} \cdot \underline{\mathbf{f}}$ in most cases. However, there is an approximate relation $\overline{\mathbf{f}}' \approx G^{-1} \cdot \underline{P} \cdot \overline{\mathbf{f}}$, since $G^{-1} \cdot \underline{P} \cdot \overline{\mathbf{f}}$ can be considered as a lower-frequency approximation to $\overline{\mathbf{f}}'$.

From now on, we define a product function $h(x) = f(x)u(x)$ and study how to obtain the approximated Gabor coefficients of $h(x)$ based on the two sets of Gabor coefficients of $f(x)$ and $u(x)$. In other words, we want to compute $\underline{\mathbf{h}}$ based on $\underline{\mathbf{f}}$ and $\underline{\mathbf{u}}$. The finite Gabor-frame expansions of $f(x)$ and $u(x)$ are given as:

$$f^{(M,N)}(x) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} f_{m,n} g_{m,n}(x), \tag{5.14}$$

$$u^{(M,N)}(x) = \sum_{m'=-M}^{M} \sum_{n'=-N}^{N} u_{m',n'} g_{m',n'}(x). \tag{5.15}$$

Assume the finite Gabor-frame expansion of $h(x)$ is

$$h^{(M,N)}(x) = \sum_{m''=-M}^{M} \sum_{n''=-N}^{N} h_{m'',n''} g_{m'',n''}(x), \tag{5.16}$$

then we can compute the Gabor coefficient of $h(x)$ according to Eq. (5.8) and get:

$$\begin{aligned} h_{m'',n''} &= \int_{\mathbb{R}} h(x) \eta^*_{m'',n''}(x) dx \\ &= \int_{\mathbb{R}} \left(\sum_{m,n} f_{m,n} g_{m,n} \right) \cdot \left(\sum_{m',n'} u_{m',n'} g_{m',n'} \right) \eta^*_{m'',n''}(x) dx \\ &= \int_{\mathbb{R}} \sum_{m,n} \sum_{m',n'} f_{m,n} u_{m',n'} g_{m,n}(x) g_{m',n'}(x) \eta^*_{m'',n''}(x) dx \\ &= \sum_{m,n} \sum_{m',n'} f_{m,n} u_{m',n'} \int_{\mathbb{R}} g_{m,n}(x) g_{m',n'}(x) \eta^*_{m'',n''}(x) dx \\ &= \sum_{m,n} \sum_{m',n'} a_{m,n} b_{m',n'} \cdot I_{m,n,m',n',m'',n''}, \end{aligned} \tag{5.17}$$

where the summation ranges $-M \le m, m' \le M$, $-N \le n, n' \le N$ are omitted, and we introduce a notation $I_{m,n,m',n',m'',n''}$ in the last step. Eq. (5.12) gives a way to compute the Gabor coefficients of a product function, which is also described in [197]. We summarize the above operations into the following definition of a multiplication operator.

Definition 1. *Let* $\underline{\mathbf{f}}$ *and* $\underline{\mathbf{u}}$ *be the Gabor coefficients of* $f(x)$ *and* $u(x)$ *under the finite Gabor-frame expansion in Eq.* (5.14) *and Eq.* (5.16). *The multiplication operator associated with* $\underline{\mathbf{f}}$ *is defined as*

$$\underline{m}(\underline{\mathbf{u}}): \ \mathbb{C}^S \to \mathbb{C}^S, \quad S = (2M+1)(2N+1), \tag{5.18}$$

where the components of $\underline{m}(\underline{\mathbf{u}})$ *are given by Eq.* (5.17).

Definition 1 implies that $\underline{\mathbf{f}}$ is a given parameter associated with the multiplication operator $\underline{m}$, and $\underline{\mathbf{u}}$ is the input to be multiplied. By recalling the formulation in [108], one can easily recognize the multiplication of the operator $\mathcal{C}_\varepsilon$ and the auxiliary field $\mathbf{F}$ is a typical example of the operator $\underline{m}$. The multiplication operator $\underline{m}$ in Definition 1 has been proven to work well for relatively low-contrast scattering problems [108–111]. However, when the contrast of a scatterer becomes high, we encounter an ill-conditioned linear system for the spatial spectral Maxwell solver, which is therefore generating difficulties in the iterative solution process to find a solution for the linear system. Furthermore, even for those cases where one can obtain a solution after a large number of iterations, the obtained "solution" does not coincide with the reference obtained via an independent Maxwell solver. To show this ill-conditioning issue, we consider the following example.

Let χ be a contrast function supported on the interval $[-50, 50]$ nm defined as

$$\chi(x) = \begin{cases} 100, & -50 \cdot 10^{-9} \le x \le 50 \cdot 10^{-9}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.19}$$

We consider the following multiplication problem with Gabor coefficients:

$$h(x) = f(x) \cdot u(x), \tag{5.20}$$

where

$$f(x) = \frac{1}{1+\chi(x)} \tag{5.21}$$

and $u(x)$ is an arbitrary function. The Gabor parameters used in this example are given in Table. 5.1, where T_x is the Gabor window length, m_x and n_x are the spatial shift index and the frequency modulation index and they are restricted to $-M \le m_x \le M$, $-N \le n_x \le N$. α_x and β_x are the oversampling parameters.

Table 5.1: Gabor parameters used to analyze and reconstruct Eq. (5.14).

T_x	M	N	α_x	β_x
$5 \cdot 10^{-8}$	5	7	$\sqrt{2/3}$	$\sqrt{2/3}$

Let $\underline{A}_{\text{org}}$ be the matrix representation of the multiplication operator $\underline{m}$ associated with the function $f(x)$ in Eq. (5.14), and the matrix $\underline{A}_{\text{org}}$ is constructed by

$$\underline{A}_{:,i} = \underline{m}(\underline{\mathbf{u}}_{\mathbf{i}}), \quad 1 \leq i \leq S, \tag{5.22}$$

where $\underline{A}_{:,i}$ presents the i-th column of the matrix $\underline{A}_{\text{org}}$, $\underline{\mathbf{u}}_{\mathbf{i}}$ is the i-th unit vector, i.e. the vector with the i-th component being 1 and the others being 0. Clearly, $\underline{A}_{\text{org}} \in \mathbb{C}^{S\times S}$, and the subscript represents the matrix corresponding to the original multiplication operator $\underline{m}$ in Definition 1.

To study the conditioning of $\underline{A}_{\text{org}}$, we consider the following linear system

$$\underline{A}_{\text{org}} \cdot \mathbf{x} = \underline{\mathbf{b}}_{\text{random}}, \tag{5.23}$$

where $\underline{\mathbf{b}}_{\text{random}}$ is a random vector in $\mathbb{C}^S$ that represents the outcome of the multiplication between the known function $f(x)$ in (5.20) and a yet unknown function $u(x)$ represented by $\mathbf{x}$. Solving this linear system thus implies an inversion on the multiplication operator $\underline{m}$ associated with the function $f(x)$. Since $f(x)$ is a positive function, this inversion should be a well-defined operation. We then use the BI-CGSTAB iterative method to solve Eq. (5.17) and plot the iterative details in Fig. 5.1 in red. From Fig. 5.1 it is clear that more than 800 iterations are required to reach a relative error below 10^{-5}. This large number of iterations indicates that the matrix $\underline{A}_{\text{org}}$ of the original multiplication operator $\underline{m}$ is ill-conditioned.

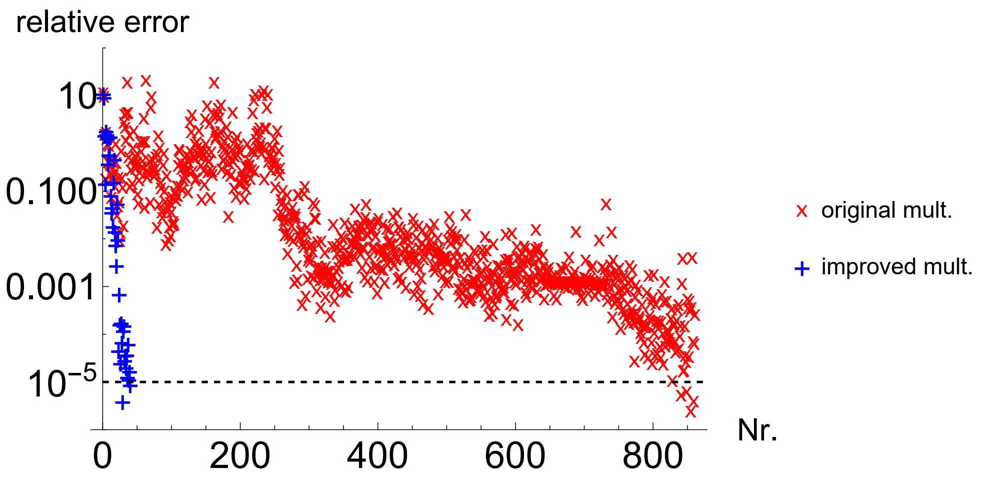


Figure 5.1: Iterative details based on the inversion of the multiplication operators $\underline{A}_{\text{org}}$ and $\underline{A}_{\text{imp}}$ for Gabor coefficients. The label "original mult." refers to the original multiplication operator $\underline{m}$ in Definition 1, and the label "improved mult." refers to the improved multiplication operator $\underline{\underline{m}}$ in Definition 2.

To address this ill-conditioning issue, we define another multiplication operator in Definition 2.

Definition 2. *Let $\underline{\mathbf{f}}'$ be the Gabor coefficients of a double-sampled function $\overline{\mathbf{f}}'$, and $\underline{\mathbf{u}}$ be the Gabor coefficients of $\overline{\mathbf{u}}$. Then the modified multiplication operator associated with $\underline{\mathbf{f}}'$ is defined as*

$$\underline{\underline{m}}(\underline{\mathbf{u}}) = \underline{U} \cdot \underline{m}(\underline{P} \cdot \underline{\mathbf{u}}). \tag{5.24}$$

where $\underline{\mathbf{f}}'$ is used as a parameter in $\underline{m}$. The operators $\underline{U}$ and $\underline{P}$ that represent a restriction operation and an extension operation to Gabor coefficients are given in (5.11) *and* (5.12)*.*

The operator $\underline{\underline{m}}$ in Definition 2 is based on the original multiplication operator $\underline{m}$ but with extra operations. First, a set of refined Gabor coefficients $\underline{\mathbf{f}}'$ is needed and this can be done with the double-sampling operator $\mathcal{S}_2$ and the finite Gabor transformation operator G. Second, the input to be multiplied must be padded to make sure two inputs have the same dimension, before using the multiplication of Eq. (5.17) for n and n' running from $-2N$ to $2N$, and this step is done with the extension operator $\underline{P}$ for Gabor coefficients. Third, an extra restriction operation on the Gabor coefficient is needed to make sure the output has the same dimension as the input.

The matrix representation of $\underline{\underline{m}}$, which is denoted by $\underline{\underline{A}}_{\text{imp}}$, can be obtained by following an analogous procedure as in Eq. (5.22). We then solve the following linear system by using the BI-CGSTAB iterative method again:

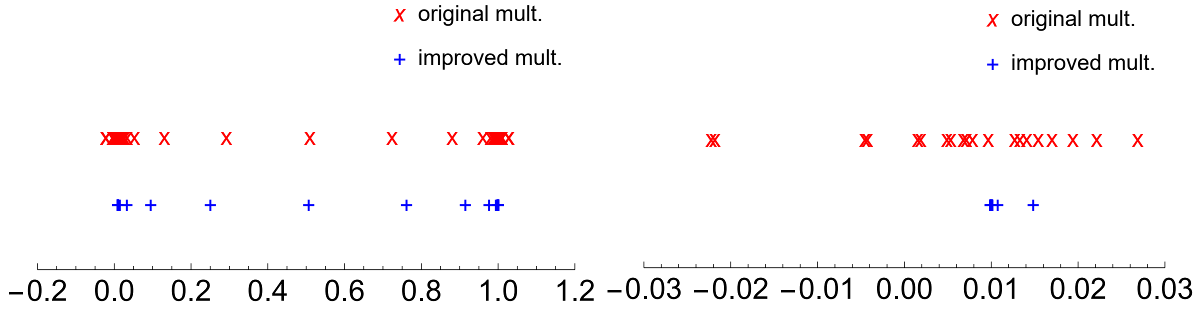
$$\underline{\underline{A}}_{\text{imp}} \cdot \mathbf{x} = \underline{\mathbf{b}}_{\text{random}}, \tag{5.25}$$

where $\underline{\mathbf{b}}_{\text{random}}$ is kept the same as in Eq. (5.16). The iterative details for this improved multiplication operator $\underline{\underline{m}}$ are displayed in Fig. 5.1 in blue.

One immediately recognizes that this improved multiplication operator $\underline{\underline{m}}$ requires much fewer iterations to reach the desired relative error of 10^{-5}: the original multiplication operator $\underline{m}$ takes 860 iterations, while the improved multiplication operator $\underline{\underline{m}}$ requires only 40.

To get a better understanding of why the number of iterations changes dramatically, we compute the eigenvalues of both $\underline{\underline{A}}_{\text{org}}$ and $\underline{\underline{A}}_{\text{imp}}$. Since all the eigenvalues have imaginary parts that are in amplitude smaller than 10^{-6}, we only focus on their real parts and display them in Fig. 5.2. We use red crosses to denote the eigenvalues associated with the original multiplication operator $\underline{m}$, and blue plus signs to denote the eigenvalues associated with the improved multiplication operator $\underline{\underline{m}}$. Fig. 5.2 (a) shows the real parts of all eigenvalues of $\underline{\underline{A}}_{\text{org}}$ and $\underline{\underline{A}}_{\text{imp}}$, and Fig. 5.2 (b) shows the distribution of the eigenvalues which real part is in $[-0.03, 0.03]$.

One can readily observe that all the eigenvalues associated with the improved multiplication operator have positive real parts, suggesting a positive definite system matrix A_{imp}, which is in line with the positivity of the function $f(x)$ in (5.20). The smallest real part of these eigenvalues associated with $\underline{\underline{m}}$ is about 0.0099, which is very close to its analytical limit $\frac{1}{1+99}$. On the contrary, the spectrum associated with $\underline{m}$ suggests an indefinite system, since there are both positive and negative real parts of those eigenvalues (the smallest one is -0.022). This change in the definiteness causes a significant difference in the conditioning of the systems: the condition number of the system associated with $\underline{m}$ is 2480, while its counterpart with $\underline{\underline{m}}$ is 566. This explains the dramatic difference in the convergence of the iterative solver BI-CGSTAB in Fig. 5.1.



(a) (b)

Figure 5.2: Comparison of the eigenvalues of the matrices A_{org} and A_{imp}. The label "original mult." refers to the original multiplication operator $\underline{m}$ and "improved mult." refers to the improved multiplication operator $\underline{\underline{m}}$. (a) real parts of all eigenvalues. (b) real parts of the eigenvalues in the range $[-0.03, 0.03]$.

5.3 Multiplication of two sampled functions

The equidistant list representation of a function was introduced in [111] to achieve a much faster multiplication operation than the multiplication of two sets of Gabor coefficients, for the spatial spectral method. We have discussed the original and the improved multiplication operators for two sets of Gabor coefficients in the preceding section. Now we focus on the multiplication of two sampled lists, which is the second type of multiplication used in the spatial spectral method.

The definition of the original multiplication operator of two uniformly sampled functions is given in a pointwise form in the following.

Definition 3. *Let $\overline{\mathbf{f}}$ and $\overline{\mathbf{u}}$ be the uniformly sampled functions of $f(x)$ and $u(x)$ according to Eq.* (5.4)*, the multiplication operator associated to $\overline{\mathbf{f}}$ is defined as*

$$\overline{m}(\overline{\mathbf{u}}): \ \mathbb{C}^{2L+1} \to \mathbb{C}^{2L+1}, \tag{5.26}$$

where the components satisfy $[\overline{m}(\overline{\mathbf{u}})]_i = \overline{\mathbf{f}}_i \cdot \overline{\mathbf{u}}_i$ for all $i = 1, 2, \ldots, 2L+1$.

We consider the problem in Eq. (5.20) again with the multiplication performed on the equidistantly sampled functions. The used Gabor parameters are given in Table 5.1. A similar procedure as in Section 5.2 has been followed to construct a matrix representation of the sampled-function-based multiplication operator associated with $\overline{\mathbf{f}}$. The matrix is denoted by $\overline{A}_{\text{org}}$, where the bar and the subscript indicate the multiplication is corresponding to the original list-based multiplication in the spatial spectral method [111]. We then consider the following linear system

$$\overline{A}_{\text{org}} \cdot \mathbf{x} = \overline{\mathbf{b}}_{\text{random}}, \tag{5.27}$$

where $\overline{\mathbf{b}}_{\text{random}}$ is a random vector in $\mathbb{C}^L$. We then use again the BI-CGSTAB iterative method to solve Eq. (5.27), and plot the iterative details in Fig. 5.3. We notice that more than 800 iterations are required to reach a relative error smaller than 10^{-5}.

Fig. 5.3 (a) suggests that the matrix $\overline{A}_{\text{org}}$ is an ill-conditioned system. When using the spatial spectral method, this ill-conditioning associated with the multiplication $\overline{m}$ implies difficulties when inverting the operator $\mathcal{C}_\varepsilon$ in Eq. (5.1). We also know that the invertibility of the operator $\mathcal{C}_\varepsilon$ is crucial since the NVF formulation relies on stable transformations from the auxiliary field $\mathbf{F}$ to the electric field $\mathbf{E}$ and vice versa [105, 107].

Moreover, the ill-conditioning caused by the multiplication operator $\overline{m}$ also has a severe impact on the entire system's conditioning. Recall Eq. (2.77), where the matrix form of the EFIE (5.26) is $(C - G \cdot M) \cdot \mathbf{u} = \mathbf{f}$. When the matrix C is already ill-conditioned, it is unlikely to get a well-conditioned system $(C - G \cdot M)$, since in most cases the Green function G and the field-material interaction operator M will not cure the spectrum for $(C - G \cdot M)$.

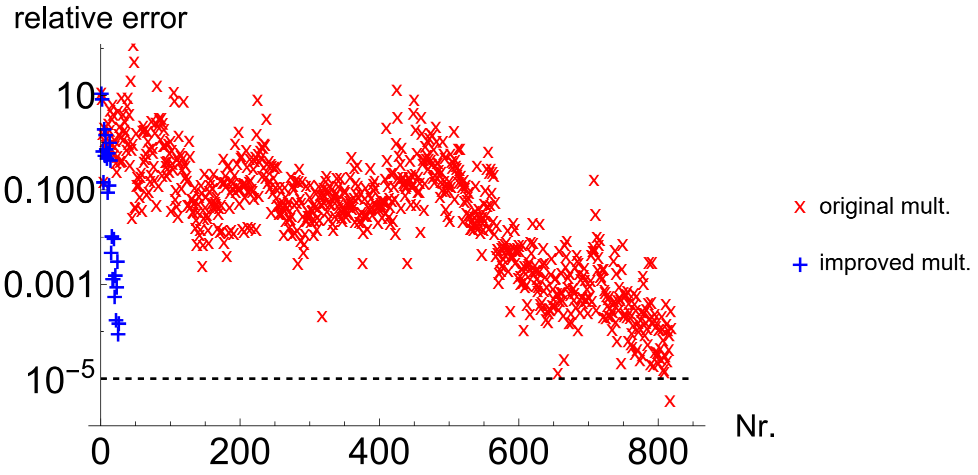


Figure 5.3: Iterative details based on the inversion of the multiplication operators $\overline{A}_{\text{org}}$ and $\overline{A}_{\text{imp}}$ for two sampled functions. The label "original mult." refers to the original multiplication operator $\overline{m}$ in Definition 3, and the label "improved mult." refers to the improved multiplication operator $\overline{\overline{m}}$ in Definition 4.

Clearly, a better multiplication operator for two sampled functions is needed for solving high-contrast scattering problems with the spatial spectral method. It is required to find the equivalent sampled-function-based representations for the Gabor-coefficient-based representations in Definition 2. Since the padding and restriction operators cured the multiplication operator for Gabor coefficients, we will look for an equivalent for the list-based representation. Based on the relation of a sampled function and its Gabor coefficients, we replace $\underline{P} \cdot \underline{\mathbf{u}}$ by $G^{-1} \cdot \underline{P} \cdot G \cdot \overline{\mathbf{u}}$, replace $\underline{U}$ by $G^{-1} \cdot \underline{U} \cdot G$, replace $\underline{\mathbf{f}}'$ by $G \cdot \underline{\mathbf{f}}'$, and then obtain $G^{-1} \cdot \underline{U} \cdot G \cdot \overline{m}(G^{-1} \cdot \underline{P} \cdot G \cdot \overline{\mathbf{u}})$, where the corresponding parameter should become $\underline{\mathbf{f}}'$. Additionally, we recall the Gabor transformations G and G^{-1} have been incorporated into a pair of Fourier transformations F and F^{-1} according to Eq. (12) in [111] to yield fast operations. Therefore the zero-padding with more harmonics in Gabor coefficients corresponds to increasing the range of the list representation in the spectral domain, and restricting the harmonics in Gabor coefficients corresponds to restricting the range of the list representation in the spectral domain. To be precise, we have the following approximations:

$$G^{-1} \cdot \underline{P} \cdot G \approx F^{-1} \cdot \tilde{P} \cdot F, \tag{5.28}$$

$$G^{-1} \cdot \underline{U} \cdot G \approx F^{-1} \cdot \tilde{U} \cdot F, \tag{5.29}$$

where we use $\tilde{P}$ and $\tilde{U}$ to represent an extension operator and a restriction operator on sampled functions in the spectral domain.

With the above analysis, we now introduce the following modified multiplication operator between two sampled functions.

Definition 4. *Let $\overline{\mathbf{f}}'$ and $\overline{\mathbf{u}}$ be the uniformly sampled functions of $f(x)$ and $u(x)$ according to Eq.* (5.4) *and Eq.* (5.6)*, respectively. The multiplication operator $\overline{\overline{m}}$ is defined as*

$$\overline{\overline{m}}(\overline{\mathbf{u}}) = \overline{U} \cdot \overline{m}(\overline{P} \cdot \overline{\mathbf{u}}), \tag{5.30}$$

where $\overline{m}$ is given in Definition 3 and takes $\overline{\mathbf{f}}'$ as a given parameter vector. Additionally, the operators $\overline{U}$ and $\overline{P}$ are defined as

$$\overline{P} = F^{-1} \cdot \tilde{P} \cdot F, \tag{5.31}$$

$$\overline{U} = F^{-1} \cdot \tilde{U} \cdot F, \tag{5.32}$$

where $\tilde{P}$ and $\tilde{U}$ represent a restriction and an extension operation to the sampled functions in the spectral domain, F and F^{-1} are the FFT-based Fourier transformations given in [111] between a spatial domain and a spectral domain.

The function $\overline{\overline{m}}$ is essentially a modification of the original multiplication operator $\overline{m}$ with a double-sampled list $\overline{\mathbf{f}}'$. We now discuss the two restriction operators $\underline{U}$ and $\tilde{U}$. From Eqs (5.11) and (5.12), we see that the operator $\underline{U}$ restricts the range of the Gabor coefficients by keeping the middle half. Suppose a spectral-domain sampled-function-based representation $F \cdot \overline{\mathbf{f}}$ has the range $[-K, K]$, for some positive integer K, then the effective range in the spectral domain is $\frac{2}{3} \cdot 2K = \frac{4}{3}K$, due to the Gabor oversampling by a factor $\frac{2}{3}$. Definition 4 takes the double-sampled function $\overline{\mathbf{f}}'$ as the associated parameter, therefore $F \cdot \overline{\mathbf{f}}'$ occupies the range $[-2K, 2K]$ in the spectral domain. To reach the original effective range $\frac{4}{3}K$ after the multiplication operator, we need a restriction factor of $\frac{1}{3}$ since $4K \cdot \frac{1}{3} = \frac{4K}{3}$. Hence, the operator $\tilde{U}$ restricts the range of the sampled function in the spectral domain by keeping the middle one third and setting the rest to zero.

Once again, we construct the matrix representation of $\overline{\overline{m}}$ and denote it by $\overline{A}_{\text{imp}}$. We then solve the following linear system by using the BI-CGSTAB iterative method, i.e. we solve

$$\overline{A}_{\text{imp}} \cdot \mathbf{x} = \overline{\mathbf{b}}_{\text{random}}, \tag{5.33}$$

where $\overline{\mathbf{b}}_{\text{random}}$ is the same vector as in Eq. (5.27). The iterative details for this improved multiplication operator $\overline{\overline{m}}$ are displayed in Fig. 5.3 (b). One can immediately recognize that inverting the matrix associated with the improved multiplication operator $\overline{\overline{m}}$ takes only 21 iterations to reach the same desired relative error.

We now compare the spectrum of $\overline{A}_{\text{org}}$ and $\overline{A}_{\text{imp}}$ in Fig. 5.4. Here we focus on the real parts of the eigenvalues again, since their imaginary parts have very small amplitude. Fig. 5.4 (a) shows the real parts of all eigenvalues of $\overline{A}_{\text{org}}$ and $\overline{A}_{\text{imp}}$, and Fig. 5.4 (b) shows the distribution of the eigenvalues which real part is in $[-0.06, 0.06]$.

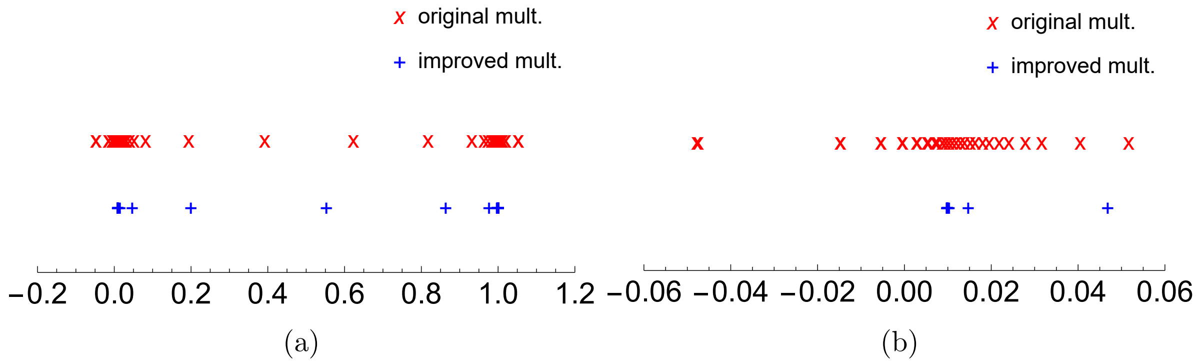


Figure 5.4: Comparison of eigenvalues of the matrices $\overline{A}_{\text{org}}$ and $\overline{A}_{\text{imp}}$. The label "original mult." refers to the original multiplication operator $\overline{m}$ and "improved mult." refers to the improved multiplication operator $\overline{\overline{m}}$. (a) Real parts of all eigenvalues. (b) Real parts of the eigenvalues in the range $[-0.06, 0.06]$.

Here we observe again that matrix $\overline{A}_{\text{org}}$ is indefinite since it has both positive and negative eigenvalues (starting from -0.05). Therefore, both the original multiplication operators $\underline{m}$ and $\overline{m}$ change in the definiteness of their matrix representations. The matrix $\overline{A}_{\text{imp}}$ associated with the improved multiplication operator $\overline{\overline{m}}$ is positive definite, with its minimum being 0.0098, which is very close to its analytical limit $\frac{1}{1+99}$. In addition, the condition number of $\overline{A}_{\text{org}}$ is 10081, while its counterpart with $\overline{A}_{\text{imp}}$ is 678. This explains the significant difference in the iterative details in Fig. 5.3.

5.4 Discussion with more examples

Fourier factorization rules for three types of products are discussed in [106]. The first type is the multiplication of two functions that have no concurrent discontinuities and

Laurent's rule should be employed to perform Fourier factorization. The second type is the multiplication of two functions that have only concurrent and complementary discontinuities, where the inverse rule should be employed instead of Laurent's factorization rule. The third type is the multiplication of two functions that have concurrent but not complementary discontinuities. This is the most difficult one, since it cannot be factorized by Laurent's rule or the inverse rule. Since there is no other rule available to overcome the difficulty, the third type of multiplication should be avoided during formulation (an example is the NVF formulation [105, 107]).

The example studied in Section 5.2 and Section 5.3 belongs to the first type, since it is essentially a multiplication operation, i.e., either $\underline{\underline{m}}$ or $\overline{\overline{m}}$, of a discontinuous function with a continuous function. In this section, we study two other examples that correspond to the second and the third type of product functions in [106] and we discuss the performances of the improved Gabor-based multiplication operator.

5.4.1 Example A: the second type of multiplication

Consider the following functions for all $x \in \mathbb{R}$:

$$f(x) = e^{-x^2} \cdot v(x), \tag{5.34}$$

$$u(x) = e^{-x^2} \cdot v(x), \tag{5.35}$$

$$h(x) = f(x) \cdot u(x). \tag{5.36}$$

where

$$v(x) = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases} \tag{5.37}$$

Note that $v(x)$ equals the standard sign function almost everywhere, except at the point $x = 0$. It is easy to notice that $f(x) \equiv u(x)$, $\forall x \in \mathbb{R}$ and the product function $h(x)$ is continuous everywhere due to the complementary discontinuities. Hence, this example belongs to the second type of multiplication in [106].

Firstly, we recall the Laurent rule and the inverse rule in Fourier factorization theory. Let $h_M(x)$ be an approximation of $h(x)$ based on a truncated Fourier series, i.e.

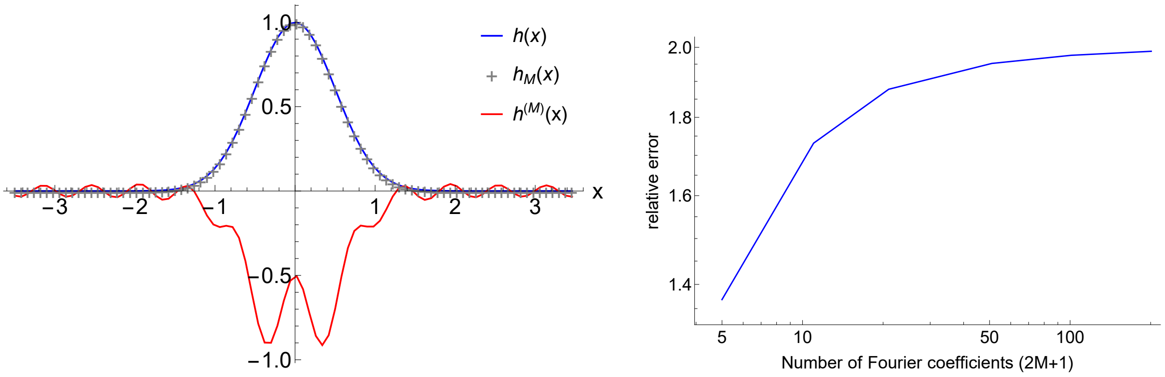
$$h_M(x) = \sum_{n=-M}^{M} h_n e^{jnx}, \tag{5.38}$$

where h_n is the n-th corresponding Fourier coefficient of $h(x)$. Laurent's rule states that $h(x)$ can be approximated by the following $h^{(M)}(x)$:

$$h^{(M)}(x) = \sum_{n=-M}^{M} h_n^{(M)} e^{jnx}, \tag{5.39}$$

$$h_n^{(M)} = \sum_{m=-M}^{M} f_{n-m} u_m, \tag{5.40}$$

where f_m and u_m are the m-th Fourier coefficients of $f(x)$ and $u(x)$, respectively.



(a) (b)

Figure 5.5: (a) Comparison of the analytical function $h(x)$, its approximation with a truncated Fourier series $h_M(x)$, and its approximation $h^{(M)}$ based on Laurent's rule. $M = 10$. (b) Relative error of $h_n^{(M)}$, i.e. the Fourier coefficients computed based on Laurent's rule, compared with h_n for different numbers of Fourier coefficients M.

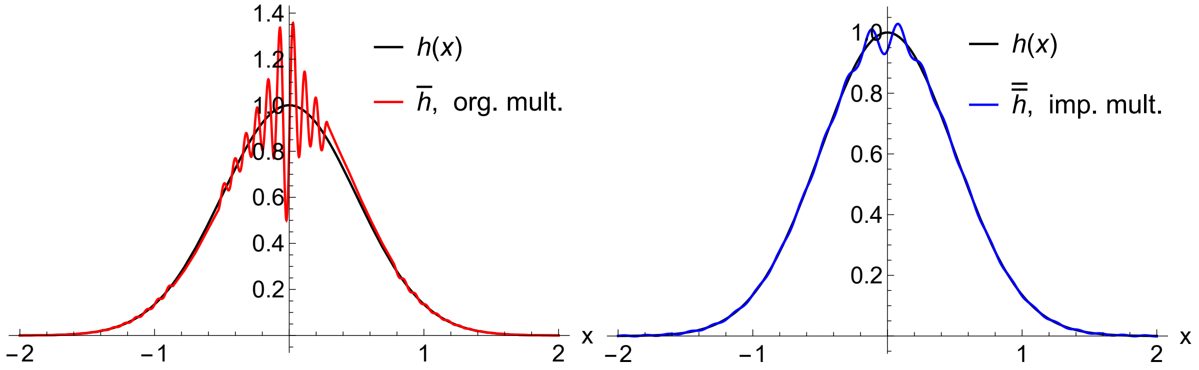
In Fig. 5.5 (a) we show the analytical function $h(x)$ given in Eq. (5.36), the finite Fourier approximation $h_M(x)$ given in (5.38), and an approximated function $h^{(M)}(x)$ given in (5.39), the latter two with $M = 10$. Clearly, Laurent's rule fails to generate accurate Fourier coefficients to approximate the product function $h(x)$. Furthermore, Fig. 5.5 (b) shows a divergent trend for the Fourier coefficients computed based on Laurent's rule when increasing the number of Fourier coefficients $2M + 1$ in the series expansion. According to the conclusions given in [106], Laurent's rule should break down on this example and the inverse rule should be used to perform the multiplication of $f(x)$ and $u(x)$. However, difficulties can occur when inverting the Toeplitz matrix corresponding to the multiplication by $f(x)$, which is generated by the Fourier coefficients of $1/f(x)$ due to the ill-conditioning of this matrix.

We now study this problem with Gabor frames. Let $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ be the uniformly sampled functions based on the original multiplication operator $\overline{m}$ and the improved multiplication operator $\overline{\overline{m}}$, respectively. The overlines indicate that we are working with equidistant lists of function values. The Gabor parameters used in this example are given in Table 5.2.

Table 5.2: Gabor parameters used in Examples A and B.

T_x	M	N	α_x	β_x
0.5	10	7	$\sqrt{2/3}$	$\sqrt{2/3}$

In Fig. 5.6 (a) and (b), we compare the approximated functions $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$, i.e., the continuous version of $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$, with the analytical reference $h(x)$. It is observed that strong oscillations occur in $\overline{\mathbf{h}}$ at $x = 0$, the position of the concurrent discontinuity. The approximated function $\overline{\overline{\mathbf{h}}}$, essentially based on the improved multiplication operator $\overline{\overline{m}}$, has fewer oscillations and the oscillations have a smaller amplitude around $x = 0$ and has a better global match to the analytical reference $h(x)$.



(a) (b)

Figure 5.6: Comparison of the approximated functions and the analytical reference function in Example A, where the equidistant sampled functions $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ are computed based on (a) the original multiplication operator $\overline{m}$. (b) The improved multiplication operator $\overline{\overline{m}}$.

We now compare the approximated functions based on the original multiplication operator $\overline{m}$ and the improved one $\overline{\overline{m}}$, i.e. the multiplications based on sampled functions. Let $\mathbf{h}_{\text{ref}}$ be the sampled function of $h(x)$, which is analytically given in Eq. (5.36), according to Eq. (5.4). Then we define the following relative errors:

$$e_{\text{org}} = \frac{||\overline{\mathbf{h}} - \mathbf{h}_{\text{ref}}||}{||\mathbf{h}_{\text{ref}}||}, \tag{5.41}$$

$$e_{\text{imp}} = \frac{||\overline{\overline{\mathbf{h}}} - \mathbf{h}_{\text{ref}}||}{||\mathbf{h}_{\text{ref}}||}, \tag{5.42}$$

where $|| \cdot ||$ is the ℓ^2 norm of a vector. A smaller relative error suggests a set of more accurate Gabor coefficients induced by the multiplication operator. By increasing the maximum of the modulation index, N, from 4 to 199 and by computing the corresponding relative errors of the Gabor coefficients, we obtain the results in Fig. 5.7, presented on a double-logarithmic scale.

From Fig. 5.7, one can readily see a convergent trend in the accuracy of the approximated functions for both the original multiplication operator $\overline{m}$ and the improved multiplication operator $\overline{\overline{m}}$. The current Example A clearly shows that the improved multiplication operator $\overline{\overline{m}}$ outperforms the original multiplication operator $\overline{m}$ in the second type of product function as well.

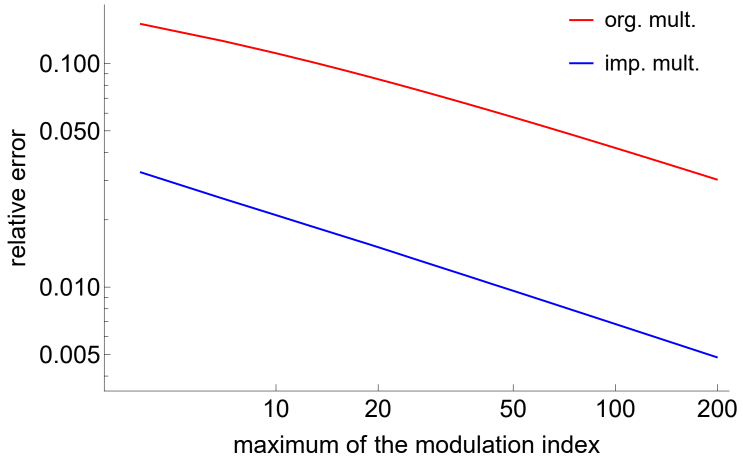


Figure 5.7: Accuracy of the approximated product functions based on the original multiplication ("org. mult.") operator $\overline{m}$ and the improved ("imp. mult.") one $\overline{\overline{m}}$, with respect to the maximum of the modulation index used in Gabor frames.

5.4.2 Example B: the third type of multiplication

In this example, we study the third type of product function with the multiplication operators $\overline{m}$ and $\overline{\overline{m}}$. Consider the following function

$$h(x,k) = f^k(x), \tag{5.43}$$

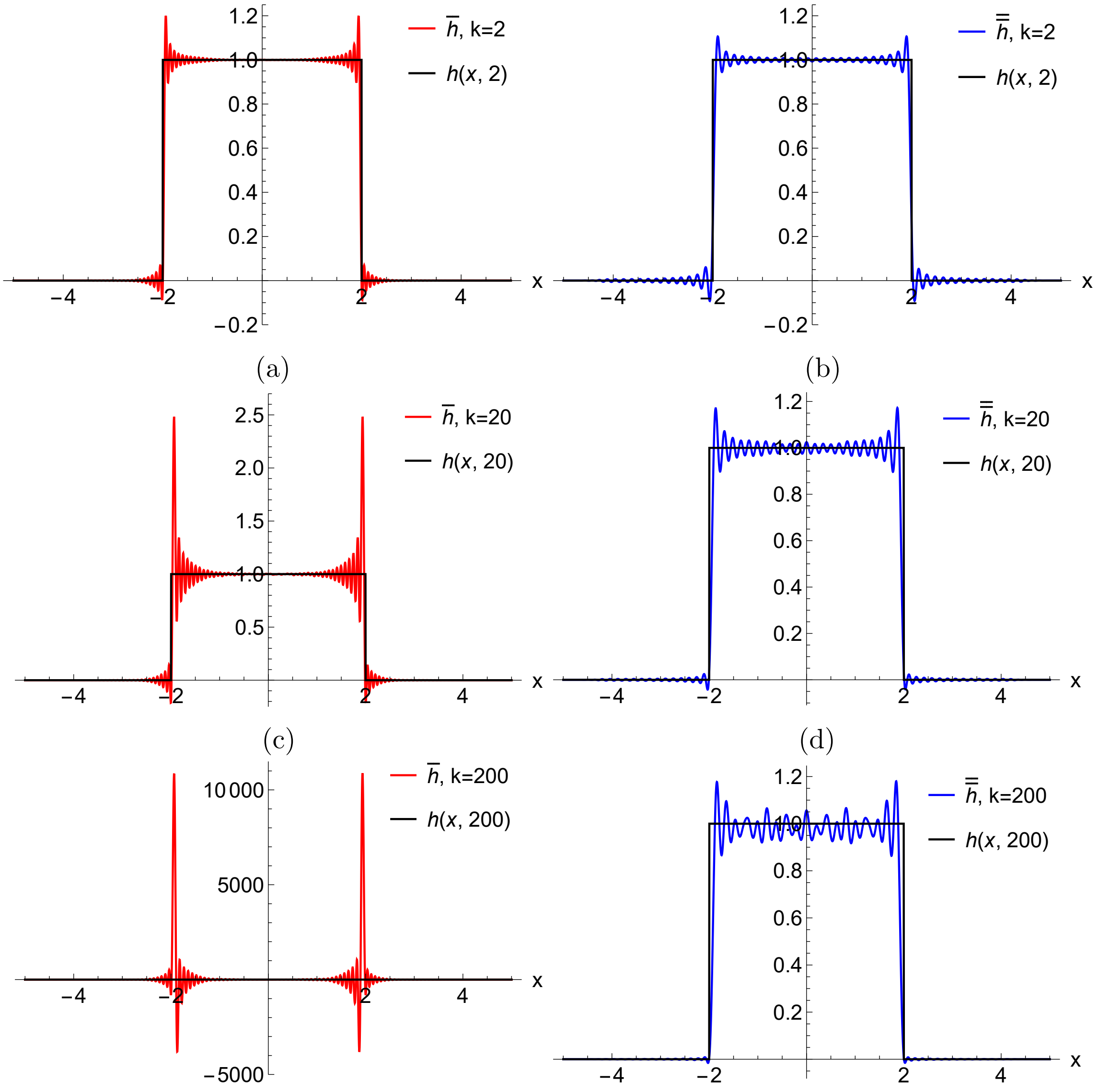
where k is a positive integer indicating the power to which the function f is elevated and

$$f(x) = \begin{cases} 1, & -2 \leq x \leq 2, \\ 0, & \text{otherwise.} \end{cases} \tag{5.44}$$

Clearly for any positive integer k we have $h(x,k) = f^k(x) \equiv f(x)$. The discontinuities involved in $h^k(x)$ are concurrent but not complimentary. Therefore, this example belongs to the third type and the most difficult product in [106].

Following the same procedure in Section 5.4.1, we are able to get the sampled functions $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ as approximations to the product function $h(x,k)$, based on the two multiplication operators $\overline{m}$ and $\overline{\overline{m}}$ and a given parameter k. The Gabor parameters used in Example B are again given in Table 5.2. Here we consider three cases specified by $k = 2$, $k = 20$ and $k = 200$, and for each case we compare the approximated functions $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ with the corresponding analytical reference $h(x,k)$. We show all the involved functions in Fig. 5.8.

From Fig. 5.8 (a), (c), and (e), it is striking that the approximated function $\overline{\mathbf{h}}$ behaves much poorer for larger k. Therefore, the original multiplication operator $\overline{m}$ is unable to perform the third type of multiplication in [106]. On the other hand, Fig. 5.8 (b), (d), and (f) suggest that the improved multiplication operator $\overline{\overline{m}}$ yields a better-stabilized operation, compared with the original operator. However, we still notice that the difference between $\overline{\mathbf{h}}$ and the reference $h(x,k)$ slowly increases with increasing k.



(a) (b) (c) (d) (e) (f)

Figure 5.8: Comparisons of the approximated functions $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ based on the original multiplication operator $\overline{m}$ (left column) and the improved multiplication operator $\overline{\overline{m}}$ (right column). (a) and (b): $k = 2$. (c) and (d): $k = 20$. (e) and (f): $k = 200$.

Recalling Fig. 5.4 (a), the maximum eigenvalue $\lambda_{\max}$ of $\overline{A}_{\text{org}}$ is larger than 1, while Fig. 5.4 (b) shows the maximum eigenvalue $\lambda_{\max}$ of $\overline{A}_{\text{imp}}$ is still bounded by 1. The eigenvalues larger than 1 of the matrix representation of $\overline{m}$ essentially cause the severe overshooting of the sampled functions $\overline{\mathbf{h}}$ in Fig. 5.8 (a), (c), and (e). This is because the maximum eigenvalue of the matrix representation of $\overline{m}$ associated with $h(x, k)$ becomes $\lambda_{\max}^{k}$, and it can be a large number for a large k when $\lambda_{\max} > 1$.

Next, we study how the accuracy of the approximated functions $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ changes for increasing k, under the two multiplication operators $\overline{m}$ and $\overline{\overline{m}}$. Let $\overline{\mathbf{f}}$ and $\overline{\mathbf{u}}$ be the equidistantly sampled functions of $f(x)$ in Eq. (5.44) and $u(x) = 1$, then we define the following references:

$$\mathbf{c}_{\text{ref1}} = \overline{m}(\overline{\mathbf{u}}), \tag{5.45}$$

$$\mathbf{c}_{\text{ref2}} = \overline{\overline{m}}(\overline{\mathbf{u}}). \tag{5.46}$$

Then we construct the vectors containing the approximated sampled functions of $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ through the following nested operations:

$$\mathbf{c}_{\text{org}}(k) = \underbrace{\overline{m}(\ldots, \overline{m}(\overline{m}(\overline{\mathbf{u}})))}_{k \text{ times}}, \tag{5.47}$$

$$\mathbf{c}_{\text{imp}}(k) = \underbrace{\overline{\overline{m}}(\ldots, \overline{\overline{m}}(\overline{\overline{m}}(\overline{\mathbf{u}})))}_{k \text{ times}}. \tag{5.48}$$

The relative errors in $\mathbf{c}_{\text{org}}(k)$ and $\mathbf{c}_{\text{imp}}(k)$ are defined as:

$$e_1(k) = \frac{\|\mathbf{c}_{\text{ref1}} - \mathbf{c}_{\text{org}}(k)\|}{\|\mathbf{c}_{\text{ref1}}\|}, \tag{5.49}$$

$$e_2(k) = \frac{\|\mathbf{c}_{\text{ref2}} - \mathbf{c}_{\text{imp}}(k)\|}{\|\mathbf{c}_{\text{ref2}}\|}, \tag{5.50}$$

where $\|\cdot\|$ stands for the ℓ^2 norm of a vector.

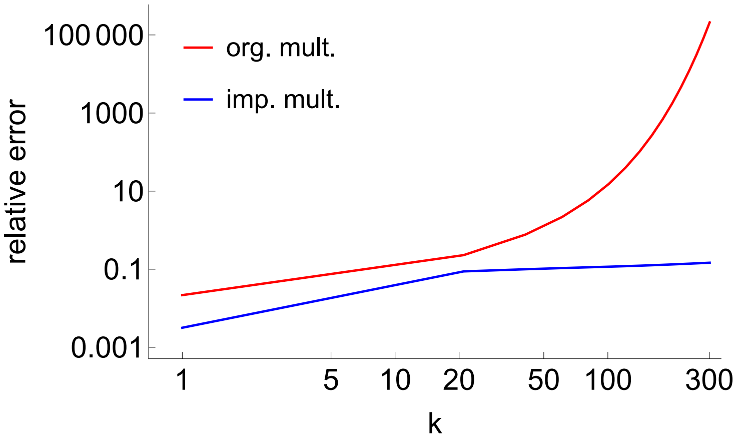


Figure 5.9: Accuracy comparison of the approximated product functions $\overline{\mathbf{h}}$ and $\overline{\overline{\mathbf{h}}}$ of $h(x, k)$, based on the original multiplication ("org. mult.") operator $\overline{m}$ and the improved multiplication ("imp. mult.") operator $\overline{\overline{m}}$, with respect to different k.

Fig. 5.9 shows how the relative error in the approximated sampled functions changes for $k = 1, \ldots, 300$ on a double-logarithmic scale. Note that the red lines represent the accuracy of the approximated sampled function computed based on $\overline{m}$ and their counterparts in blue represent the accuracy of the approximated sampled function computed based on $\overline{\overline{m}}$. Here we notice that the results with the original multiplication operator diverge rapidly away from the desired result after $k = 20$, while the results with the improved multiplication operator $\overline{\overline{m}}$ suggests a slowly increasing error when k increases.

5.5 Numerical experiment

To show the effectiveness of the improved field material interaction operator introduced in Section 5.3, we have tested the spatial spectral Maxwell solver [24] with the improved field material interaction operator on a 3D scattering problem with high contrast. We compare the solution with an independent reference and perform a convergence study by refining the discretization.

5.5.1 Geometry configuration and discretization

A bar-shaped scatterer with a high relative permittivity $\varepsilon_r = 17$ is placed in free space, see Fig. 5.10. The scatterer's dimensions are given as 300 × 200 × 100 nm. A normally incident plane wave with a wavelength $\lambda = 425$ nm illuminates the bar from above, with the electric field polarized in the x-direction and with unit amplitude. The wave vector of the incident plane wave is $\mathbf{k} = (0, 0, k_0)$, where $k_0 = 2\pi/\lambda$ is the wave number in free space.

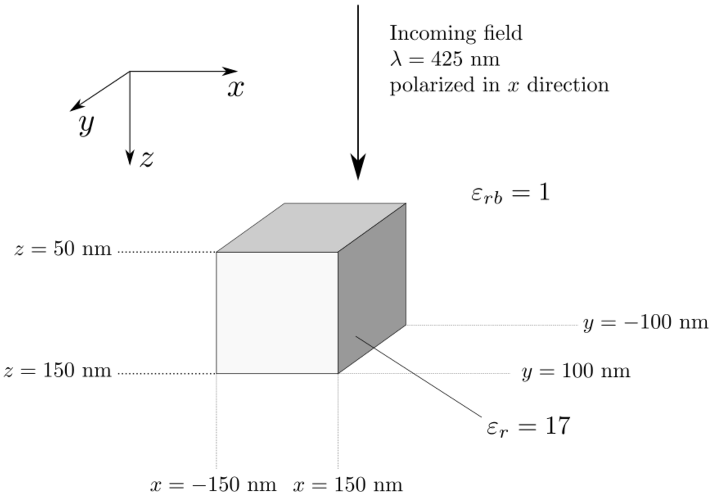


Figure 5.10: Normal incident plane wave illuminated and a bar-shaped scatterer with relative permittivity $\varepsilon_r = 17$ in free space.

Table 5.3 displays all the discretization parameters used for this example. Following the notations used in Section 2.3.2, we use T_x, T_y to denote the Gabor window lengths, m_x,

m_y to denote the spatial shift numbers, and n_x, n_y to denote the frequency modulation numbers. The subscripts x and y indicate that the parameters are associated with the x- and y-direction, respectively. We use N_z to represent the number of PWL functions that are used in the z direction. Based on the parameters in the second row, we are going to show the near-field results in Section 5.5.2, and far-field results in Section 5.5.3. The values of n_x, n_y, and N_z in the third and the fourth rows will be varied, since we are going to consider different discretizations in a convergence study in Section 5.5.3. Additionally, the oversampling parameters of the Gabor frame, i.e. α_x, α_y and β_x, β_y, are specified in the table. We use IDR(16) as the iterative method to find the solution, and the maximum number of iterations is set to 1250.[2]

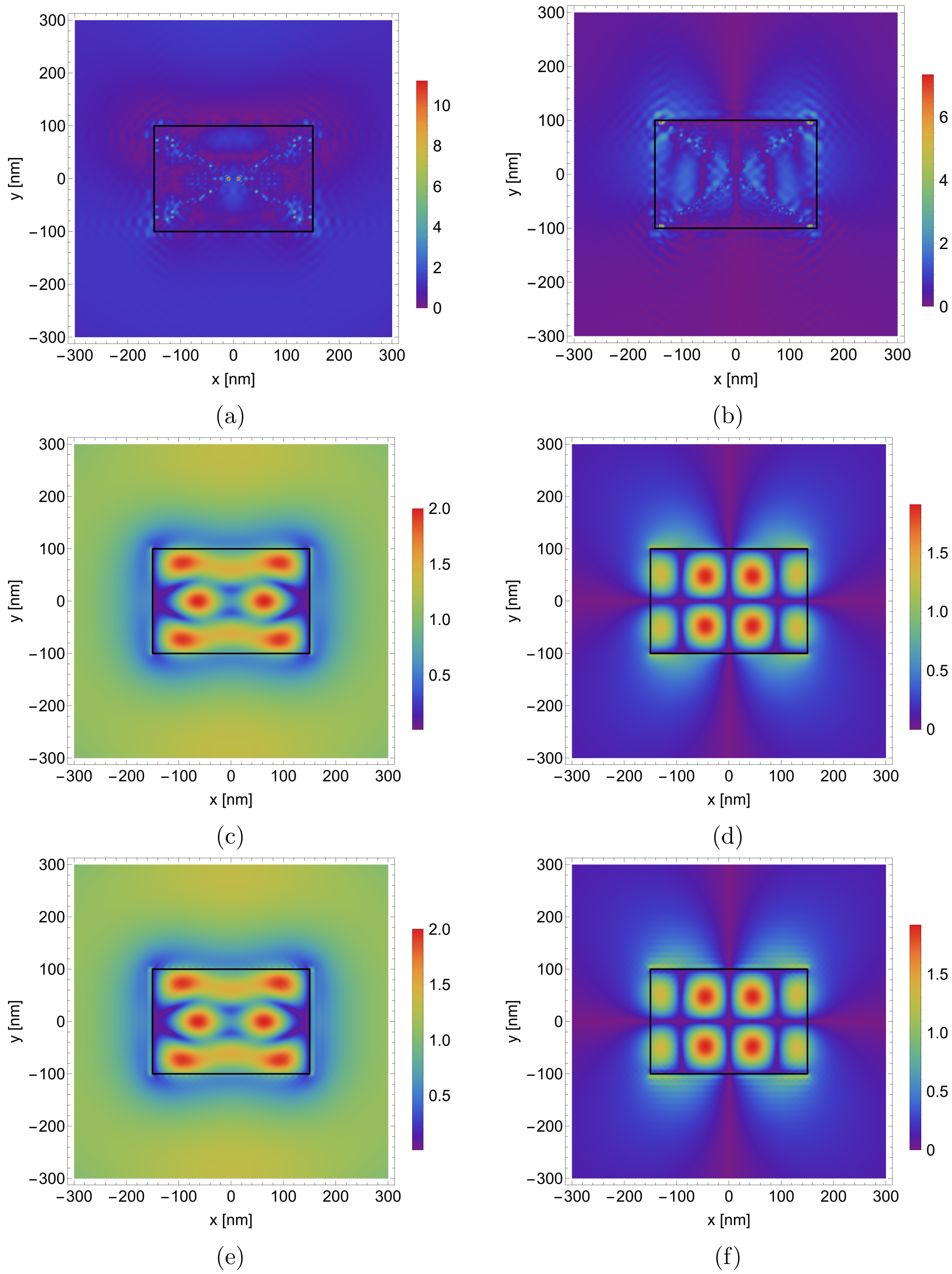
Table 5.3: Discretization parameters used in Section 5.4. Parameters contained in the second row are used for the simulations in Section 5.5.2. The third row and the fourth row include the parameters used for the convergence study in Fig. 5.14 (a) and (b).

case	T_x [nm]	T_y [nm]	m_x, m_y	n_x, n_y	N_z	α_x, α_y	β_x, β_y
Section 5.5.2	130	130	$-4:4$	$-28:28$	201	$\sqrt{2/3}$	$\sqrt{2/3}$
Fig. 5.14 (a)	130	130	$-4:4$	vary	201	$\sqrt{2/3}$	$\sqrt{2/3}$
Fig. 5.14 (b)	130	130	$-4:4$	$-28:28$	vary	$\sqrt{2/3}$	$\sqrt{2/3}$

5.5.2 Comparison of near-field results against a reference

We have computed the solution for the same 3D scattering problem in the spatial spectral Maxwell solver with both the original field-material interaction operator, given in Definition 3 in Section 5.3, and the improved field-material interaction operator, given in Definition 4 in Section 5.3. We compare the solutions with the commercial reference JCMWave, which employs FEM [198]. Discretization parameters are specified in Table 5.3 and the total number of unknowns for the corresponding matrix equation (2.77) is 1.6×10^8. In particular, we consider the total electric fields at $z = 120$ nm. Fig. 5.11 (a) and (b) show the absolute values of $E_x(x, y)$ and $E_y(x, y)$ as obtained from the original multiplication operator $\overline{m}$, where black rectangles are added to denote the boundary of the scatterer. Absolute values of $E_x(x, y)$ and $E_y(x, y)$ obtained from the JCMWave, which are considered as a numerical reference, are displayed in Fig. 5.11 (c) and (d). It is apparent that the results based on the original multiplication operator are totally different from the reference. In Fig. 5.11 (e) and (f) we show absolute values of $E_x(x, y)$ and $E_y(x, y)$ based on the improved multiplication operator $\overline{\overline{m}}$. One can easily see that the near-field solutions based on the improved multiplication operator match the reference mostly well, except for some oscillations that can be observed near the boundary of the scatterer, which are due to the Gibbs phenomenon.

[2] In this example we are only interested in the accuracy of the solution that can be achieved with the improved multiplication operator. A reduction in the number of iterations is discussed in Section 6.4 where a preconditioner is employed.



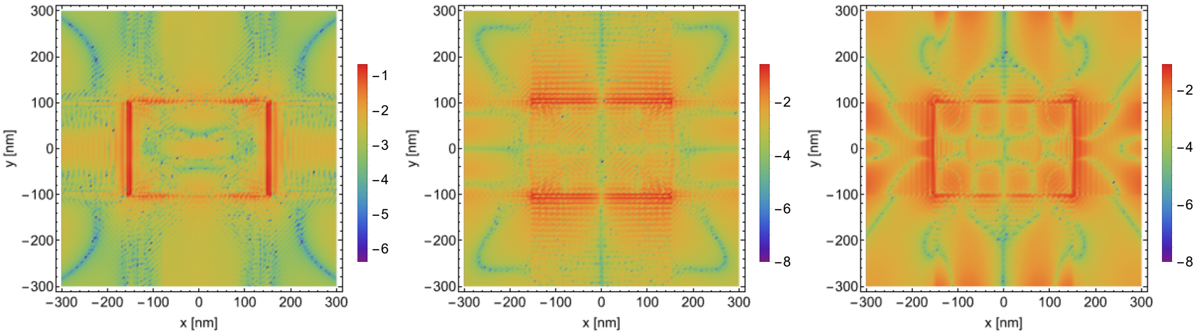
(a) (b)

(c) (d)

(e) (f)

Figure 5.11: Cartesian components of the total electric field at $z = 120$ nm. Left column: $|E_x(x, y)|$, right column: $|E_y(x, y)|$. (a) and (b): the multiplication is performed with the original field material interaction operator. (c) and (d): JCMWave reference. (e) and (f) the multiplication is performed with the improved field material interaction operator.

Additionally, we show the absolute error, i.e. for the x component $||E_x| - |E_{x,JCM}||$ and analogous absolute errors for the y and z components, of the solutions computed based on the improved operator in Fig. 5.12. The ringing effect of the Gibbs phenomenon is observed again.



(a) (b) (c)

Figure 5.12: Absolute error in the total electric field between the solution obtained with the improved multiplication operator and the result from the JCMWave reference. All are shown on a $\log_{10}$ color scale. (a) Absolute error in $|E_x(x,y)|$. (b) Absolute error in $|E_y(x,y)|$. (c) Absolute error in $|E_z(x,y)|$.

5.5.3 A convergence study in the spectral domain and far-field results

To illustrate the convergence of the proposed method under refinement of the discretization, we test the same scattering problem as given in Section 5.5.1 with refinements of the discretization parameters in the x, y and z directions.

Table 5.4: Refined discretization along x and y directions.

Discretization	1	2	3	4
$-N_x : N_x$, $-N_y : N_y$	$-10 : 10$	$-16 : 16$	$-22 : 22$	$-28 : 28$
Δ_x, Δ_y [nm]	5	3.2	2.3	1.8
Nr. of unknowns	$2.2 \cdot 10^7$	$5.3 \cdot 10^7$	$9.9 \cdot 10^7$	$1.6 \cdot 10^8$

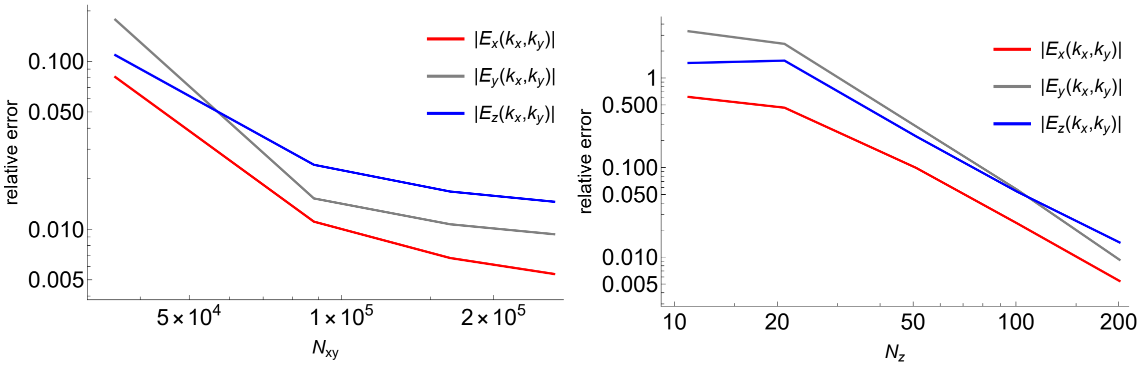
Firstly, we apply the discretizations specified in the third row in Table 5.3, and set up four cases with an increasing number of modulated Gabor functions per window length, i.e., n_x and n_y in Table 5.3 range from 21 to 57. We compute the resolution Δ_x along the x-direction according to the following formula:

$$\Delta_x = \frac{\alpha_x T_x}{2N_x + 1}, \tag{5.51}$$

where $-N_x : N_x$ is the range for n_x. Analogously, we get Δ_y. The resolutions in the x and y directions and the number of unknowns in the corresponding linear system are given in Table 5.4. We now compute and analyze the relative errors of the four simulations in Table 5.4. We use $E^s_{\mathrm{ref},x}(k_x, k_y)$ to denote the far-field JCMWave reference in the spectral domain, and $E^s_x(k_x, k_y)$ to denote the far-field solution obtained within the spatial spectral method with the improved multiplication operator $\overline{\overline{m}}$. Let Λ be a uniform grid in the spectral domain containing $124,301$ sample points within the Ewald circle. Then we can define vectors $\mathbf{v}_x$, $\mathbf{v}_{\mathrm{ref},x}$ by evaluating $E^s_x(k_x, k_y)$ and $E^s_{\mathrm{ref},x}(k_x, k_y)$ on the grid Λ, respectively. Then we define the relative error of the amplitude in the x-direction according to:

$$r_x = \frac{\big\| |\mathbf{v}_x| - |\mathbf{v}_{\mathrm{ref},x}| \big\|}{\big\| |\mathbf{v}_{\mathrm{ref},x}| \big\|_\infty}, \tag{5.52}$$

where $\|\cdot\|$ and $\|\cdot\|_\infty$ stand for the ℓ^2 norm and the ℓ^∞ norm of a vector, respectively. Analogously, we can compute r_y and r_z. Fig. 5.13 (a) shows these relative errors with corresponding discretizations as specified in Table 5.4 on a double-logarithmic scale. The horizontal axis represents the total number of sample points in the transverse plane, where $N_{xy} = (2M_x + 1) \cdot (2M_y + 1) \cdot (2N_x + 1) \cdot (2N_y + 1)$. The vertical axis represents the relative errors r_x, r_y, and r_z. The decay of the relative error is clearly observed with respect to the refinement of the discretization parameters in the transverse plane. On the other hand, we also notice that the convergent trend in Fig. 5.13 (a) levels off after some point. This is a sign that another factor, e.g., the discretization along the longitudinal direction, might limit the far-field accuracy thereafter. Therefore we do a convergence study along z-direction in the following experiment.



(a) (b)

Figure 5.13: Convergence study of the spectral-domain Cartesian components computed based on the improved operator. (a) Refined discretization along x and y directions as specified in Table 5.4. (b) Refined discretization along z direction as specified in Table 5.5.

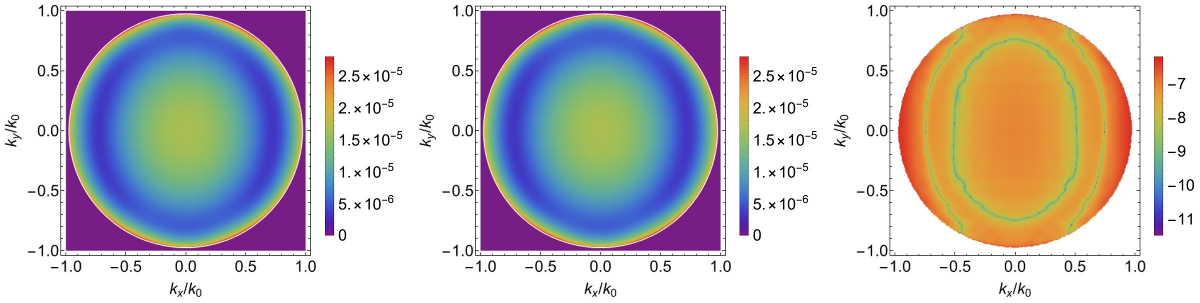
Secondly, we apply the discretizations specified in the fourth row in Table 5.3. The number of Gabor frames per window length is fixed by setting $N_x = N_y = 28$, and the resolution in the transverse plane is $\Delta_x = \Delta_y = 1.8$ nm. We then set up five cases by varying the number of the PWL functions in the z direction, i.e. N_z is changing from 11 to 201. The increasing resolution due to the refined longitudinal discretization and corresponding dimensions of the system are shown in Table 5.5.

Table 5.5: Refined discretization in z direction.

Discretization	1	2	3	4	5
N_z	11	21	51	101	201
Δ_z [nm]	10	5	2	1	0.5
Nr. of unknowns	$8.7 \cdot 10^6$	$1.7 \cdot 10^7$	$4 \cdot 10^7$	$8 \cdot 10^7$	1.6×10^8

Fig. 5.13 (b) shows the relative errors due to different discretizations according to Table 5.5, also in a double-logarithmic scale. Here we observe again a convergent trend with respect to the refinement of the discretization along the z-direction. The results before $N_z = 21$ yield a large relative error. The convergent trend starts after $N_z = 21$. Be aware that the case in Table 5.4 with Discretization 4 is identical to the case in Table 5.5 with Discretization 5. Therefore, the relatively slow decay in Fig. 5.13 (a) and the relatively fast decay in Fig. 5.13 (b) around the last case indicate that the far-field accuracy is more limited by the discretization in the longitudinal direction than the discretization in the transverse directions.

In particular, we consider the case with the finest resolution in all directions, i.e. Discretization 4 in Table 5.4 or Discretization 5 in Table 5.5. The far-field amplitude in the plane $z = 0$ nm is given in Fig. 5.14 (a), based on the improved multiplication operator $\overline{\overline{m}}$ given in Section 5.3 Definition 4 in the spatial spectral solver, and Fig. 5.14 (b), based on the JCMWave reference. Only the fields within the Ewald circle in the spectral domain are displayed. The absolute error in this far-field solution is displayed on a $\log_{10}$ color scale in Fig. 5.14 (c).



(a)

(b)

(c)

Figure 5.14: Far-field results in comparison. (a) far-field amplitude $\|\mathbf{E}^s(k_x, k_y)\|$, based on the improved multiplication operator, (b) far-field amplitude of the electric field based on the JCMWave reference, and (c) absolute error of the far-field amplitude $\|\mathbf{E}^s(k_x, k_y)\|$ on a $\log_{10}$ color scale compared to the JCMWave reference.

5.6 Conclusion

We have introduced two improved multiplication operators: one between two sets of Gabor coefficients and one between two equidistantly sampled functions. Each improved multiplication operator implies a modification to the result obtained from the original multiplication operator, as defined in [111,197]. The modification is performed in the spectral domain and yields a smoother approximated product function in the spatial domain.

As a further analysis of the improved multiplication operators, we studied three examples corresponding to the three types of multiplication classified in [106]. In the first example, the improved multiplication operators (for both two sets of Gabor coefficients and two sampled functions) correspond to a well-conditioned linear system, resulting in a lower cost when inverting the operator $\mathcal{C}_\varepsilon$ in the spatial spectral method. In the second and the third example, we focused on the improved multiplication for two sampled functions and showed the accuracy of the Gabor coefficients of the product function is significantly improved with the improved multiplication operator. Finally, we studied a 3D scattering problem with the improved multiplication operator in the spatial spectral method and validate both the near-field and far-field results with a commercial FEM-based reference.

Chapter 6

A normal-vector-field-based preconditioner for a spatial spectral domain-integral equation method for multi-layered electromagnetic scattering problems[1]

A normal-vector-field-based block diagonal-preconditioner for the spatial spectral integral method is proposed for an electromagnetic scattering problem with multi-layered medium. This preconditioner has a block-diagonal matrix structure for both 2D TM polarization and 3D cases. Spectral analysis shows that the preconditioned system has a more clustered eigenvalue distribution, compared to the unpreconditioned system. For the cases with high contrast or negative permittivity, numerical experiments illustrate that the preconditioned system requires fewer iterations than the unpreconditioned system. The total computation time is reduced accordingly while the accuracy based on the normal-vector field formulation of the solution is preserved.

6.1 Introduction

In electrical engineering Maxwell solvers for electromagnetic scattering problems have wide and important applications, which range from semiconductor metrology in integrated circuits (ICs) production [7, 200, 201], to designing elements on nanophotonic chips [202, 203], and to analysing metamaterials [204, 205]. In these cases, it is required to have fast and accurate Maxwell solvers, especially for the cases where the number of unknowns is large.

Different types of Maxwell solvers have been developed in the past decades to solve electromagnetic scattering problems. When the incident fields and solutions are station-

[1] This chapter was published as [199]

ary or time-harmonic, one can solve the problem with a frequency-domain Maxwell solver. The frequency-domain solver can be more computationally efficient than a time-domain Maxwell solver, and it can be divided into two categories. The first kind relies on a differential form of Maxwell's equations, popular methods in this first category are the finite-difference (FD) [34] and finite element methods (FEM) [65]. The second category depends on an integral-equation formulation of Maxwell's equations, which incorporates the Green function and the volume is restricted to the support of the sources of the electromagnetic field. Both domain integral equations [206,207] and surface integral equations [208] belong to the latter category.

In [108–110], a spatial spectral method is proposed to solve two-dimensional (2D) transverse electric (TE), 2D transverse magnetic (TM) and three-dimensional (3D) scattering problems in a layered medium, respectively. The main differences between this method and other volume integral equation solvers are: (1) a Gabor frame is used as a discretization in the transverse plane, which brings a fast and accurate Fourier transformation; and (2) a spectral integration path is chosen to avoid the singularities of the Green function in the spectral domain. The accuracy is improved by introducing an auxiliary field based on the local normal-vector field (NVF) formulation [107].

The above spatial spectral discretization approach leads to a high-dimensional linear system of equations. Usually iterative methods such as GMRES [138], BiCG-type methods [85,140,209], or IDR(s) [86] are deployed to solve these large linear systems instead of a direct method [135]. For each iteration, this spatial spectral solver reaches a computational complexity of $O(N \log N)$ in terms of the matrix-vector product. However, convergence difficulties are observed in terms of a large number of iterations when the underlying physical problem has high-contrast or negative-permittivity scatterers embedded in the layered medium, or when the scatterer is large. Preconditioning is usually a vital component for high-dimensional linear systems with a poor convergence rate, to enable practical computations within a reasonable time [139]. A good preconditioner transforms the original system into a system that has the same solution, but exhibits better convergence performance. Furthermore, constructing and executing this preconditioner should be fast because it will be performed in every iteration as an extra matrix-vector product (MVP).

Optimal circulant preconditioners have been successfully used in domain integral equations in one-dimensional (1D) and 2D TE, or E-polarized cases to accelerate the iteration, see [210] and [211]. Circulant-type preconditioners have also proved effective to solve the system in the form of $I - GX$ with multi-level Toeplitz structure [212]. For scattering in periodic setups, the integral-equation formulation in the transverse directions exploits a continuous auxiliary field formulation together with a normal-vector field around object boundaries [105, 107]. In that case, the linear system corresponding to the integral equation can be written in the form $(C - GM)\mathbf{u} = \mathbf{f}$, where the matrices C and M are block-Toeplitz-Toeplitz-block (BTTB) matrices, and the matrix G represents the Green operator. In [213], the matrix C^{-1} and its approximations have been proposed as preconditioners and promising improvements were obtained after deploying these preconditioners. For the nonperiodic case, the spatial spectral method based on Gabor frames and an auxiliary field in combination with a normal-vector field formulation [108, 110] bears a close resemblance to the case of fully spectral methods for periodic structures. Therefore, it is a natural idea to extend the application of the C^{-1} preconditioner in [213] to the Gabor-frame based spatial spectral solver, which is the main objective of this paper. To be specific, we show that this NVF-based preconditioner has a block-diagonal structure and we illustrate that this preconditioner can reduce the number of iterations, while preserving the accuracy of the solution for high contrasts or negative permittivities.

This paper is organized as follows. In Section 6.2 we recall the most important details of the 2D TM and 3D spatial spectral Maxwell solver and we establish the NVF-based preconditioner. In Section 6.3 we discuss the effects of this NVF-based preconditioner based on spectral analysis. Numerical experiments are discussed in Section 6.4, which contains three experiments for which we show the reduction in the number of iterations, an accuracy validation, and a comparison in computation time. Section 6.5 contains the conclusions.

6.2 Formulation

Consider the following 2D or 3D scattering problem in Fig. 6.1. A multi-layered dielectric medium is placed in between two dielectric half-spaces. We define a Cartesian coordinate system such that all layers are stacked along the z direction as background materials, and each layer i ($1 \leq i \leq N$) has a constant relative permittivity ε_{rbi}. A scattering object, which is made of a different material, is located in a finite domain $D \in \mathbb{R}^3$ and is completely embedded within layer i. The relative permittivity of the scatterer is ε_{rs} and one can define a global relative permittivity function $\varepsilon_r(\mathbf{x})$ to distinguish all materials, where $\mathbf{x} = (x, y, z)$ denotes the spatial coordinates. In the absence of the scatterer, the incident electric field $\mathbf{E}^i(\mathbf{x})$ can be calculated as in [10].

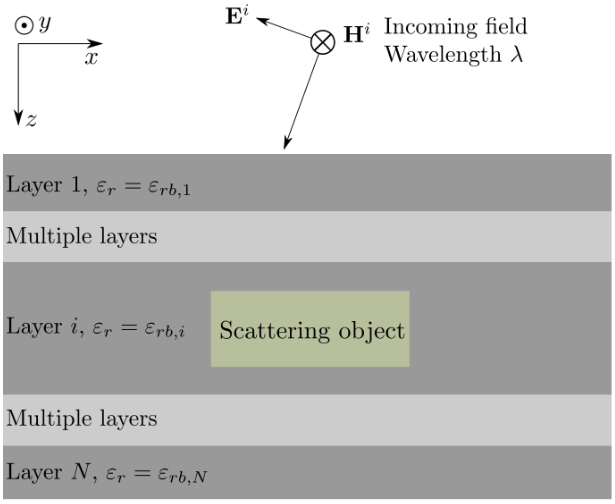

y
x
z
$\mathbf{E}^i$
$\mathbf{H}^i$
Incoming field
Wavelength λ
Layer 1, $\varepsilon_r = \varepsilon_{rb,1}$
Multiple layers
Layer i, $\varepsilon_r = \varepsilon_{rb,i}$
Scattering object
Multiple layers
Layer N, $\varepsilon_r = \varepsilon_{rb,N}$

Figure 6.1: Geometric setting for a multi-layer medium with embedded scattering object.

6.2.1 Summary of the spatial spectral method

The spatial spectral method [24] is developed based on the following domain integral representation:

$$\begin{aligned}\mathbf{E}^i(\mathbf{x}_T, z) =& \mathbf{E}(\mathbf{x}_T, z) \\ & - \mathcal{F}_T^{-1}\left\{\int_{\mathbb{R}} G(z'|\mathbf{k}_T, z) \cdot \mathcal{F}_T[\mathbf{J}(\mathbf{x}_T, z')]dz'\right\}\end{aligned} \tag{6.1}$$

where $\mathbf{x}_T$ denotes the spatial Cartesian coordinates in the transverse plane (i.e., $\mathbf{x}_T = (x, y)$ in the 3D case and $\mathbf{x}_T = x$ in the 2D case), similarly, $\mathbf{k}_T$ denotes the spatial Fourier transform variables in the transverse direction, i.e. with respect to $\mathbf{x}_T$. Note that $\mathcal{F}_T$ and $\mathcal{F}_T^{-1}$ denote a pair of Fourier transformations in the transverse plane between $\mathbf{x}_T$ and $\mathbf{k}_T$. G is the spectral-domain Green operator in the multi-layered medium. $\mathbf{E}(\mathbf{x}_T, z)$ represents the unknown total electric field, $\mathbf{J}(\mathbf{x}_T, z)$ is the contrast current density given by the spatial field-material interaction:

$$\mathbf{J}(\mathbf{x}_T, z) = j\omega\varepsilon_0\varepsilon_{rbi}\chi(\mathbf{x}_T, z)\mathbf{E}(\mathbf{x}_T, z), \tag{6.2}$$

where ε_0 is the permittivity of free space, ε_{rbi} is the relative permittivity of the ith homogeneous layer in the background medium that contains the scatterers. Given the relative permittivity function $\varepsilon_r(\mathbf{x}_T, z)$, the contrast function $\chi(\mathbf{x}_T, z)$, bound to layer i, is defined as

$$\chi(\mathbf{x}_T, z) = \frac{\varepsilon_r(\mathbf{x}_T, z)}{\varepsilon_{rbi}(z)} - 1, \tag{6.3}$$

which is only supported on the domain of the scatterer.

One important feature of this spatial spectral method is that a Gabor frame is used to perform discretization in the transverse plane in both spatial and spectral domains. A Gabor-frame expansion for any $f(\mathbf{x}_T) \in \mathrm{L}^2(\mathbb{R}^2)$ in the spatial domain is

$$f(\mathbf{x}_T) = \sum_{\mathbf{m},\mathbf{n}} f_{\mathbf{m},\mathbf{n}} g_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T), \tag{6.4}$$

in which $\mathbf{m}$ and $\mathbf{n}$ represent the spatial and the spectral shift number, respectively, $g_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)$ is a Gabor frame function and $f_{\mathbf{m},\mathbf{n}}$ is a Gabor coefficient. Gabor coefficients are computed via the Gabor transformation:

$$f_{\mathbf{m},\mathbf{n}} = \int f(\mathbf{x}_T)\eta^*_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)d\mathbf{x}_T, \tag{6.5}$$

where $\eta_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)$ is the dual frame function and is computed via the Moore–Penrose inverse [132]. Full representations of the Gabor frame function and its dual frame function can be found in [108, 110]. The main advantage of this Gabor-frame-based discretization is that it establishes a fast relation between the spatial domain and the spectral domain. The Fourier transform of a spatial Gabor frame function $g_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)$ yields a Gabor frame in the spectral domain, and the Gabor coefficients of the spectral function $\hat{f}(\mathbf{k}_T)$ can be readily obtained via simple operations on the spatial Gabor coefficients $f_{\mathbf{m},\mathbf{n}}$ [197]. This property guarantees fast transformations between the spatial and the spectral domains and eventually contribute to the $O(N \log N)$ computational complexity for the matrix-vector product of the spatial spectral method, where N represents the total number of unknowns after discretization. In [111], a set of basis functions is calculated based on equidistant Dirac delta test functions and an approximation of the exact Gabor-based discretization is introduced in [110] for 3D scattering problems. These new basis functions yield faster operations like multiplication and FFT-based Fourier transformation, which reduces the computation time and preserves accuracy.

In the z direction, the integral in Eq. (6.1) is discretized in terms of piecewise-linear (PWL) expansion functions:

$$\Lambda(z) = \begin{cases} 1 - \frac{|z-p\Delta_z|}{\Delta_z} & \text{if } |z-p\Delta_z| < \Delta_z \\ 0 & \text{if } |z-p\Delta_z| > \Delta_z \end{cases}, \tag{6.6}$$

where Δ_z is the discretization step in the z direction and $1 \leq p \leq N_z$ denotes the index of the sample points along the z direction. N_z denotes the total number of sample points in the z direction. Another feature of the spatial spectral method is that a deformed integration path on the complex plane is chosen as alternative to an integration path on the real axis, to properly handle the branch cuts of the dielectric half-spaces and the poles that represent guided waves of the layered medium. Based on the reflection interfaces within a multi-layered medium [10], effective reflection coefficients are defined in [108–110]. Representing the Green function in the spectral domain along this integration path avoids the tedious calculation of Sommerfeld integrals.

To improve the accuracy and efficiency of the Gabor expansion in the presence of discontinuous permittivities in the transverse plane, a local normal-vector field formulation is used in this spatial spectral method. Based on the Li rules [106], which provide a framework to assess whether functions with discontinuities can be multiplied or not, the normal-vector field formulation was introduced by Popov and Nevière [105] to improve the convergence in Fourier analysis. The main idea of the normal-vector field formulation is to perform spatial multiplications on the continuous components of the electric field $\mathbf{E}$ and the electric flux density $\mathbf{D}$, which together constitute the auxiliary field $\mathbf{F}$, and then derive their discontinuous components from the multiplication by the field-material interactions. The normal-vector field $\mathbf{F}$ can then be transformed to the total electric field $\mathbf{E}$ and the contrast current function $\mathbf{J}$ through

$$\begin{aligned} \mathbf{E} &= C\mathbf{F}, \\ \mathbf{J} &= M\mathbf{F}. \end{aligned} \tag{6.7}$$

Explicit expressions of components of matrices C and M expressed in Cartesian coordinates are given in [108], [110] and [107].

6.2.2 The NVF-based block-diagonal preconditioner

Based on the domain integral representation (6.1) and the normal-vector field formulation (6.7), the spatial spectral method can be represented by the following linear system:

$$L\mathbf{u} = \mathbf{f}, \tag{6.8}$$

where $L \in \mathbb{C}^{N\times N}$ is the system matrix, the inhomogeneous term $\mathbf{f} \in \mathbb{C}^N$ represents the incident field $\mathbf{E}^i$, $\mathbf{u} \in \mathbb{C}^N$ contains the expansion coefficients of the auxiliary field $\mathbf{F}$ to be determined, and N represents the number of unknowns. The system matrix A can be decomposed as

$$L = C - G \cdot M, \tag{6.9}$$

where C and M transform the normal-vector field $\mathbf{F}$ into the total electric field $\mathbf{E}$ and the contrast current $\mathbf{J}$ through Eq. (6.7), and G denotes the Green tensor operation in combination with a pair of Fourier transformations. In the spatial spectral solver [24], the matrix L is implemented implicitly to avoid storing a full system matrix.

The structures of matrices L, C, G, M depend on the order of the discretization indexes associated with either the transverse plane or the z direction. When choosing the index associated to z-samples as the outermost one, i.e. the slowest changing index when moving row-wise or column-wise, matrices C and M have a block-diagonal structure with each block containing the Gabor coefficients of the operators related to the contrast χ (defined in [108] and [110]). The block-diagonal structure essentially comes from the direct (spatial) multiplication between the χ-related operators and the auxiliary field $\mathbf{F}$ per z sample. The Green matrix G contains the Gabor transformation of the homogeneous-medium Green tensor and the reflected waves from the layer interfaces [109, 110] and therefore it has a denser structure at the block level. On the other hand, the fact that the Gabor frames have effectively a finite support in the spectral domain yields some sparsity per block of the matrix G. Simplified structures of matrices L, C, G and M are given in Fig. 6.2.

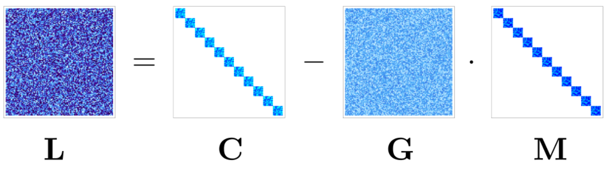


Figure 6.2: Sparsity patterns of the matrices L, C, G, and M.

In [213], the system of a light scattering problem for a 2D-periodic structure was also represented in the form $C - GM$, but then for an expansion in terms of (discrete) Fourier modes. For that case, the matrix C is a block-diagonal matrix and each diagonal block of the matrix C is a so-called BTTB-block matrix. Preliminary investigations have shown that the number of iterations can be reduced significantly by taking the full inverse matrix C^{-1} as a preconditioner. Since the Gabor-frame transformation is a unitary transformation

[214], it is natural to transfer the idea of this C^{-1} preconditioner to the spatial spectral method, where the Gabor representation is used in the transverse plane.

The N_z block matrices of C in Fig. 6.2 come from the N_z sampling points in z direction. Each block corresponds to the Gabor transformation of the function $\chi(\mathbf{x}_T, z_p)$ and the normal-vector field in the transverse plane with some fixed $z = z_p$ $(1 \leq z_p \leq N_z)$. Therefore, for a dielectric scatterer that has a uniform cross section in the z direction, the block submatrices of C are identical to each other. Together with the sparsity, owing to the block-diagonal structure, one can readily see that the matrix C^{-1} can be constructed by inverting one block submatrix of C. This simplifies the computational procedure in practice and makes C^{-1} a good candidate to precondition the original system $L\mathbf{u} = \mathbf{f}$. Hence we refer to the matrix C^{-1} as the normal-vector-field-based block-diagonal (NVF-BD) preconditioner for the spatial spectral method.

6.3 An indication of a clustered spectrum

It is well known that the convergence rate of an iterative method highly depends on the distribution of the eigenvalues of the system matrix: the more clustered the spectrum is, the faster the convergence rate will be, see [215, Chapter. 1]. Hence, a good preconditioner should yield a clustered spectrum for the preconditioned system matrix and result in an increased convergence rate. The clustering effect has been studied in detail for various types of preconditioners, e.g. circulant preconditioners [216–218], Toeplitz preconditioners [219–221], and block Toeplitz preconditioners [222]. Following the analysis for the above preconditioners, we compare the eigenvalue distributions of the systems with and without applying the NVF-BD preconditioner.

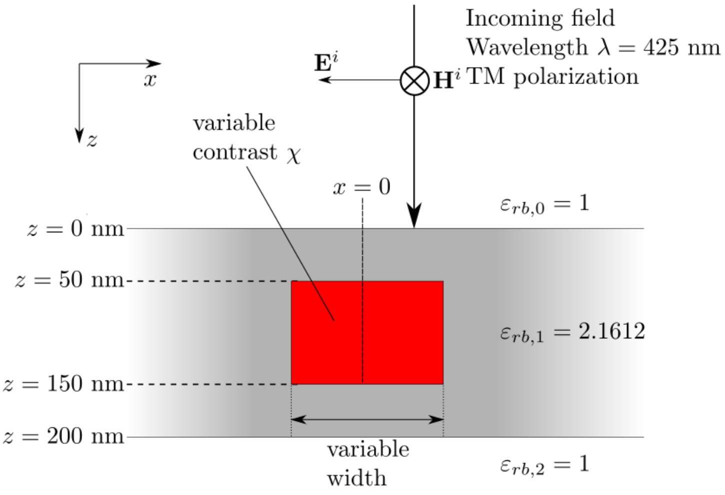


Figure 6.3: Scattering setup: a 2D TM polarized field is incident on a dielectric object (in red) embedded in a layered medium composed of SiO$_2$ and vacuum. $\varepsilon_{rb,i}$, for $0 \leq i \leq 4$, denotes the relative permittivities of these layers.

To this end, we consider a 2D scattering problem as presented in Fig. 6.3, where one rectangular scatterer is embedded in a layer of SiO_2 enclosed by two vacuum half-spaces. The incident field is a plane wave with wavelength 425 nm that is normally incident with respect to the xy plane and the incident electric field $\mathbf{E}^i$ is polarized along the x direction. The substrate medium SiO_2 has a relative permittivity $\varepsilon_{rb} = 2.16$, and the scatterer has a relative permittivity $\varepsilon_r = 54$. Therefore the scatterer has contrast $\chi = 24$ and its length in the x direction is 200 nm. With such a high-contrast case we expect a better conditioned system after applying the NVF-BD preconditioner. In this example there are 15990 unknowns.

In Fig. 6.4 we compare the absolute values and real parts of the eigenvalues. Both the original system and the preconditioned system are indefinite but not strongly: among the 15990 eigenvalues only 14 of them have negative real parts. Throughout the rest of this article, 'org' represents the original system and 'pdr' represents the system after applying the NVF-BD preconditioner. In Fig. 6.4 (a), the horizontal axis denotes six intervals ranging from 0.003 to 32, and the vertical axis represents the number of the eigenvalues, which absolute values belong to each of the corresponding intervals, on a log scale. A significant difference is observed when comparing their minimum absolute eigenvalues: the minimum absolute eigenvalue is shifted away from the origin from 3.8×10^{-3} to 7.6×10^{-2}. In Fig. 6.4 (b), we see the real part of those eigenvalues that satisfy $-0.5 \leq \text{Re}(\lambda_i) \leq 2$. It is clear that without the NVF-BD preconditioner, the original system has much more eigenvalues close to 0, while after applying the NVF-BD preconditioner, only a few eigenvalues around 0 remain and there are much more eigenvalues clustered around 1. The distribution of eigenvalues plays a crucial role in a system's conditioning, especially when the maximum eigenvalue does not change dramatically.

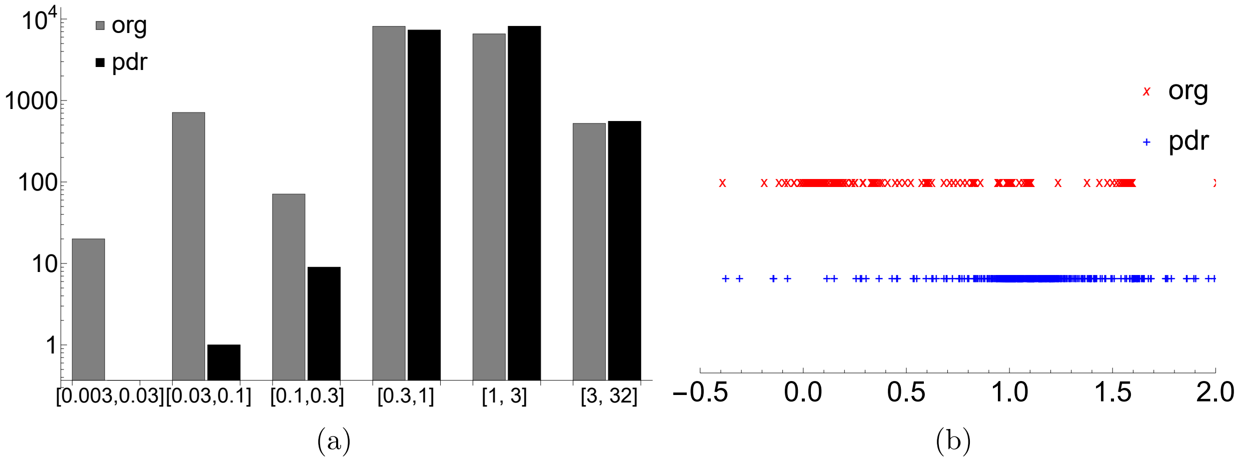


Figure 6.4: Comparison of eigenvalue distributions: (a) number of eigenvalues with absolute value located in per indicated interval, for the original system (org) and the preconditioned system (pdr). Note that the minimum and the maximum absolute eigenvalues of the original system are 0.0038 and 25, and the counterparts for the preconditioned system are 0.076 and 25. (b) Real parts of the eigenvalues $\text{Re}(\lambda_i)$ in the range $[-0.5, 2]$.

Table 6.1: Percentage comparison for eigenvalues located within the interval $[1-\delta, 1+\delta]$, given δ as a parameter.

δ	% of org.	% of pdr.
10^{-2}	89.5	93.0
10^{-4}	85.4	91.6
10^{-6}	67.0	81.4
10^{-8}	36.5	71.5

Fig. 6.4 also shows that the preconditioned system's spectrum is more clustered around 1. To verify this observation, we counted how many eigenvalues are located within the interval $[1-\delta, 1+\delta]$ for a given $\delta > 0$. We obtain the percentages by dividing by the total number of eigenvalues and compare them in Table 6.1. Note that in the original system only 36.5% of the eigenvalues were located within the disc centered at 1 with radius 10^{-8} in the complex plane, while this number becomes 71.5% after applying the NVF-BD preconditioner. Clearly, there is a stronger clustering of the eigenvalues around 1 in the preconditioned system. Analogous to the clustering effects studied in other preconditioners [216–222], we expect this promising indicator of the NVF-BD preconditioner can reduce the number of iterations as well.

6.4 Numerical Results

To show the effectiveness of the NVF-BD preconditioner, we have tested the preconditioned system on the following three scattering problems: (A) a 2D TM rectangular object with high contrast, (B) a 2D TM metal grating problem with negative permittivity and (C) a 3D bar-shaped object with high contrast. In all cases we mainly focus on the reduction in the number of iterations after applying the NVF-BD preconditioner. We also show the reduction in computation time in case (A), and compare the solution in case (B) with an independent reference. The iterative method used in all three cases is the BiCGstab(2) algorithm [85], with the maximum number of iterations set to 1250. In Table 6.2 we summarize all Gabor parameters used in these three problems. Note that case (A-1), case (A-2) and case (A-3) are three variants of case (A), which are used to demonstrate the NVF-BD preconditioner's effects on larger-scatter cases and computation time. Following the notations in [108–110], we use X, m and n to denote the Gabor window length, the spatial shift number, and the frequency modulation number, respectively, and we use N_z to represent how many PWL functions are used in the z direction. Further, $p = 3$ and $q = 2$ are the oversampling parameters for the Gabor frames. Note that case (C) is a 3D problem and we use the same discretization parameters in both x and y directions.

We define the relative error in step k, with corresponding solution vector $\mathbf{u}_k$, as $e_k = \frac{\|L\mathbf{u}_k - \mathbf{f}\|}{\|\mathbf{f}\|}$, with the system matrix L and the inhomogeneous term $\mathbf{f}$ introduced in Section 6.2, and $\|\cdot\|$ denotes the ℓ^2 norm of a vector. The iterative procedure is terminated once a relative error of 10^{-5} or less is reached. It is known that different iterative methods yield differences in convergence behaviour, especially for high-contrast cases. However, comparing the difference in convergence of the various iterative methods is not the aim of this paper.

Table 6.2: Discretization parameters used in simulation cases (A), (B) and (C).

case	X [nm]	m	n	N_z
(A)	100	$-7:7$	$-40:40$	41
(A-1)	100	varying	$-40:40$	41
(A-2)	100	$-5:5$	varying	101
(A-3)	100	$-5:5$	$-100:100$	varying
(B)	500	$-12:12$	$-40:40$	29
(C)	100	$-4:4$	$-10:10$	21

6.4.1 Case (A): a 2D TM high-contrast problem

In the first case we consider the 2D scattering problem in Fig. 6.3 again and keep the geometry parameters as introduced in Section 6.3. Table 6.2 displays the discretization parameters we used. Note that there are 81 frame functions used in each Gabor window length X, which yields a resolution of 1 nm in the x direction. In the z direction the PWL functions are employed with sample distance $\Delta_z = 2.5$ nm. Discretization parameters are given under case (A) in Table 6.2.

Table 6.3: Total number of iterations for Simulation case (A) for a scatterer with different contrast χ but the same geometric size. Note that "1250+" means the iterative solver fails to reach the desired relative error within 1250 iterations.

χ	org	pdr
2	7	4
4	17	9
8	50	22
16	245	66
24	1227	112
32	1250+	247
48	1250+	446
64	1250+	743

To see the effect of the NVF-BD preconditioner on the number of iterations, we fix the object's size by taking its width $w = 200$ nm and change the value of the contrast χ. The contrast ranges from 2 to 64 and we are more interested in the high-contrast cases, since they are more challenging. We compare the number of iterations for the original solver and the preconditioned solver in Table 6.3. Due to the nature of the BiCGstab(2) algorithm, one iteration represents four matrix-vector products (MVPs). Note that in the low-contrast cases such as $\chi \leq 4$ the NVF-BD preconditioner saves about 50% of the iterations. For the cases where $8 \leq \chi \leq 24$ the total number of iterations is reduced by up to 90%, when $\chi \geq 32$ the unpreconditioned system fails to converge within 1250 iterations, whereas the NVF-BD preconditioner makes the solver converge within an acceptable number of iterations.

Fig. 6.5 shows the evolution of the relative error versus the iteration count for the original system and the preconditioned system for the specific case $\chi = 24$, which corresponds to a relative permittivity $\varepsilon_r = 51.87$ for the rectangular scatterer. The horizontal axis denotes the number of iterations within the iterative solver, and the vertical axis denotes the relative error of the approximated solution at each iteration. Clearly, the preconditioned system significantly outperforms the original system in this high-contrast case. One possible reason for this significant reduction in number of iterations is that the NVF formulation plays a dominant role in the behavior of the iterative solver acting on the original system. The NVF-BD preconditioner improves the distribution of eigenvalues, as observed in Fig. 6.4, and also yields a much better conditioned system. The reduction in the number of iterations also saves a significant amount of computation time. We recorded the total computation times for this case on a single-core Intel(R) Xeon(R) Gold 6148 CPU at 2.40 GHz with 755 GB RAM. The original system takes 11,829.5 seconds (3 hours, 17 minutes and 12 seconds), while the preconditioned system only needs 1,151.5 seconds (19 minutes and 12 seconds) to reach the desired relative error of $1 \cdot 10^{-5}$.

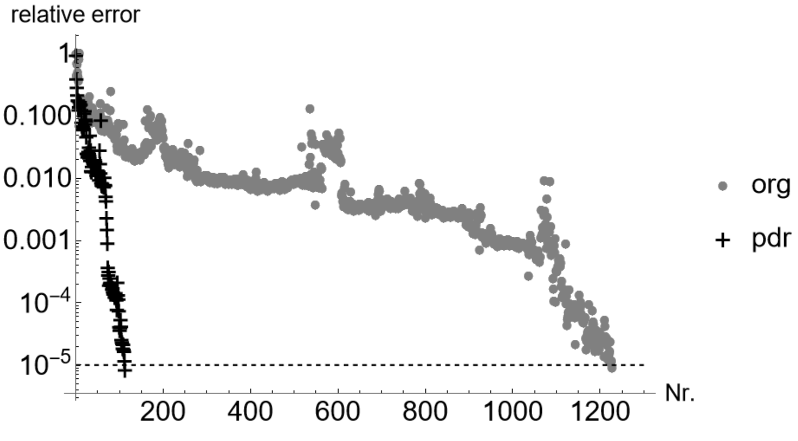


Figure 6.5: Convergence of the iterative solver for the high contrast case in Fig. 6.3 with $\chi = 24$. The dashed line denotes the desired relative error $1 \cdot 10^{-5}$.

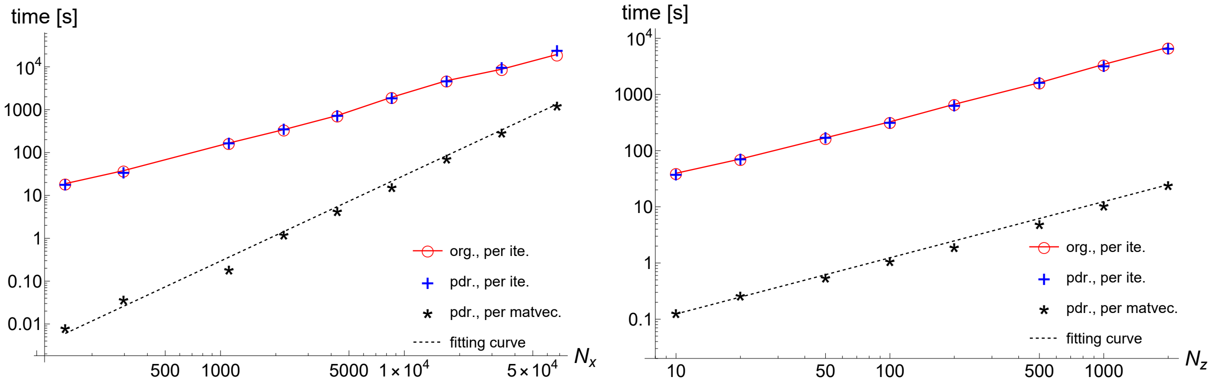
To explore the performance of the NVF-BD preconditioner on a larger scattering object, we change the scatterer's size and keep its contrast constant. The scatterer's width is changed from 200 nm to 1100 nm, which implies that the range of the x coordinate of the scatterer changes from $[-100, 100]$ nm to $[-550, 550]$ nm. We set the spatial shift index m of the Gabor frame in Table 6.2 from $m = -8 : 8$ to $m = -12 : 12$, and therefore the corresponding computation domain is increased from $[-500, 500]$ nm to $[-850, 850]$ nm in the x direction, which covers the scatterer's domain and a part of its near field. Other discretization parameters are given under case (A-1) in Table 6.2. For all cases, the scatterer's contrast is kept at $\chi = 16$, which corresponds to a relative permittivity $\varepsilon_r = 36.74$. Table 6.4 presents the number of iterations for the original system and the preconditioned system. We observe that the preconditioned system can handle a significantly larger range of widths of the scatterer. Hence, the NVF-BD preconditioner reduces the number of iterations not only in cases of high contrast but also in cases where the scatterer has a larger width.

Table 6.4: Total number of iterations recorded for the scatterers with different width but with a constant contrast $\chi = 16$.

width [nm]	org	pdr
200	232	66
300	432	91
400	1102	229
500	970	258
600	1250+	333
700	1250+	527
900	1250+	930
1100	1250+	1030

Next, we investigate the NVF-BD preconditioner's effect on computation time. The total computation time equals to the initialization time and solution time. During the initialization of the NVF-BD preconditioner, the matrix C^{-1} is computed based on Doolittle LU factorization, and one only has to compute the inverse of a single block matrix of C since the contrast χ is a constant along its height. The total solution time is equal to the product of the average solution time per iteration and the total number of iterations. One may not obtain a significant reduction in the total computation time for the preconditioned system if the time per iteration increases a lot due to the extra four MVPs induced by the preconditioner per iteration. We compare the computation time per iteration for the original solver with that for the preconditioned system by considering the single scatterer case in Fig. 6.3 with $\chi = 8$ and width $w = 200$ nm. Note that each block matrix C_i $(1 \leq i \leq N_z)$ has dimension N_x, and the MVP of matrix block C_i has a quadratic complexity. Hence we expect the MVP with matrix C^{-1} should have a complexity of $\mathcal{O}(N_z N_x^2)$.

In Fig. 6.6, we compare the solution time per iteration and the extra computation time per MVP of the preconditioner by changing the number of unknowns in both x and z directions, respectively. The vertical axes denote the average solution time per iteration in seconds. The horizontal axis in Fig. 6.6 (a) shows the number of unknowns N_x in the x direction. Note that $N_x = (2m+1)\cdot(2n+1)$, where the spatial shift index m satisfies $-5 \leq m \leq 5$, the frequency modulation index n satisfies $n \in \{-N, \ldots, N\}$ and N ranges from 6 to 3080. This corresponds to a resolution in the x direction that ranges from 6.231 nm to 0.013 nm. In the z direction, a total of 101 PWL functions are used with sample distance $\Delta_z = 1$ nm. All discretization parameters used in Fig. 6.6 (a) are summarized under case (A-2) in Table 6.2. The horizontal axis in Fig. 6.6 (b) denotes that N_z PWL functions are used in the calculation. N_z ranges from 10 to 2000, which corresponds to a resolution Δ_z in the z direction from 10 nm to 0.05 nm. In this case we have $-5 \leq m \leq 5$ and $-100 \leq n \leq 100$. All discretization parameters used in Fig. 6.6 (b) are summarized under case (A-3) in Table 6.2. In both Fig. 6.6 (a) and (b), the red dots and the blue crosses are computed based on the total solution time divided by the total number of iterations, and the gray stars are the computation time per MVP due to the preconditioner only. All the simulations were performed on a dual 20-core Intel(R) Xeon(R) Gold 6148 CPU at 2.40 GHz with 755 GB RAM.



(a) (b)

Figure 6.6: Average computation time per iteration for (a) N_x and (b) N_z unknowns. Note that "org. per ite." means the computation time per iteration for the unpreconditioned system, "pdr. per ite." means the computation time per iteration for the system after applying the NVF-BD preconditioner, "pdr. per matvec." means the time penalty due to the extra MVP in the preconditioned system, and "fitting curve" means the best fit based on the data points in the N_x and N_z cases, respectively.

The average computation time per iteration for the original system and the preconditioned system, and the average computation time per MVP of the preconditioned system are displayed in Fig. 6.6 (a) and (b). It is clear that the average computation time increases when a finer discretization is taken in either the x or z direction. The analytical representation of the fitting curve in Fig. 6.6 (a) is $T(N_x) = 2.94 \cdot 10^{-7} N_x^2$ and in (b) it is $T(N_z) = 0.012 N_z$, where T is the average computation time per iteration. The recorded data points of the MVP time coincide with the fitting curve well. Therefore we confirm our prediction that the extra MVP operation in the preconditioned system has a complexity

$\mathcal{O}(N_z N_x^2)$. Furthermore, we observe that both original system and preconditioned system have similar average computation time per iteration for almost the entire range of N_x and N_z cases, except for the last two data points in Fig. 6.6 (a), where the time of the extra MVP due to the preconditioner is non-negligible compared with MVP of the original system and the other operations in the BiCGstab(2) iterative solver. In the z direction, a much larger N_z would be required to observe a similar effect, owing to the $\mathcal{O}(N_z N_x^2)$ complexity for the MVP of the preconditioner.

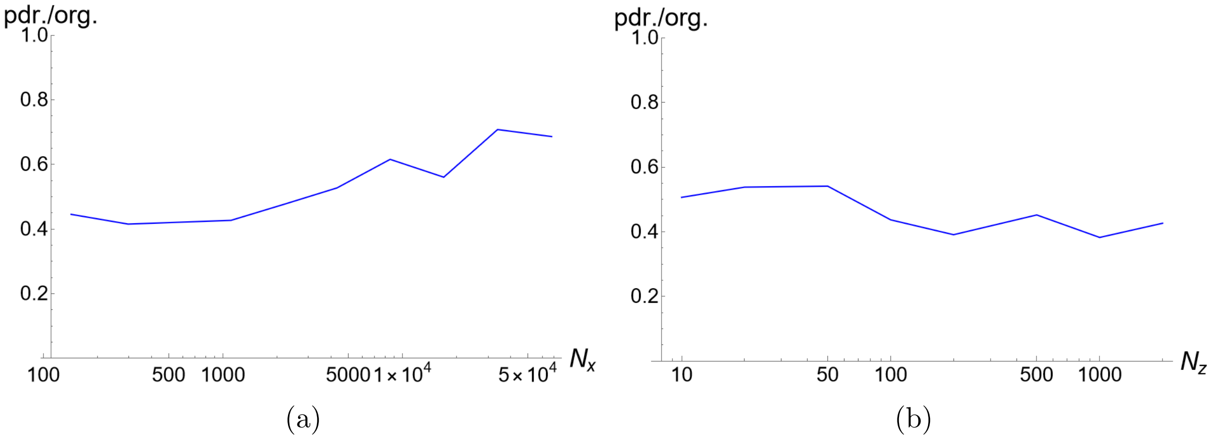


Figure 6.7: Comparison of computation time between systems ”org” and ”pdr”. (a) different discretization in the x direction. (b) different discretization in the z direction.

In Fig. 6.7 we compare the total solution time for the original system and the preconditioned system. The vertical axes denote the total solution time in seconds. The horizontal axes represent the discretization parameters N_x and N_z in Fig. 6.7 (a) and (b), respectively. Both figures suggest that in most cases (except for the cases with extremely large N_x) the total solution time can be reduced by a factor larger than 2, which is corresponding to the gained reduction factor in terms of the number of iterations for the $\chi = 8$ case in Table 6.3. For other cases in Table 6.3, the reduction factor in computation time is expected to be comparable to the reduction factor in terms of the number of iterations, since the computation time per iteration in Fig. 6.6 (a) and (b) is independent of the contrast χ.

We conclude that the total solution time can be reduced by applying the NVF-BD preconditioner. We also observe that the reduction in computation time for the preconditioned system gets lowered for large N_x cases due to the computational complexity of the preconditioner.

6.4.2 Case (B): a 2D TM metal grating problem

In the second case, a grating device made of aluminium is embedded in air and supported by an aluminium half-space, see Fig. 6.8. A plane wave with wavelength 700 nm is incident under an angle of 22.9° with respect to the z-axis and the incident electric field $\mathbf{E}^i$

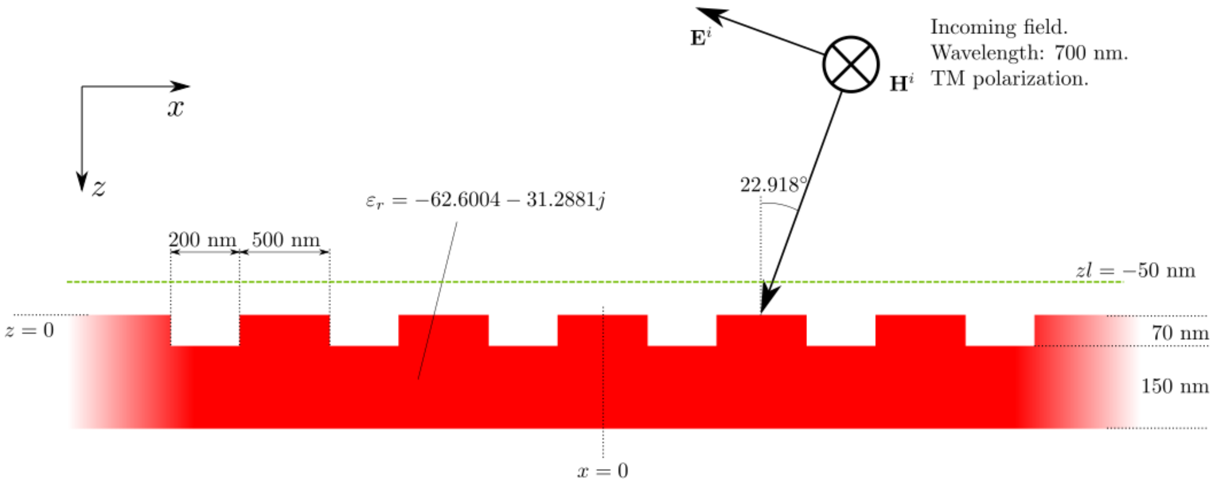


Figure 6.8: Geometry setting of a 2D TM metal grating problem.

is polarized in the xz plane. The relative permittivity of aluminum is $-63.6004-31.2881j$. The air in the six grooves are considered as the scatterers, which have negative contrast $\chi = -1.0108 + 0.0064j$. Table 6.2 case (B) displays the discretization parameters we used in this simulation. Notice that in total 81 Gabor frame functions are used for each Gabor window length X, which yields a resolution of 5 nm in the x direction. In the z direction PWL functions are employed with sampling distance $\Delta = 2.5$ nm. Our goal in this example is to demonstrate the effectiveness of the NVF-BD preconditioner as compared with the original system in terms of convergence with an acceptable relative error. Also, in this 2D TM case, computing the NVF-BD preconditioner requires only moderate memory requirements due to the relatively low-dimensional block matrix C_i and we compute C^{-1} directly based upon the Doolittle LU factorization.

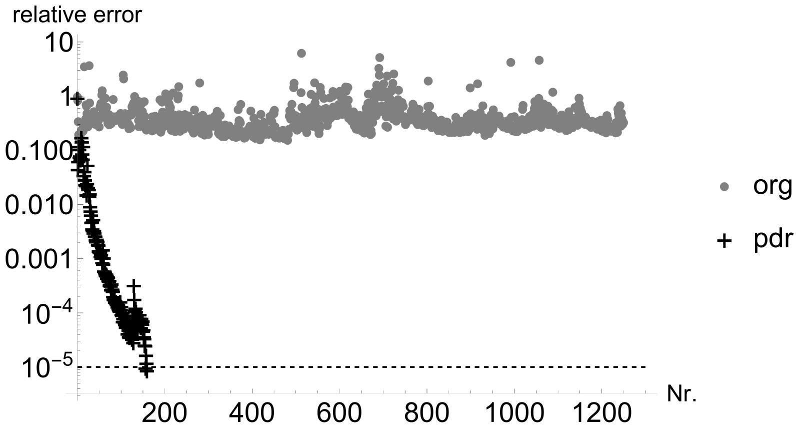


Figure 6.9: Iteration details for the negative permittivity case in Fig. 6.8 with $\chi = -1.01 + 0.0064j$. The horizontal dashed line denotes the desired accuracy goal $1 \cdot 10^{-5}$.

Fig. 6.9 shows the convergence of the relative error versus the number of iterations for the original system and for the preconditioned system. It is clear that without applying the NVF-BD preconditioner, the system does not converge, while the preconditioned system reaches the desired relative error of 10^{-5} in 159 iterations.

We have validated the preconditioned system's solution against the commercial FEM code JCMWave [198]. Fig. 6.10 (a) presents the x-component of the total electric field E_x, where the red line denotes the JCMWave reference and the blue dashed lines denote the solution from the preconditioned system, so the solutions can be compared. Fig. 6.10 (b) displays the absolute error between the solution from the preconditioned system in the near field for $z = -50$ nm, just above the upper interface. One can observe that some high-frequency Gibbs ringings occur near the grooves' boundaries, where the contrast function is discontinuous. Gibbs phenomena can be the dominant contribution to the error in the near field, just as observed in the original system [108,110], but it does not propagate over a long distance. Figures (a) and (b) together suggest that the solution obtained from the preconditioned system matches the reference well.

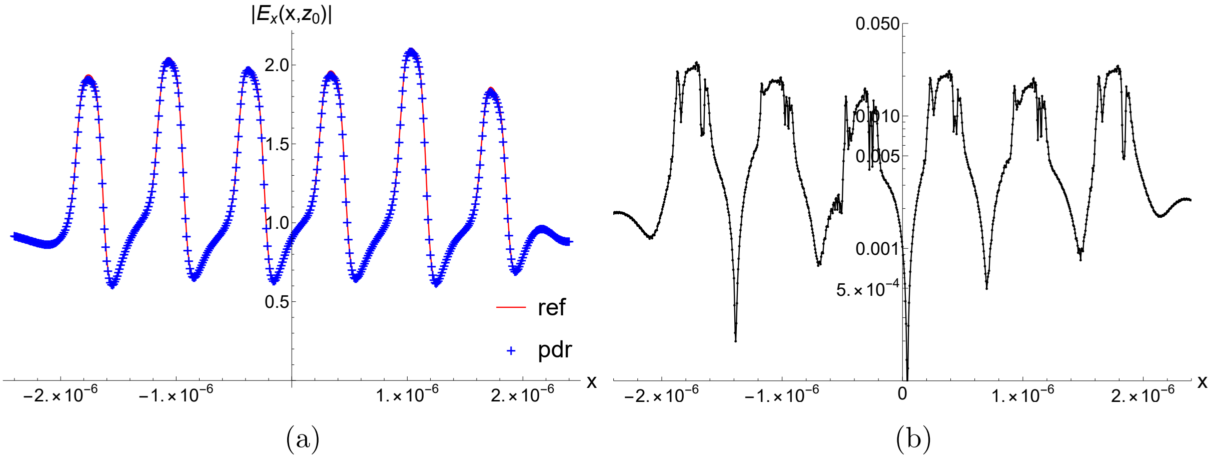


Figure 6.10: The electric fields for the case in Fig. 6.8. In (a) both reference and the solution $E_x(x, z_0)$ from the preconditioned system are displayed at $z_0 = -50$ nm, indicated by the horizontal green line in Fig. 6.8. In (b) the absolute error in $E_x(x, z_0)$ between the solutions from the preconditioned system and the JCMWave reference is displayed on a log scale.

6.4.3 Case (C): a 3D high contrast problem

In the third case we consider a bar-shaped scatterer with a high relative permittivity $\varepsilon_r = 17$ in free space. The scatterer's dimensions are 300 × 200 × 100 nm. The incident plane wave is characterized by the Cartesian wavevector $\mathbf{k} = (0, 0, k_0)$, with the electric field polarized in the x direction and with unit amplitude. The plane wave has a wavelength of $\lambda = 425$ nm. The geometry setting is given in Fig. 6.11. Table 6.2, case (C) displays the discretization parameters that are used in this simulation. Note that the frequency modulation number $-10 \leq n_x, n_y \leq 10$ and therefore there are 21 frame functions used per Gabor window length X and Y, which yields a resolution of 3.86 nm in both x and y directions. In the z direction PWL functions are employed with sample distance $\Delta = 5$ nm.

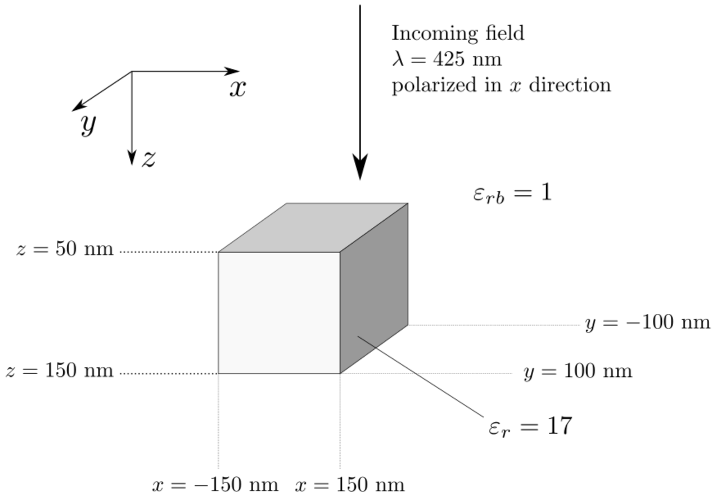


Figure 6.11: 3D scattering problem for a bar-shaped scatterer with relative permittivity $\varepsilon_r = 17$ embedded in air.

Usually the dimension of the system equation in 3D cases is huge. In this simulation there are $2.3 \cdot 10^6$ unknowns after performing the discretization based on Table 6.2 case (C), and the dimension of the block submatrices C_i in Fig. 6.2 is 1.0×10^5. Therefore it is unrealistic to store the full submatrix C_i due to its excessive memory requirement. As an alternative strategy we have implemented the preconditioned system such that the extra MVP for C^{-1} is executed based on an inner iterative solver. We also use the BiCGstab(2) algorithm in the inner iterative process and this inner iterative process is terminated once a relative error of less than or equal to 10^{-15} is reached. The inner iterative solver takes much fewer MVPs than the outer solver. However, this double-iterative method should be improved in future work, to make the entire solution process more efficient. Therefore, we focus on the effect of the NVF-BD preconditioner on the reduction in the number of iterations, instead of computation time, in this 3D case.

Fig. 6.12 shows the evolution of the relative error versus the iteration count for the original system and the preconditioned system. It is clear that the preconditioned system outperforms the original system in this 3D high contrast problem with $\chi = 16$. The preconditioned system takes 454 iterations to reach the required relative error with a relatively fast rate of convergence. However, the original system failed to converge to the desired relative error within 1250 iterations. Notice that, from iteration 550 to 1250, the residual vector of the original system gained less than 1-digit accuracy. This is a clear example that shows how the NVF-BD preconditioner can reduce the number of iterations in a 3D high-contrast case.

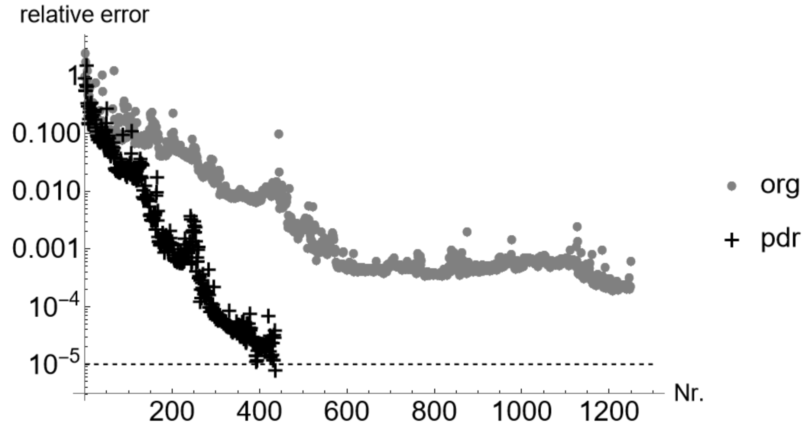


Figure 6.12: Iteration details for the high contrast case with $\chi = 16$. The dashed line denotes the default accuracy goal 1×10^{-5}.

6.5 Conclusion

We proposed a normal-vector-field-based block-diagonal (NVF-BD) preconditioner for the original system of a spatial spectral solver with Gabor discretization for 2D TM polarization and 3D cases. The block-diagonal structure of the matrix that incorporates the normal-vector field formulation and previous work motivated us to apply this preconditioner to this spatial spectral Maxwell solver. We observed a more clustered eigenvalue distribution after applying this NVF-BD preconditioner, which is a good sign in the sense of expecting a reduction in the number of iterations. The NVF-BD preconditioner is either computed via a direct LU decomposition, in the 2D TM cases, or performed via an inner iterative procedure in the 3D problem.

We tested this NVF-BD preconditioner on three types of problems: (A) a 2D TM scattering problem with high contrast values and large geometry size, (B) a 2D TM metal grating problem, and (C) a 3D high contrast problem. The numerical experiments reveal that the number of iterations can be significantly reduced by applying the NVF-BD preconditioner, which therefore extends the capability of the original spatial spectral solver to cases with higher contrast, negative permittivity, or larger geometrical dimension. Computation-time analysis shows that the total solution time can also be reduced after applying the NVF-BD preconditioner, even though the reduction effect can be dampened when a large number of transverse basis functions N_x is used, due to the extra MVP for the preconditioner with $\mathcal{O}(N_z N_x^2)$ computational complexity. The proposed NVF-BD preconditioner itself can readily benefit from parallel computing, since the NVF-BD preconditioner has the same per-z-sample block-diagonal structure as the matrices C and M. However, a similar speed increase due to parallelization at the z-sampling level will not readily obtained for the original system due to the communication overhead associated with the Green function, for which many z-samples need to be combined.

Acknowledgment

This work was funded by NWO-TTW as part of the HTSM program under project number 16184. The authors thank Artur Palha from ASML Netherlands B.V. for providing the references based on JCMWave. The authors are grateful to the anonymous reviewers for their helpful comments and suggestions.

Chapter 7

Applications of the spatial spectral Maxwell solver

7.1 Scattering by single-pad resist-only metrology targets

7.1.1 Introduction

An optical scatterometer is an important optical sensor and it has been widely used in optical wafer metrology [17]. After the resist layer has been exposed and developed (see e.g. the 5th step in Fig. 1.2), an important step is to measure the uniformity of the critical dimension (CDU) and overlay of the printed patterns. This is because a large critical dimension (CD) variation (e.g. CDU is larger than 10% of the CD) or a large overlay (e.g. overlay is larger than 30% of the CD) significantly degrades the performances of the manufactured device on the wafer [17]. The measurement of CDU and overlay in optical metrology can be performed by optical scatterometry.

In Fig. 7.1 we show a simplified conceptual sketch of a sensor model for wafer metrology. Incoherent or partially coherent light is processed by an optical system and then illuminates a repetitive metrology target. The scattered light is collected by a set of optical elements and captured by a charge-coupled device (CCD) camera for intensity and spectrum measurements. More detailed schematic diagrams of a scatterometer for CD and overlay metrology can be found in Figure 10 and Figure 11 in [17].

The blue dashed box in Fig. 7.1 indicates an electromagnetic scattering process, which can therefore be modeled by a Maxwell solver. In real wafer metrology applications, this step is often performed by a periodic Maxwell solver, which implies the assumption that the geometry has a 2D periodicity in the directions perpendicular to the direction of the stratification of the background layers. On the other hand, the spatial spectral Maxwell solver has been developed for aperiodic scattering problems [108–111, 199, 223], and therefore it can also be used for solving this finite scattering problem. Therefore, even with the same excitation, the scattered electric field obtained from an aperiodic solver can differ from its counterpart in a periodic solver due to the absence of a periodicity assumption in an aperiodic solver. Hence, a natural idea, and also the goal of this section, is to use the spatial spectral Maxwell solver to simulate and solve this type of scattering problem and detect the difference with respect to the results obtained from a periodic solver. A comparison between the aperiodic solver and a periodic solver should yield insights into what kind of role the spatial spectral Maxwell solver can play in a real optical scatterometry application.

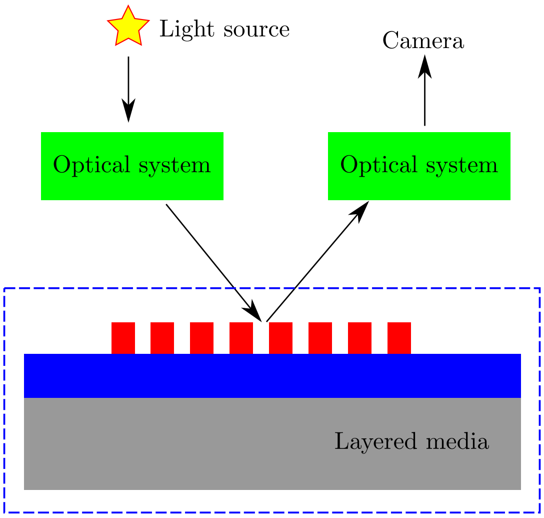
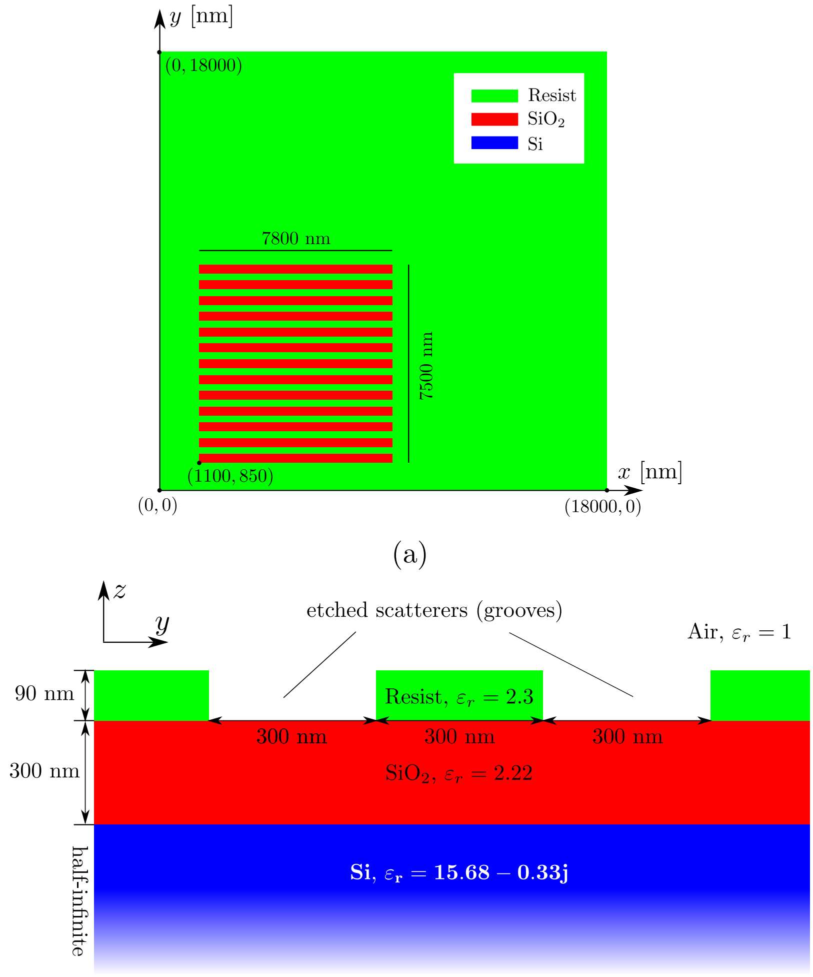


Figure 7.1: Conceptual sketch of an optical metrology sensor. A repetitive pattern (in red) is illuminated by incoherent or partially coherent light processed by an optical system. The reflected light is also processed by an optical system and the spectrum is measured by a camera. The blue dashed box indicates a scattering problem that can be solved by a Maxwell solver.

The structure of this section is as follows. In Section 7.1.2, we state the scattering problem of a single-pad resist-only metrology target by giving the geometry configuration and the incident electric field. In Section 7.1.3, we outline the solution strategy before using the spatial spectral method. Section 7.1.4 contains two numerical examples with an obliquely incident plane wave excitation and a beam excitation. The results from the spatial spectral Maxwell solver are compared with an external reference from a periodic solver. We give the conclusions in Section 7.1.5.

7.1.2 Statement of the problem

We consider a single-pad resist-only metrology grating target consisting of 13 grooves in the top layer of a layered medium. The grooves are bar-shaped and they are considered to be the scatterers. The dimension of each scatterer is $7800 \times 300 \times 90$ nm, and the distance between two subsequent scatterers is 300 nm. In Fig. 7.2 (a) we show the location of all scatterers with a Cartesian coordinate system in the xy-view. Note that the background medium is assumed to be of infinite extent in the transverse directions, and Fig. 7.2 (a) only shows the finite domain $[0, L_D] \times [0, L_D]$ with $L_D = 18000$ nm. In the longitudinal direction, the top layer of the layered medium consists of a 90 nm high resist, on top of a 300 nm high silicon dioxide layer, and a silicon substrate occupying the bottom half-space. In Fig. 7.2 (b) shows the stratification and the relative permittivities of the materials, for the illumination wavelength $\lambda_0 = 550$ nm. Additionally, we define the top of the resist layer as $z = 0$, and the z direction is pointing upwards.



(b)

Figure 7.2: A single-pad resist-only metrology grating target and a layered medium. (a) xy-view of the scatterers. Note that all the scatterers are in the first quadrant of the given coordinate system, and the center of the scatterer domain is at $(5000, 4600)$ nm. (b) yz-view, with each layer's material and relative permittivity for the free-space wavelength $\lambda_0 = 550$ nm.

To have a full statement of the scattering problem, the excitation must also be provided, apart from the geometry configuration. As an example, we consider an excitation with a single plane wave coming from the upper half-space. Given a standard Cartesian coordinate system with Cartesian unit vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$, we assume a plane wave with wavelength $\lambda_0 = 2\pi/k_0$ coming in from the upper half space as visualized in 7.2. The wave vector specifying the direction of propagation of the plane wave is denoted by $\mathbf{k}$ and we define $\hat{\mathbf{k}} = \mathbf{k}/k_0$. Further, $\mathbf{k}_T$ be the transverse part of $\mathbf{k}$ in the xy plane. The wave vector $\mathbf{k}$ of this plane wave is determined by a polar angle θ and an azimuthal angle ϕ defined with respect to the z and x axis, respectively. In Fig. 7.3 we show an example of an oblique plane-wave incidence with corresponding incident angles.

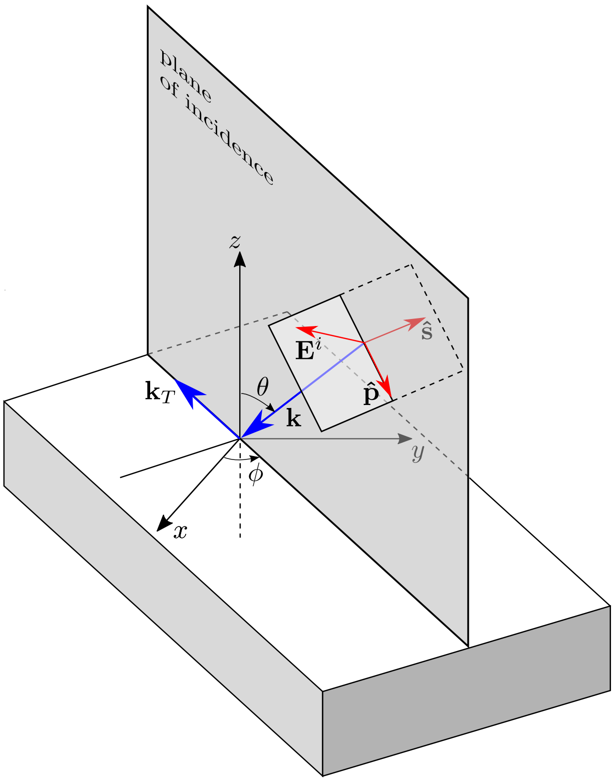


Figure 7.3: The $s - p$ polarization of an incident plane wave $\mathbf{E}^i$ with wave vector $\mathbf{k}$.

Assuming an $e^{j\omega t}$ time convention, the incident electric field $\mathbf{E}^i$ in the upper half space, which is associated with the plane-wave excitation, can be represented by:

$$\begin{aligned} \mathbf{E}^i(\mathbf{x}) &= \left(E_s^i \hat{\mathbf{s}} + E_p^i \hat{\mathbf{p}}\right) e^{-j\mathbf{k}\cdot\mathbf{x}}, \\ &= \left(E_s^i \hat{\mathbf{s}} + E_p^i \hat{\mathbf{p}}\right) e^{-j\mathbf{k}_T\cdot\mathbf{x}_T} e^{-jk_z z} \end{aligned} \tag{7.1}$$

where

$$\begin{aligned}\mathbf{k} =& -k_0 \sin\theta\cos\phi\hat{\mathbf{x}} - k_0\sin\theta\sin\phi\hat{\mathbf{y}} - k_0\cos\theta\hat{\mathbf{z}} \\ =& \mathbf{k}_T + k_z\hat{\mathbf{z}} \quad \text{and} \quad \mathbf{k}_T = k_x\hat{\mathbf{x}} + k_y\hat{\mathbf{y}},\end{aligned} \tag{7.2}$$

$$\hat{\mathbf{s}} = \frac{\hat{\mathbf{k}}\times\hat{\mathbf{z}}}{\|\hat{\mathbf{k}}\times\hat{\mathbf{z}}\|}, \quad \hat{\mathbf{p}} = \hat{\mathbf{k}}\times\hat{\mathbf{s}}, \quad \mathbf{x}_T = x\hat{\mathbf{x}} + y\hat{\mathbf{y}}. \tag{7.3}$$

Here, two auxiliary unit vectors $\hat{\mathbf{s}}$ and $\hat{\mathbf{p}}$ are defined on the plane perpendicular to the wave vector $\mathbf{k}$, such that the polarization of the incident electric field $\mathbf{E}^i$ is fixed. The plane containing both $\mathbf{k}$ and $\hat{\mathbf{z}}$ is called the plane of incidence. Clearly, $\hat{\mathbf{s}}$ is normal to the plane of incidence, and $\hat{\mathbf{p}}$ is contained in the plane of incidence. Furthermore, E_s^i and E_p^i are the independent complex amplitudes of $\mathbf{E}^i$ along the $\hat{\mathbf{s}}$ and $\hat{\mathbf{p}}$ directions, respectively. Mathematically, the wave vector $\mathbf{k}$ in Eq. (7.2) can be considered as a parameter, and both $\hat{\mathbf{s}}$ and $\hat{\mathbf{p}}$ are fixed once $\mathbf{k}$ is given. Therefore, the incident electric field $\mathbf{E}^i(\mathbf{x})$ is uniquely determined by the complex amplitudes E_s^i and E_p^i according to Eq. (7.1).

The s and p polarization definitions of the incident electric field in Eq. (7.1) yield a removable singularity for modes that are close to normal incidence where $\theta = 0$. To avoid this inconvenience, we define

$$\hat{\mathbf{x}}_p = -\sin\phi\,\hat{\mathbf{s}} + \cos\phi\,\hat{\mathbf{p}}, \tag{7.4}$$

$$\hat{\mathbf{y}}_p = \cos\phi\,\hat{\mathbf{s}} + \sin\phi\,\hat{\mathbf{p}}. \tag{7.5}$$

The unit vectors $\hat{\mathbf{x}}_p$ and $\hat{\mathbf{y}}_p$ are essentially an improper rotation of $\hat{\mathbf{s}}$ and $\hat{\mathbf{p}}$ in the plane perpendicular to $\mathbf{k}$ and they approach the directions $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ respectively when θ is approaching 0. Consequently, the incident electric field $\mathbf{E}^i$ can be represented as

$$\mathbf{E}^i(\mathbf{x}) = \left(E_{xp}^i\hat{\mathbf{x}}_p + E_{yp}^i\hat{\mathbf{y}}_p\right)e^{-j\mathbf{k}_T\cdot\mathbf{x}_T}e^{-jk_z z}, \tag{7.6}$$

for all $\mathbf{x}$ in the upper half space. Hence, for a given plane wave with wave vector $\mathbf{k}$, the incident electric field $\mathbf{E}^i(\mathbf{x})$ is uniquely determined by the complex amplitudes E_{xp}^i and E_{yp}^i and the auxiliary unit vectors $\hat{\mathbf{x}}_p$ and $\hat{\mathbf{y}}_p$.

A realistic illumination pattern consists of multiple mutually independent and incoherent beams and each of these beams needs to be simulated separately to obtain a complete image formation in an optical metrology sensor. To generate individual beams expressed in terms of a plane-wave spectrum, we sample the $\mathbf{k}_T$ spectrum of the plane waves in a uniform manner. The incident electric field of the scattering problem described in Fig. 7.2, after uniform sampling, is defined in the plane $z = 0$ through

$$\mathbf{E}^i(\mathbf{k}_T, z=0) = E_{xp}^i(\mathbf{k}_T)\hat{\mathbf{x}}_p(\mathbf{k}_T) + E_{yp}^i(\mathbf{k}_T)\hat{\mathbf{y}}_p(\mathbf{k}_T), \tag{7.7}$$

for all $\mathbf{k}_T \in \Lambda$, and Λ is a uniform grid in $\mathbf{k}_T$-space defined as

$$\Lambda = \left\{(k_x, k_y)\in\mathbb{R}^2 | (k_x, k_y) = (k_x^c + l_1\cdot\Delta_k, k_y^c + l_2\cdot\Delta_k)\right\}. \tag{7.8}$$

Here we denote by $(k_x^c, k_y^c) = (0, 0)$ the center of this spectral grid, and $\Delta_k = k_0\lambda_0/L_D$ is the grid spacing, with L_D the length of the computational domain in the xy plane, see Fig. 7.2 (a). Additionally, $-L_1 \leq l_1 \leq L_1$, $-L_2 \leq l_2 \leq L_2$ for positive integers L_1 and L_2.

Throughout the rest of this section, we will focus on the following single-pad resist-only metrology scattering problem: given the geometry configuration in Fig. 7.1, a spectral grid defined in (7.8), and corresponding complex amplitudes E_{xp}^i, E_{yp}^i in Eq. (7.7), compute the scattered electric field at the top of the resist layer on the same spectral grid Λ, i.e., $\mathbf{E}^s(\mathbf{k}_T, z = 0)$ for all $\mathbf{k}_T \in \Lambda$.

7.1.3 Solution strategy

We now discuss the preprocessing steps to use the spatial spectral Maxwell solver. The spatial spectral method requires an incident electric field represented in the spatial domain. Therefore, the first preprocessing task is to transform the incident electric field given in Eq. (7.7) into the following form

$$\mathbf{E}^i(\mathbf{x}_T, z_n) = E_x^i(\mathbf{x}_T, z_n)\hat{\mathbf{x}} + E_y^i(\mathbf{x}_T, z_n)\hat{\mathbf{y}} + E_z^i(\mathbf{x}_T, z_n)\hat{\mathbf{z}}, \tag{7.9}$$

where $z_n \in [z_a, z_b]$ are sample points in the PWL discretization, z_a and z_b bound all the scatterers in the longitudinal direction. To achieve this goal, we first represent the incident electric field in Eq. (7.7) in Cartesian components as

$$\mathbf{E}^i(\mathbf{k}_T, z = 0) = E_x^i(\mathbf{k}_T)\hat{\mathbf{x}} + E_y^i(\mathbf{k}_T)\hat{\mathbf{y}} + E_z^i(\mathbf{k}_T)\hat{\mathbf{z}}, \tag{7.10}$$

for all $\mathbf{k}_T \in \Lambda$, where the complex amplitudes $E_x^i(\mathbf{k}_T), E_y^i(\mathbf{k}_T), E_z^i(\mathbf{k}_T)$ can be obtained readily, once $\hat{\mathbf{x}}_p$ and $\hat{\mathbf{y}}_p$ in Eq. (7.6) are expressed in terms of $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$.[1] In a subsequent step, the incident electric field $\mathbf{E}^i(\mathbf{k}_T, z = 0)$, for all $\mathbf{k}_T \in \Lambda$, needs to be propagated to the top position of the scatterers at $z = z_a$. Later on, the propagation of the incident electric field from $z = z_a$ to all $z_n \in [z_a, z_b]$ is performed by applying a Fourier transformation on a deformed integration manifold, as introduced in [109, 110], and by computing the tensorial transmission coefficients as in [24]. The detailed process of extending Eq. (7.10) to Eq. (7.9) can be found in Appendix A.

While transforming from Eq. (7.10) to Eq. (7.9), we also need to represent the discretely sampled incident electric field in the spectral domain by Gabor frames. The Gabor-based spectral-domain grid spacing (or resolution) is determined by

$$\Delta'_{k,x} = \frac{K_x}{\beta_x \cdot (2M_x + 1)}, \tag{7.11}$$

where $K_x \in \mathbb{R}$ is the spectral Gabor window length, $\{-M_x, \ldots, M_x\}$ is the range of the translation index m_x, and β_x is one of the oversampling parameters. Analogously, we have $\Delta'_{k,y}$.

[1] Since the z-direction is pointing downwards in the Cartesian coordinate system defined in [24], the coordinates of a vector (a_1, a_2, a_3) represented in Fig. 7.3 gets coordinates $(a_1, a_2, -a_3)$ in the coordinate definition in the spatial spectral solver.

From Eq. (7.7) and (7.8), we observe that the incident electric field is expressed in terms of $(2L_1 + 1)(2L_2 + 1)$ planar waves of the same frequency but with different polarization and propagation direction. When transferring all the information to the spatial spectral Maxwell solver, the best way is to choose the Gabor-based grid Λ' the same as the given grid Λ for the plane-wave samples. By doing so, we can use the exact plane waves in the spatial spectral solver, without artificially creating interpolation artifacts.

The second preprocessing task is therefore to determine the Gabor coefficients such that $\Delta'_{k,x} = \Delta'_{k,y} = \Delta_k$. This can be done[2] by choosing a suitable pair of K_x and M_x according to Eq. (7.11) and set $K_y = K_x$ and $M_y = M_x$. In Fig. 7.4 (a) we give an example where the Gabor-based spectral grid Λ' matches the plane-wave spectral grid Λ for the incident electric field.

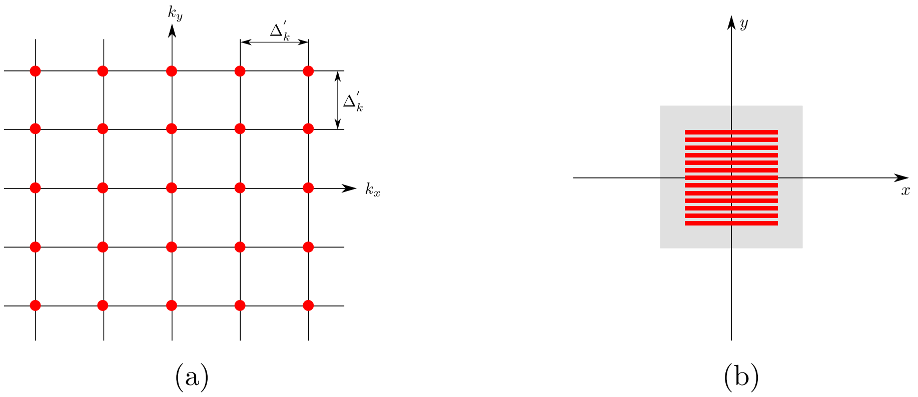


Figure 7.4: (a) An example of two spectral grids having the same spacing. The red dots represent the original grid defined in Eq. (7.8) with spacing Δ_k. The black lines are used to represent the Gabor-based spectral grid with spacing $\Delta'_{k,x} = \Delta'_{k,y} = \Delta'_k$. (b) The scatterers are shifted to a location that is symmetric about the origin and the gray area depicts the new computational domain.

The third preprocessing task concerns the computational domain. The spatial spectral method [24] works on a symmetric computational domain in the transverse plane, but from Fig. 7.2 (a) we can easily see that the grating targets are not symmetric about the origin, which results in a much larger symmetric computational domain, which reduces the efficiency of the solver. To over overcome this problem, we spatially shift the scatterers with the translation vector $\mathbf{a} = (-5000, -4600)$ nm to make them symmetric about the origin. Additionally, we apply the same spatial shift to the incident electric field, but now performed in the spectral domain as

$$\mathbf{E}^i_{\text{sym}}(\mathbf{k}_T, z = 0) = \mathbf{E}^i(\mathbf{k}_T, z = 0)e^{-j\mathbf{k}_T \cdot \mathbf{a}}. \tag{7.12}$$

[2] Other requirements for the Gabor parameters must also be satisfied: the spatial Gabor window length relies on its spectral counterpart and the combination of the spatial window length and the range of the spatial shift index should be such that a sufficiently large computational domain is covered. The choices of M_x, M_y, N_x, and N_y are restricted by the application of the FFT-based fast operations [111] and the resolutions in both spatial and spectral domains should be fine enough.

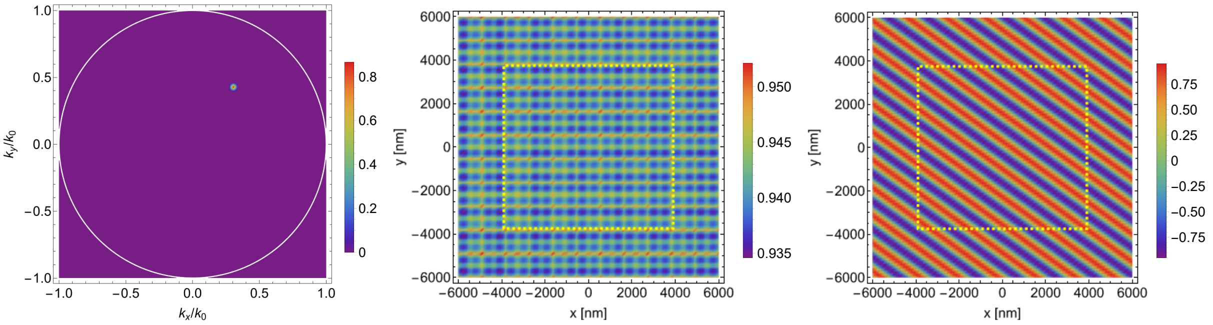
In practice, the above spatial shift is performed immediately after the Cartesian representation in Eq. (7.10) and before the propagation steps in Appendix A. After shifting the scatterers and the original incident electric field, a much smaller computational domain is obtained for the spatial spectral Maxwell solver, see Fig. 7.4 (b), which implies less computational effort.

7.1.4 Numerical results

We now give two examples of the scattering problem stated in Section 7.1.2: Example 1 is with a single plane-wave excitation, Example 2 is with a beam excitation. The Gabor parameters used in these two examples are given in Table 7.1. Following the notations used in Section 2.3.2, T_x, T_y represent the spatial Gabor window lengths, K_x, K_y represent the spectral Gabor window lengths, m_x, m_y are the spatial translation indices for the Gabor frames, and n_x, n_y are the frequency modulation indices for the Gabor frames. N_z represents the number of the PWL functions used in the z direction. Oversampling parameters are denoted by α_x, α_y and β_x, β_y. The resulting resolution is 36.4 nm in the transverse plane and 30 nm in the longitudinal direction. The total number of unknowns in the pertaining matrix system is 2.94×10^6.

Table 7.1: Discretization parameters used in both Example 1 and Example 2.

T_x, T_y	K_x, K_y	m_x, m_y	n_x, n_y	N_z	α_x, α_y	β_x, β_y
1.47×10^{-6}	4.28×10^{6}	$-7:7$	$-16:16$	4	$\sqrt{2/3}$	$\sqrt{2/3}$



(a) (b) (c)

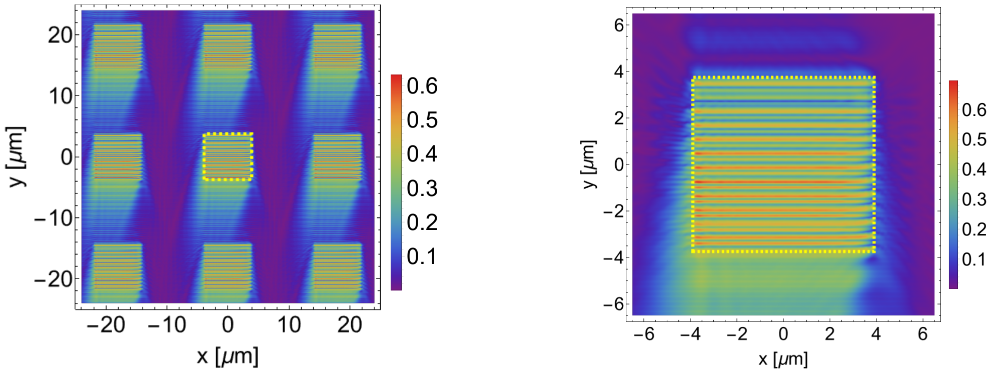
Figure 7.5: Incident electric field in the plane $z = 0$. (a) $\|\mathbf{E}^i(k_x, k_y)\|$, intensity in the spectral domain. (b) $|E_x^i(x, y)|$. (c) $\mathrm{Re}[E_x^i(x, y)]$. Note that the yellow dashed box in (b) and (c) indicates the location of the shifted scatterers.

In Example 1, we consider a single plane-wave excitation with normalized wave vector $\hat{\mathbf{k}} = (0.31, 0.43, -0.85)$, and the complex amplitudes of the incident electric field are $E_{xp}^i = 0.71 + 0.71j$ and $E_{yp}^i = 0$. After representing $\mathbf{E}^i(\mathbf{x})$ according to Eq. (7.9) and performing the spatial shift according to Eq. (7.12), we get

$$\mathbf{E}^i(\mathbf{x}_T, z = 0) = (0.12 - 0.94j, -0.009 + 0.07j, -0.04 + 0.3j)e^{-j\mathbf{k}_T \cdot \mathbf{x}_T}, \tag{7.13}$$

for all $\mathbf{x}_T$ is in the $z = 0$ plane, and $\mathbf{k}_T = (3.4 \times 10^6, 4.89 \times 10^6)$. In Fig. 7.6 we plot the incident electric field in the plane $z = 0$. Fig. 7.5 (a) shows the incident electric field of (7.12) is a dot within the Ewald circle in the spectral domain. Fig. 7.5 (b) shows the absolute value of $E_x^i(x, y)$ in the spatial domain. Here we observe that $|E_x^i(x, y)|$ is more or less a constant but not completely. This is because a Dirac delta function in the spectral domain requires an infinite number of Gabor coefficients (see Eq. (A.3) in Appendix A). Therefore, truncation to a finite number of coefficients results in a non-constant amplitude in the spatial domain. Fig. 7.5 (c) shows the real part of $E_x(x, y)$.

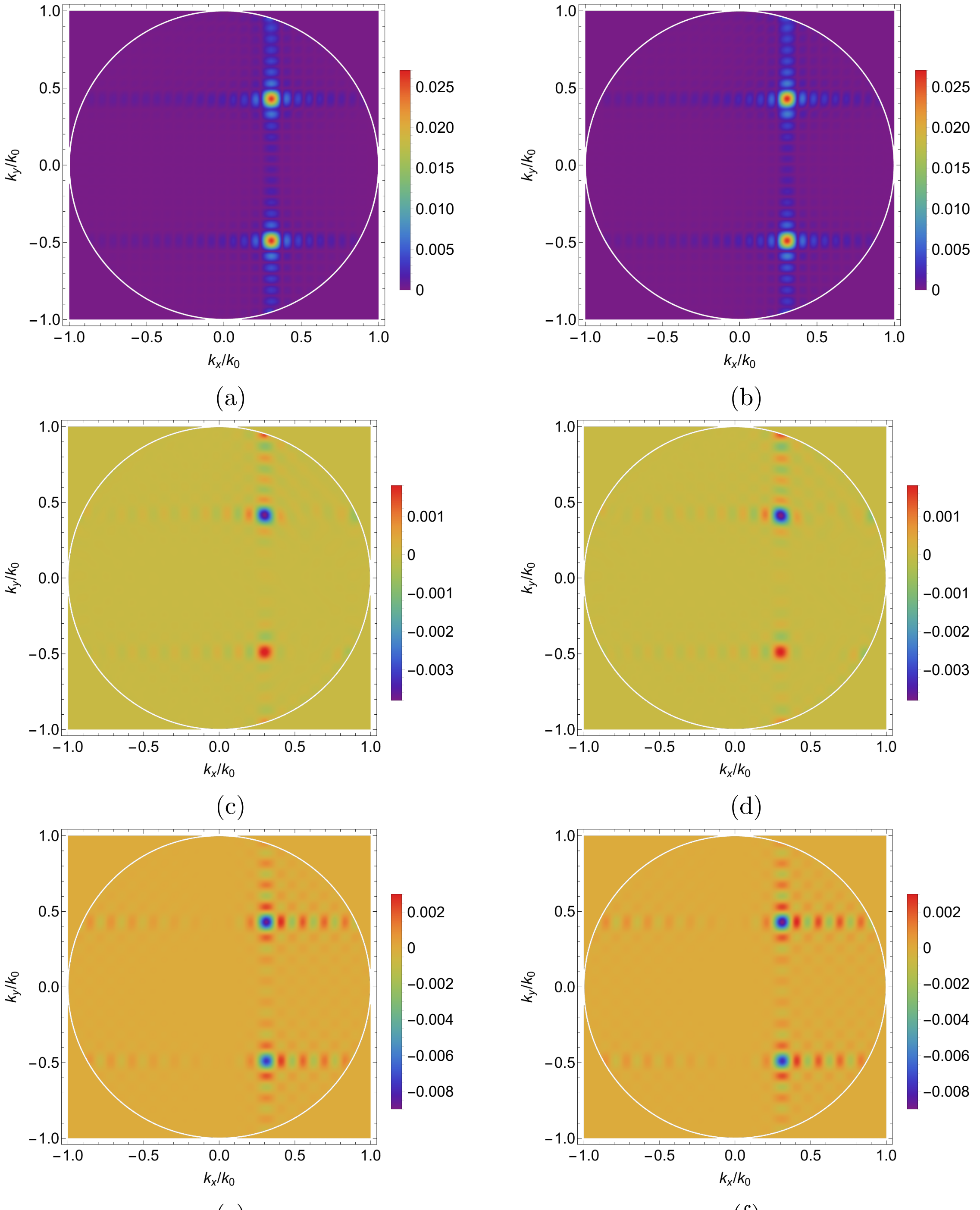
With the discretization parameters given in Table 7.1, we obtain the solutions of Example 1 by the spatial spectral Maxwell solver. We then compare the solutions with an external reference computed based on the periodic assumption. The periodic solver assumes a 2D periodicity in 3D space with the periodic directions orthogonal to $\hat{\mathbf{z}}$. The 18000 × 18000 nm geometry shown in Fig. 7.1 (a) can be seen as the unit cell that is repeated to fill the entire xy plane. The intensity of the scattered electric field of the reference (denoted by $\mathbf{E}^s_{\text{ref}}(x, y)$) is shown in Fig. 7.6 (a) on a linear color scale. Furthermore, we show the field intensity of the scattered electric field by the spatial spectral method in Fig. 7.6 (b). Note that the plotting domain in (b) is smaller than the domain in (a).



(a) (b)

Figure 7.6: Intensity of the scattered electric fields at $z = 0$ in Example 1. (a) $\|\mathbf{E}^s_{\text{ref}}(x, y)\|$, based on the periodic assumption (reference). (b) $\|\mathbf{E}^s(x, y)\|$ based on the spatial spectral method. All figures are on a linear color scale, and the yellow dashed box denotes the location of the scatterers.

From Fig. 7.6 (a) and (b), one observes that the scattered field $\mathrm{E}^s_{\text{ref}}(x, y)$ propagates to the adjacent periodically repeated scatterers, while in Fig. 7.6 (c) there is no other scatterer. This slight difference in the spatial domain is essentially caused by the difference between a periodic solver and an aperiodic solver, which might cause differences in the far field as well. This is an indication of the differences in the approximation of an aperiodic scattering problem by a periodic solver with the supercell approach.



(a) (b) (c) (d) (e) (f)

Figure 7.7: Comparison of the far-field solutions at $z = 0$ in Example 1. The reference results are displayed in the left column and the results from the spatial spectral Maxwell solver are displayed in the right column. (a) and (b): $|E_x^s(k_x, k_y)|$. (c) and (d): real part of $E_y^s(k_x, k_y)$. (e) and (f): imaginary part of $E_z^s(k_x, k_y)$.

In order to quantitatively compare the far-field results, we had to rescale the results from the spatial spectral method by an empirically determined scaling constant $c_{\text{ff}} = 2.33k_0^2\Delta_k^2/(4\pi^2)$, to match the scaling factor of the periodic solver[3]. We now show the far-field results in Fig. 7.7. Note that only the fields within the Ewald circle in the spectral domain have been shown. The absolute differences in the far-field solutions computed by the spatial spectral finite solver are shown in 7.8, in the spectral domain.

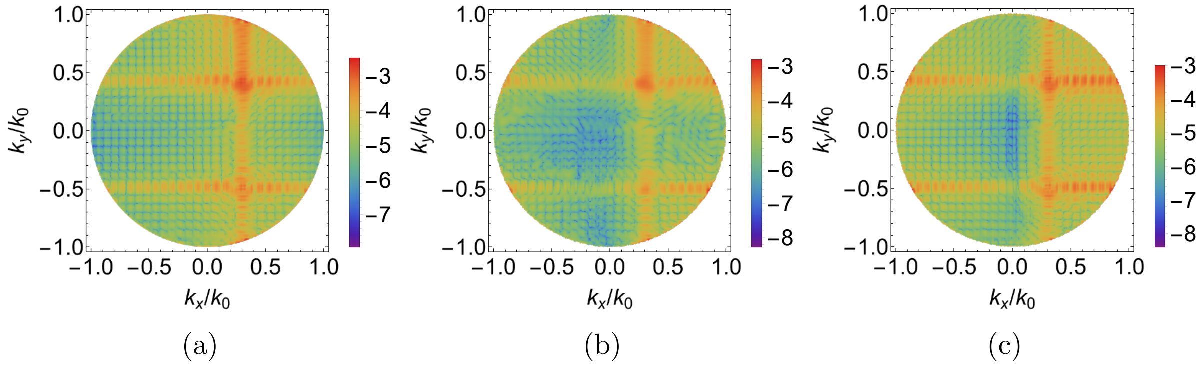


Figure 7.8: Absolute differences in the far-field results. (a) $E_x^s(k_x, k_y)$, (b) $E_y^s(k_x, k_y)$ and (c) $E_z^s(k_x, k_y)$, obtained from the spatial spectral Maxwell solver. All on a $\log_{10}$ color scale.

To show the similarities in the far-field results, we defined the following relative differences. Let the vectors $\mathbf{v}_x$, $\mathbf{v}_y$ and $\mathbf{v}_z$ contain the scattered electric fields $E^s_{\text{ref},x}(k_x, k_y)$, $E^s_{\text{ref},y}(k_x, k_y)$ and $E^s_{\text{ref},z}(k_x, k_y)$ evaluated on the spectral grid Λ, respectively. Let $\mathbf{u}_x$, $\mathbf{u}_y$, and $\mathbf{u}_z$ be the reference data from the periodic solver on the same grid Λ. Clearly, $\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z, \mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z \in \mathbb{C}^{32400}$, where $32400 = 180 \times 180$ is the number of sample points contained in the grid Λ. We define the following relative difference for $E_x^s(k_x, k_y)$:

$$r_x = \frac{\|\mathbf{u}_x - \mathbf{v}_x\|}{\|\mathbf{u}_x\|}, \tag{7.14}$$

where $\|\cdot\|$ is the ℓ^2 norm for a vector. Analogously, we define r_y and r_z. Furthermore, we define the following weighted relative difference:

$$r_w = w_x \cdot r_x + w_y \cdot r_y + w_z \cdot r_z, \tag{7.15}$$

where $w_i = \|u_i\|/(\|u_x\| + \|u_y\| + \|u_z\|)$ for $i = x, y, z$. The relative differences of the solutions obtained by the spatial spectral Maxwell solver and the periodic reference are: $r_x = 0.05$, $r_y = 0.13$, $r_z = 0.04$, and $r_w = 0.058$.

We now study Example 2, the scattering problem stated in Section 7.1.2 with a beam excitation. We use the same discretization parameters, given in Table 7.1. After repeating the preprocessing steps as described in Section 7.1.3, we show the incident electric field of Example 2 in Fig. 7.10 in both the spectral domain and the spatial domain. From

[3] Unfortunately, the exact mathematical expression of the scaling factor of the periodic solver was unknown. More validation data is needed to determine the origin of this scaling factor.

Fig. 7.9 (a) we observe that the beam-type excitation contains a set of plane waves with different propagating directions. The field intensity of $\|\mathbf{E}^i(x,y)\|$ is shown in Fig. 7.9 (b). Note that only a part of the scatterer is illuminated by the beam. Fig. 7.9 (c) shows the real part of $E^i_x(x,y)$.

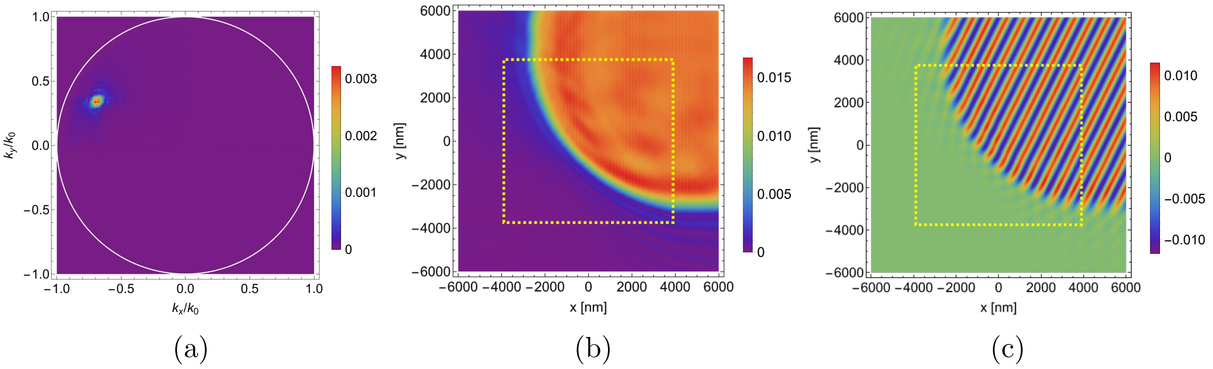


Figure 7.9: Incident electric field of Example 2 at $z = 0$. (a) $\|\mathbf{E}^i(k_x,k_y)\|$, intensity in the spectral domain. (b) $\|\mathbf{E}^i(x,y)\|$. (c) $\mathrm{Re}[E_x(x,y)]$. Note that the yellow dashed box in (b) and (c) indicates the location of the shifted scatterers.

The scattered electric fields are shown in Fig. 7.10. The reference from a periodic solver is displayed in (a), where we observe a similar behavior of the scattered electric field within the layered medium. The solution from the spatial spectral solver is displayed in (b).

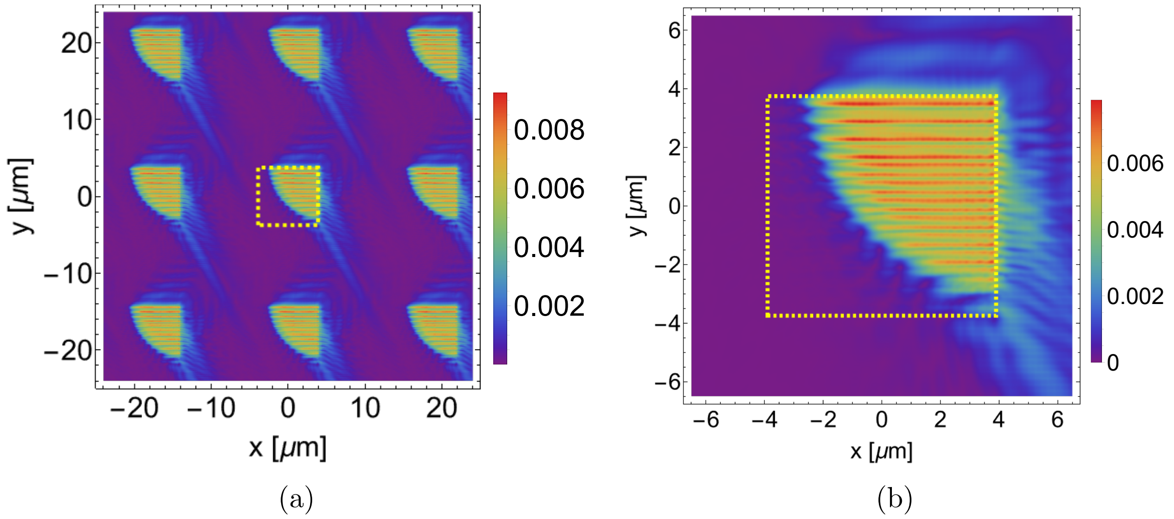
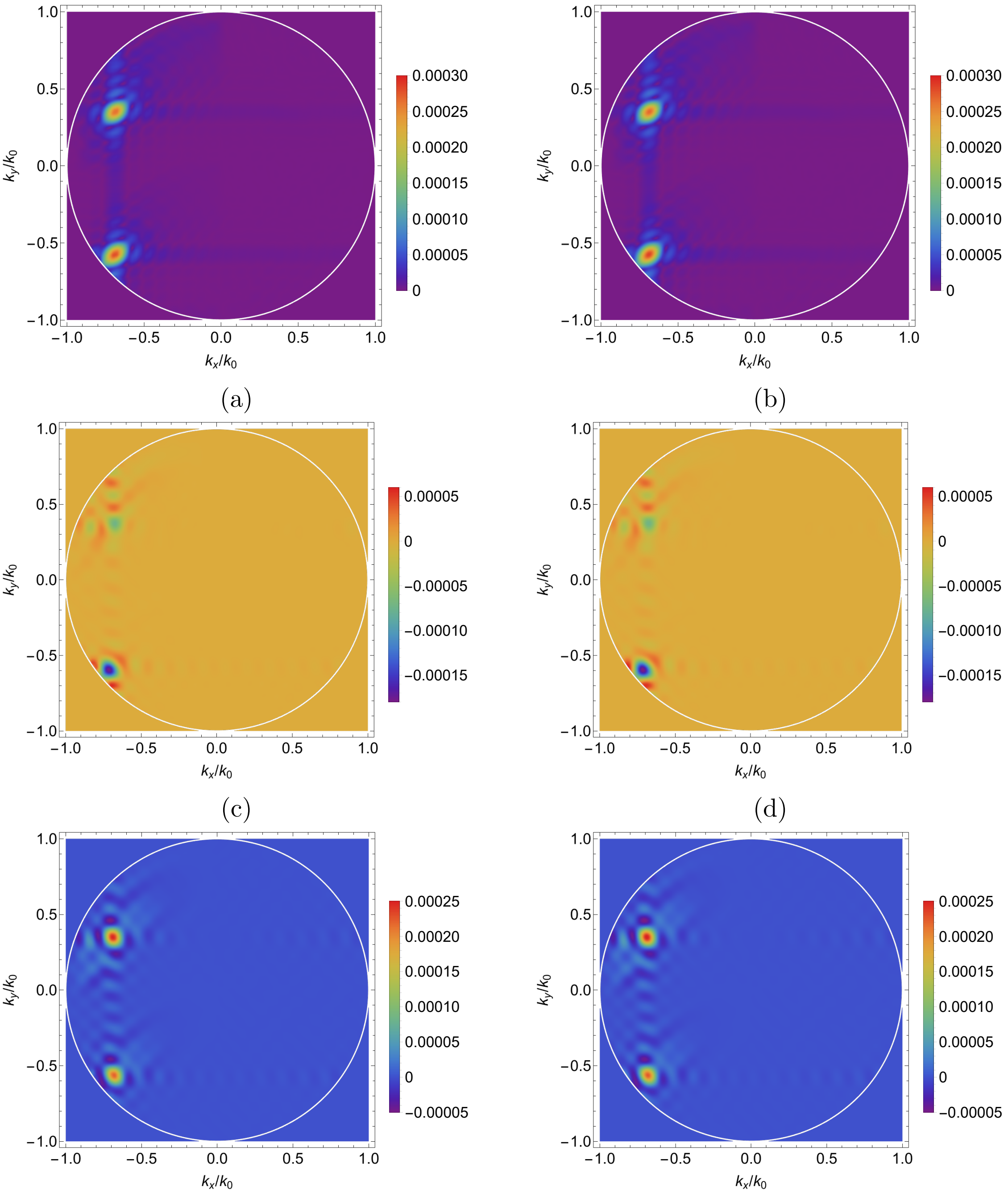


Figure 7.10: Amplitude of the scattered electric fields at $z = 0$ in Example 2. (a) $\|\mathbf{E}^s_{\mathrm{ref}}(x,y)\|$, based on the periodic assumption (reference). (b) $\|\mathbf{E}^s(x,y)\|$ based on the spatial spectral method. All figures are on a linear color scale, and the yellow dashed box denotes the location of the scatterers.



(a)

(b)

(c)

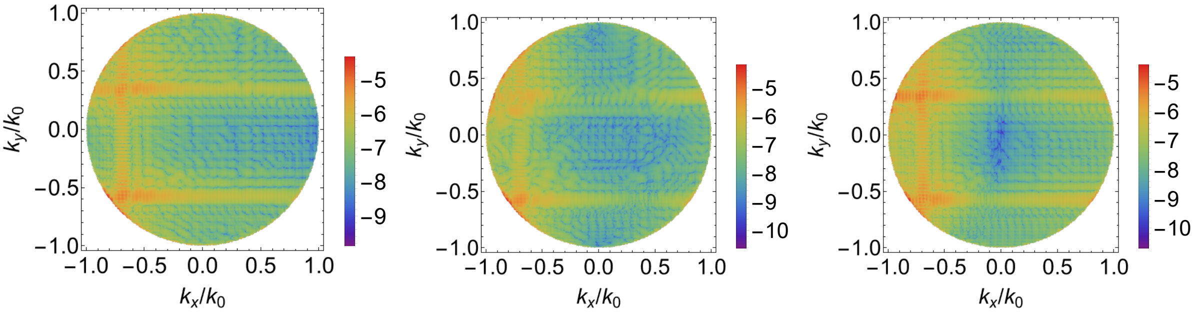
(d)

(e)

(f)

Figure 7.11: Comparison of far-field solutions at $z = 0$ in Example 2. The reference results are displayed in the left column and the results from the spatial spectral Maxwell solver are displayed in the right column. (a) and (b): $|E_x^s(k_x, k_y)|$. (c) and (d): real part of $E_y^s(k_x, k_y)$. (e) and (f): imaginary part of $E_z^s(k_x, k_y)$.

In Fig. 7.11 we compare the far fields from the spatial spectral Maxwell solver, using the same scaling constant c_{ff}, with the reference, by showing the absolute values of $E_x^s(k_x, k_y)$, the real parts of $E_y^s(k_x, k_y)$ and the imaginary parts of $E_z^s(k_x, k_y)$. Note that the color ranges are set the same in each comparison.



(a)

(b)

(c)

Figure 7.12: Absolute differences of (a) $E_x^s(k_x, k_y)$, (b) $E_y^s(k_x, k_y)$ and (c) $E_z^s(k_x, k_y)$ in Example 2.

The absolute differences in the far-field solutions computed by the spatial spectral finite solver are shown in 7.12, on a $\log_{10}$ color scale. The relative differences computed based on Eqs (7.14) and (7.15) are: $r_x = 0.014$, $r_y = 0.022$, $r_z = 0.012$, and $r_w = 0.015$.

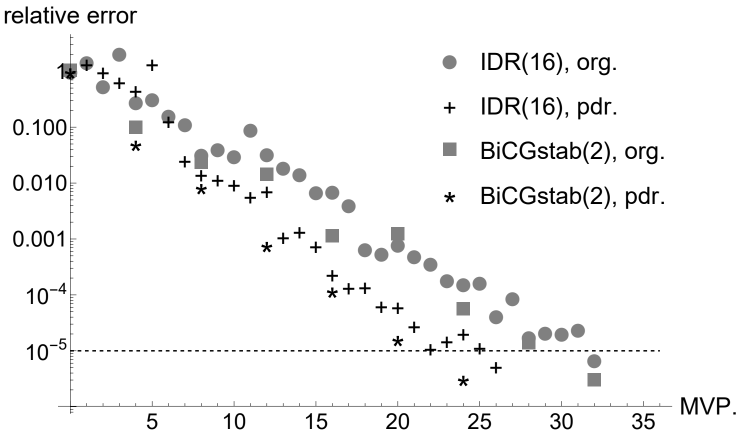


Figure 7.13: Iteration details of IDR(16) and BICGstab(2) in Example 2, where “org” stands for the original unpreconditioned system, and “pdr” indicates that the NVF-BD preconditioner proposed in Chapter 6 has been used.

We complete this subsection by discussing the computational cost. All computations were performed on a dual 20-core Intel(R) Xeon(R) Gold 6148 CPU at 2.40 GHz with 755 GB RAM. Here we focus on the beam-excited Example 2, and the recorded initialization time was 162 seconds and the solution time was 160.5 seconds for IDR(16). Further, 33 MVPs were counted in the iterative process to reach the desired relative error of 1.0×10^{-5}.

Besides using IDR(16), we also solved the same scattering problem with BICGstab(2). BICGstab(2) achieved a solution time of 167.4 seconds and required 32 MVPs to attain the same desired relative error. Consequently, both IDR(16) and BICGstab(2) take about 5 seconds per MVP in this example. The details of the iterative process are shown in Fig. 7.13.

We now consider the same scattering problem of the 13 grooves, but now the excitation consists of an incoherent sum of 9031 beams. Each beam is specified by a different set of complex amplitudes E^i_{xp} and E^i_{yp}, on the same grid Λ in the spectral domain according to Eq. (7.8). Due to the incoherence, we have to compute each beam individually and consequently the total number of simulations is 9031. We compute $\sim$ 45 simulations simultaneously on a dual 20-core Intel(R) Xeon(R) Gold 6148 CPU at 2.40 GHz with 755 GB RAM.

After computing, we obtain the far-field electric fields $\mathbf{E}^s(k_x, k_y, z = 0)$ for the 9031 beams, and they are all represented by complex amplitudes on the plane-wave sample grid Λ and decomposed in terms of the unit vectors $\hat{\mathbf{x}}_p$ and $\hat{\mathbf{y}}_p$, defined in Eqs (7.4) and (7.5). To simulate the optical system in Fig. 7.1, the data set containing all the complex amplitudes for the reflected light has been processed with proprietary internal software by ASML Netherlands B.V.. This software simulates the optical configuration of YieldStar, an optical metrology tool commercialized by ASML, the working principle of which can be found in [224] and [225]. As a result, three simulated camera images are generated showing the three lowest diffraction orders of this grating structure as detected by YieldStar. These images are shown in Fig. 7.14.

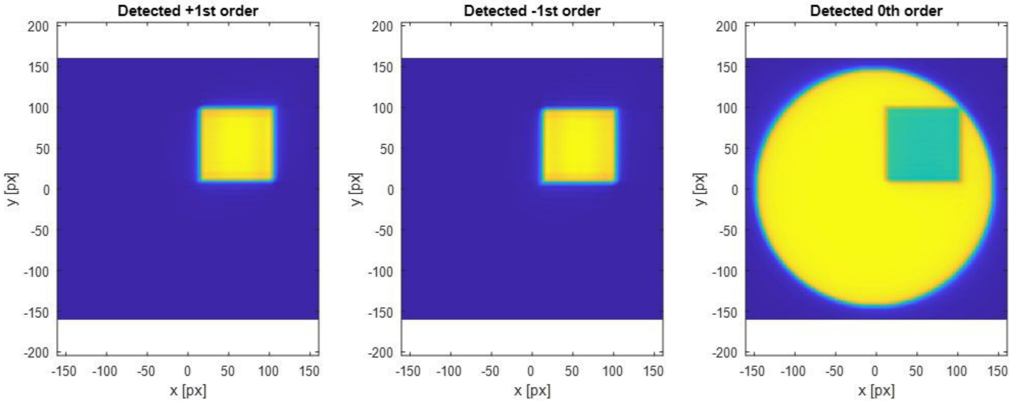


Figure 7.14: Detected diffraction orders of the grating structure specified in Fig. 7.2. Left: the $+1^{\text{st}}$ diffraction order. Middle: the -1^{st} diffraction order. Right: the 0^{th} diffraction order. For all figures, the horizontal axis represents the x-direction and the vertical axis represents the y-direction. The square regions indicate the grating area containing the 13 scatterers. The yellow disc in the right figure represents the size of the spot that results from the specular reflection of the incident light composed of the incoherent summation of all the 9031 incident beams.

7.1.5 Conclusion on the single-pad resist-only metrology target

We have applied the spatial spectral Maxwell solver developed in [24] and solved a single-pad resist-only metrology target scattering problem. The solution strategy was aimed at using the spatial spectral method efficiently. We then studied two examples with plane-wave excitation and beam excitation and compared the far-field results with a solution obtained from a periodic solver. In the plane-wave incidence example, we reached a relative difference of 5.8% in the far field. In the beam-excitation example, the average relative difference was 1.5% and the computation time was about 6 minutes. These results suggest that the spatial spectral Maxwell solver has the ability to solve similar type of scattering problems accurately and efficiently, and can hence contribute to the optical metrology sensor model for non-periodic metrology targets. For future work, it is recommended to formulate additional physics-based interpretations for comparison and conduct experiments in order to further understand the differences between the spatial spectral Maxwell solver and the periodic solver.

Acknowledgements

The results in this section are based on a collaboration with ASML. We acknowledge Frank Buijnsters (ASML Netherlands B.V.) for the continuous support to the project, the frequent discussions, the help with data preparation and for setting up the interface with the internal ASML optical modeling software. We also acknowledge Filippo Alpeggiani (ASML Netherlands B.V.) for help in generating Fig. 7.14.

7.2 Simulating a computer-generated waveguide hologram scattering problem with an artificial 2D Gaussian beam source[4]

We study a computer-generated waveguide hologram scattering problem. An artificial current density function with Gaussian profile is defined on a source plane that is close to the scatterers. A Gaussian beam field is induced based on the artificial current density function, and can be controlled to approximate the original incident field. The computer-generated waveguide hologram scattering problem is then solved with a spatial spectral method.

7.2.1 Introduction

A method of designing a large-area Computer-Generated Waveguide Hologram (CGWH) with a long working distance is given in [226]. Subsequently, an integrated CGWH is presented to emit or receive optical beams in free space [227]. An accurate and efficient Maxwell solver is necessary for the design and optimization before the fabrication, since it can be computationally expensive for large-area devices such as grating couplers. Here we consider solving a 3D CGWH scattering problem numerically with the spatial spectral method developed in [110, 111].

In Section 7.2.2 we present a scheme to artificially generate an inner-layer Gaussian beam (GB) source, which yields the possibility to approximate the real incident field by optimizing the beam parameters. We show numerical results in Section 7.2.3 and conclude in Section 7.2.4.

7.2.2 Methodology

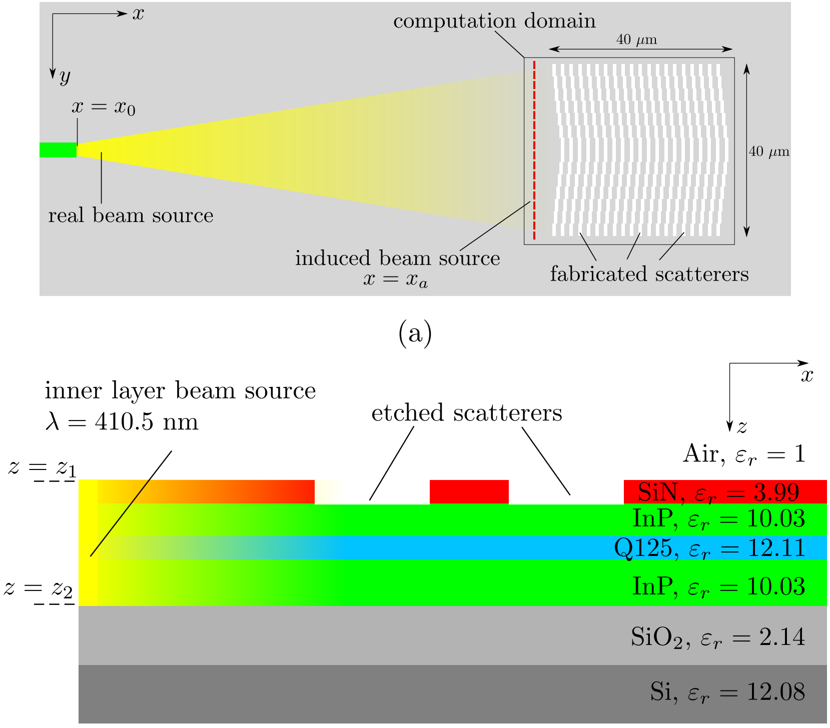
7.2.2.1 Statement of the problem

We consider a similar CGWH scattering problem as in [226], but with a smaller hologram area being 40 × 40 μm. Fig. 7.15 (a) shows the layout of hologram area and the light source in the xy plane. The CGWH has a grating structure and it contains 10126 bar-type scatterers in total. All scatterers have uniform lengths in the y and z directions. In the x direction, the lengths of the scatterers vary and they are designed specifically to project a focused beam into free space [227].

Far from the hologram area, an optical input is released from a single-mode waveguide. The input light has a vacuum wavelength $\lambda_0 = 1300$ nm and an inner-layer wavelength $\lambda_1 = 410.5$ nm. The incoming light is guided within a high-contrast indium phosphide (InP) layer. The hologram scatterers have a uniform height since they are all embedded in a layer above the InP slab waveguide. Fig. 7.15 (b) shows the details of the multi-layered

[4] This section was published as [223]

medium in the xz view. To reduce the computational cost, we will only simulate the CGWH structure and approximate the source by a GB.



(a)

(b)

Figure 7.15: Geometry setup: (a) layout of the CGWH and the sources in xy view. The dashed line represents the location where the artificial current is defined. (b) layered medium in xz view with each layer's material and the relative permittivity under $\lambda_0 = 1300$ nm.

7.2.2.2 Induced Gaussian beam source

Inspired by the method used in [228, Chapter 3], we define an artificial current density function $\mathbf{J}^{GB}(x, y, z)$ in a Gaussian form that is centered at (x_0, y_0, z_0), see Fig. 7.15 (a). We make the following assumptions: (1) the fundamental mode is the only propagating mode in the multi-layered medium, and (2) this mode can be excited by a current that is uniform in the z direction. Therefore we use a 2D Gaussian function as a current source generate the Gaussian beam. Let $\mathbf{J}^{GB}(x, y, z) = (0, 0, J_z^{GB})^T$, and define:

$$J_z^{GB}(x, y, z) = A(x, y)P(x, y)\delta(x - x_a)\Pi_{[z_1, z_2]}(z), \tag{7.16a}$$

$$A(x, y) = J_0 \frac{w_0}{w(x)} \exp\left\{-\frac{(y - y_0)^2}{w(x)^2}\right\}, \tag{7.16b}$$

$$P(x, y) = \exp\left\{-\frac{jk(y - y_0)^2}{2R(x)} - jk(x - x_0)\right\}. \tag{7.16c}$$

Note that $A(x, y)$ represents an amplitude factor, where J_0 is the initial amplitude, w_0 is the beam waist, $x_R = \pi w_0^2/\lambda$ is the Rayleigh length, and $w(x) = w_0\sqrt{1 + (x - x_0)^2/x_R^2}$ is the beam radius at x. $P(x, y)$ in (7.16) is a phase factor, where $R(x) = (x-x_0)\left\{1 + x_R^2/(x - x_0)^2\right\}$ is the radius of curvature of the phase front. The longitudinal phase is ignored since it vanishes after a long propagation distance. In (7.16a), $\delta(x-x_a)$ is the Dirac delta function to generate a pulse current on x_a (see Fig. 7.15 (a)), and in practice it is approximated by a Gaussian function. $\Pi_{[z_1,z_2]}(z)$ is a rectangular pulse function to bound the infinite 2D GB to the computation domain only.

Then one can compute the scattered electric field by taking the pre-defined $\mathbf{J}^{GB}$ as a generating function:

$$\mathbf{E}^{GB}(x, y, z) = \mathcal{F}_T^{-1}\left\{\int_{\mathbb{R}} G(k_x, k_y, z, z')\mathcal{F}_T[\mathbf{J}^{GB}(x, y, z')]dz'\right\}, \tag{7.17}$$

where $\mathcal{F}_T, \mathcal{F}_T^{-1}$ denote 2D Fourier transformations, and G is the spectral-domain Green operator in the multi-layered medium [110]. In Fig. 7.16 we show an example of the induced electric field based on a 2D GB source defined within the InP waveguide in Fig. 7.15 (a). The beam waist is $w_0 = 700$ nm and the beam center is $(-23.5, 0, 0.16)$ μm.

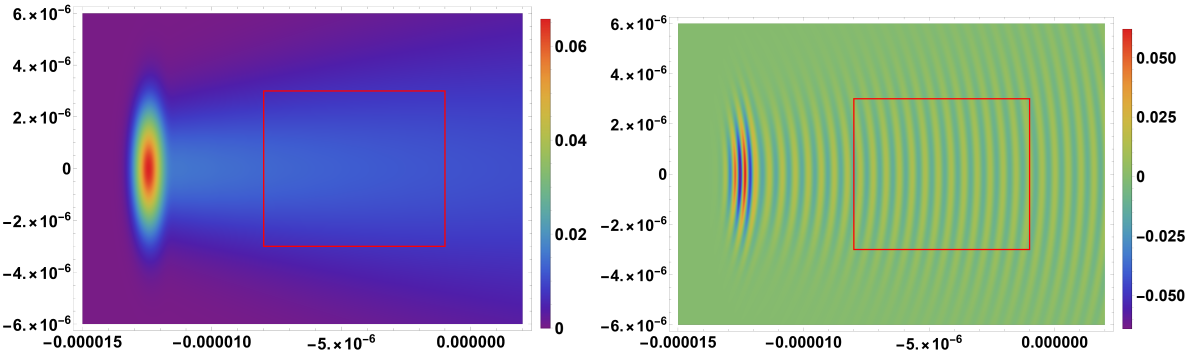
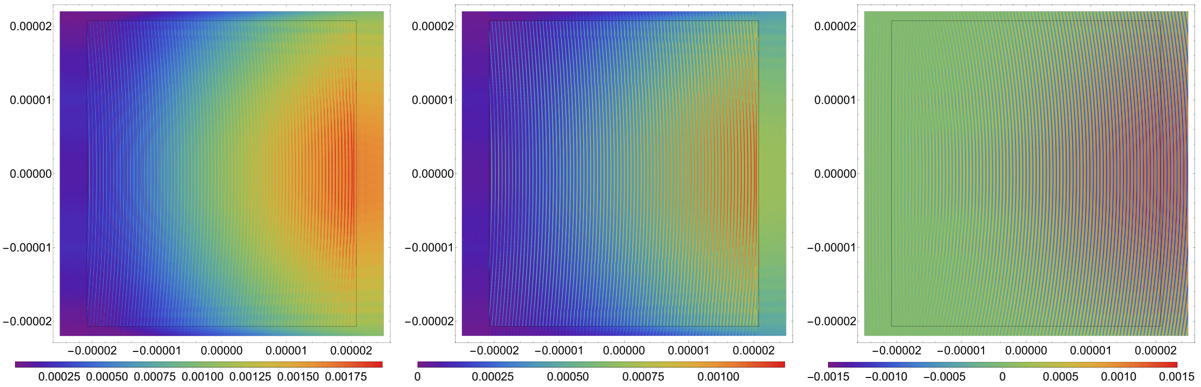


Figure 7.16: Induced 2D GB and the domain of scatterers (red box). Left: $|E_x^{GB}(x, y)|$. Right: $E_x^{GB}(x, y)$, real part.

7.2.3 Numerical Results

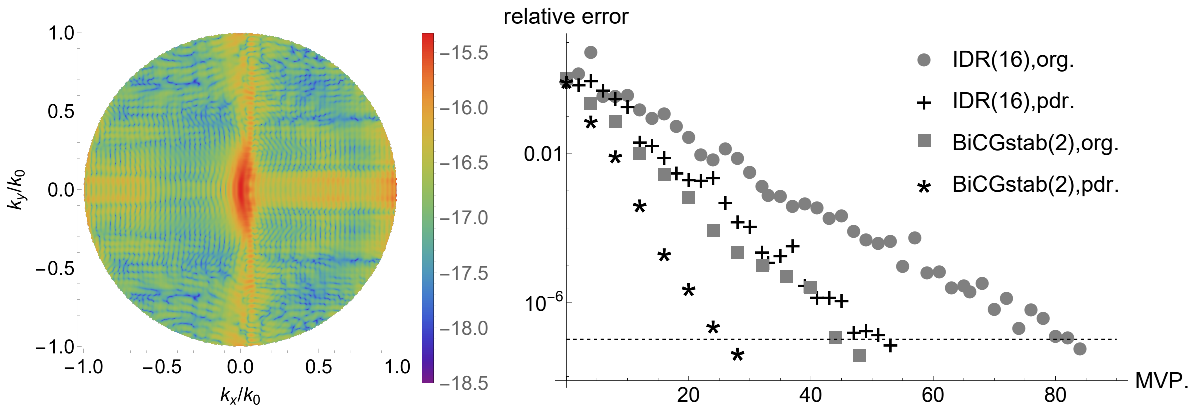
We regard the induced scattered field $\mathbf{E}^{GB}$ in (7.17) as the new incident field, i.e., $\mathbf{E}^i = \mathbf{E}^{GB}$, and test the performance of the spatial spectral method [110, 111] on this CGWH scattering problem. The beam waist is set to $w_0 = 600$ nm. The computation domain is 47.8×42.9 μm, and discretization parameters determine a resolution of 27.9 nm in the transverse plane and 16.5 nm in the z direction. Convergence study on the far field shows a relative error of 10^{-3} under these settings.

We show the near-field scattered electric field in Fig. 7.17. Fig. 7.18 (a) shows far-field intensity in the spectral domain above the top interface. The distribution of the far field represents the total electric field after leaving the CGWH into free space, and it suggests a beam with a dominant direction in the xz plane. Fig. 7.16 (b) shows convergence performances of iterative methods IDR(16) and BiCGstab(2), with (pdr.) or without (org.) preconditioning, respectively. With BiCGstab(2) and a NVF-BD preconditioner [199], only 28 matrix-vector products are required to reach the desired relative error.



(a) (b) (c)

Figure 7.17: $\mathbf{E}^s(x,y)$ at $z = 22$ nm: (a) field intensity. (b) real part of $E_z^s(x,y)$. (c) imaginary part of $E_z^s(x,y)$.



(a) (b)

Figure 7.18: (a) $k_0^2 \cdot \|\mathbf{E}(k_x,k_y)\|$ on a $\log_{10}$ scale within the Ewald circle, where $k_0 = 2\pi/\lambda_0$. (b) Convergence comparisons.

7.2.4 Conclusion

A CGWH scattering problem with multi-layered medium is simulated via the spatial spectral method. A GB incidence is induced based on an artificial current function. Near-field and far-field solutions are shown and discussed. The fast convergence suggests the spatial spectral method can be particularly useful for the CGWH design and similar large-area devices.

7.3 Simulating a metasurface scattering problem

7.3.1 Introduction

Dielectric metasurfaces based on dielectric resonators have interesting characteristics and have been shown an effective tool for light manipulation [229–231]. Conventionally, wavefront shaping is performed by designing the curvature of the surfaces in diffractive optics. However, with a metasurface, variation in the alignment and geometrical details of the subwavelength structures in the transverse directions can manipulate the phase, amplitude, and polarization of the electromagnetic waves in a highly controllable manner, and the dimension of the metasurface is reduced as compared to conventional optical designs [232, 233]. For example, a robust design scheme is given in [234] for single-layer metasurface lenses based on dielectric resonators. A dielectric resonator metasurface usually contains high-contrast scatterers to generate strongly resonant responses [233, 235], which yields difficulties when using a full-wave simulation. Naturally, an efficient Maxwell solver can contribute to the design process.

In this section, we use the spatial spectral Maxwell solver to simulate a single-layer dielectric resonators metasurface lens. In Section 7.3.2 we present the geometry configuration of the metasurface lens and give the discretization parameters used in the spatial spectral method. We show the near-field results and the iteration details in Section 7.3.3. The challenges when solving this metasurface case and concluding remarks are discussed in Section 7.3.4.

7.3.2 Methodology

We consider a metasurface with multiple circular-cylindrical scatterers positioned on a substrate and surrounded by the air. Fig. 7.19 (a) shows the top view of the layout of a case with 13×13 scatterers in the x and y directions. The scatterers are made of amorphous silicon, while the substrate material is $\mathrm{SiO_2}$. The side view of this metasurface is displayed in Fig. 7.19 (b). Note that all the cylindrical scatterers have a uniform height of 800 nm, but a non-uniform diameter ranging from 161 nm to 257 nm. A normally incident plane wave with the electric field polarized along the x-direction is impinging from the bottom substrate and travels along the z direction. The free-space wavelength of the incident plane wave is 1000 nm. Relative permittivities of the scatterers and the substrate are shown in Fig. 7.19 (b) as well, and the consequent contrast of these cylinder-shaped scatterers is $\chi = 12.1769$.

In particular, we consider 4 cases of the metasurface scattering problem as described in Fig. 7.19 with different numbers of scatterers. Table 7.2 shows the discretization parameters used for these 4 cases. Note that the $\ell \times \ell$, for some positive integer ℓ, represents a case with ℓ scatterers per row and ℓ scatterers per column. Therefore, Fig. 7.19 (a) corresponds to the case 13×13. T_x, T_y in Table 7.2 denote the Gabor window lengths in the spatial domain,

m_x, m_y are the spatial shift numbers, n_x, n_y are the spatial frequency modulation numbers. Note that the spatial shift numbers are increased to generate a larger computational domain when more scatterers are taken into account. In the z direction, N_z represents the number of PWL functions used. The oversampling parameters are $\alpha_x = \alpha_y = \sqrt{2/3}$, and $\beta_x = \beta_y = \sqrt{2/3}$. The total number of unknowns for each case is given in the last column, based on Eq. (2.78). Additionally, the discretization parameters, as listed in Table 7.2, imply a transverse resolution of about 14 nm and a longitudinal resolution of 10 nm.[5]

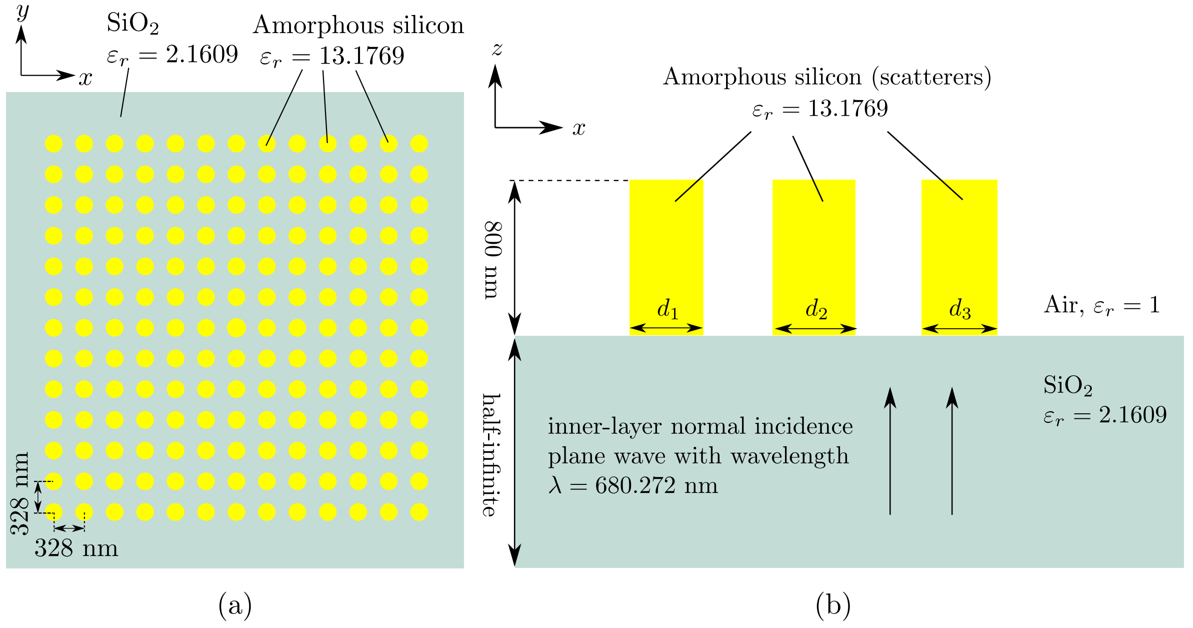


Figure 7.19: A metasurface example. (a) top view of the scatterers and the background medium. Here 13 × 13 circular-cylindrical scatterers are displayed on a rectangular grid with a center-to-center spacing of 328 nm. (b) Side view of three scatterers and the layered background medium. The relative permittivities are given for the free-space wavelength $\lambda_0 = 1000$ nm. The scatterers of this metasurface can have different diameters.

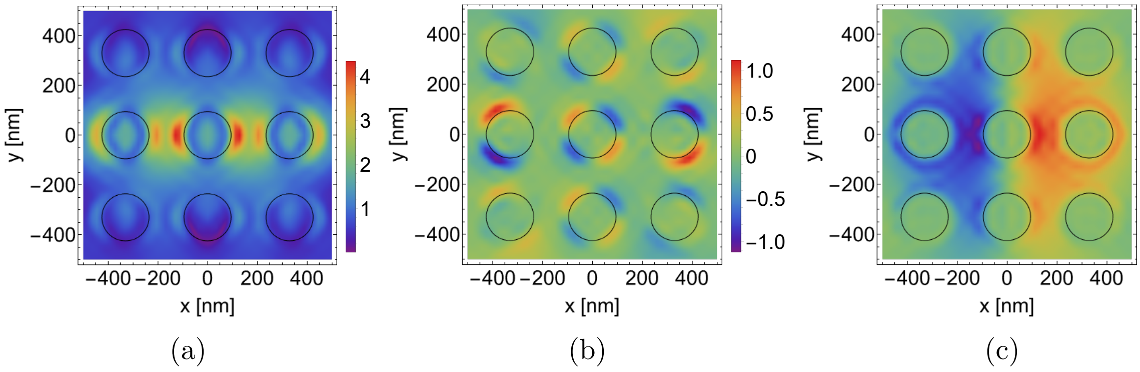
Table 7.2: Discretization parameters used in the metasurface cases.

# scatterers	T_x, T_y [nm]	m_x, m_y	n_x, n_y	N_z	q	p	Nr. of unknowns
3×3	500	$-4:4$	$-16:16$	81	2	3	21.4×10^6
9×9	500	$-7:7$	$-16:16$	81	2	3	59.5×10^6
13×13	500	$-7:7$	$-16:16$	81	2	3	59.5×10^6
27×27	500	$-13:13$	$-16:16$	81	2	3	192.9×10^6

[5] On a single-scatterer test, these parameters yield an average relative error of 10^{-3} in the far field, compared to a self-reference with a transverse resolution of 2.3 nm and a longitudinal resolution of 2 nm.

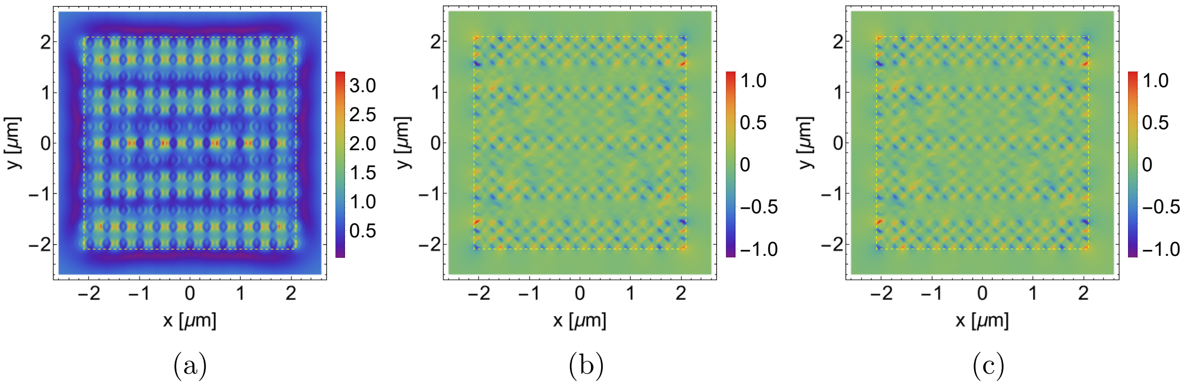
7.3.3 Numerical results

Now we show the near-field solutions of two metasurface cases with different numbers of scatterers, as described in Table 7.2. The first case contains 3×3 scatterers distributed over the transverse domain $[-450, 450] \times [-450, 450]$ nm^2. The second case contains 13×13 scatterers distributed over the domain $[-2100, 2100] \times [-2100, 2100]$ nm^2. The plane that separates the scatterer and the substrate is located at $z = z_a = 0$ nm and we are interested in the total electric field at the top of the scatterers, where $z = z_b = 800$ nm. The near-field solutions of these two cases are shown in Fig. 7.20 and Fig. 7.21, respectively.



(a) (b) (c)

Figure 7.20: Total electric field $\mathbf{E}(x, y)$ at $z = z_b = 800$ nm of a metasurface with 3×3 scatterers. (a) Absolute value of $E_x(x, y)$, (b) real part of $E_y(x, y)$, and (c) real part of $E_z(x, y)$. The black circles denote the boundaries of the circular-cylindrical scatterers in the transverse plane.



(a) (b) (c)

Figure 7.21: Total electric field $\mathbf{E}(x, y)$ at $z = z_b = 800$ nm of a metasurface with 13×13 scatterers. (a) Absolute value of $E_x(x, y)$, (b) real part of $E_y(x, y)$, and (c) real part of $E_z(x, y)$. The yellow dashed box indicates smallest square domain that contains all the scatterers.

The iteration details of the Krylov-subspace solver for the 4 cases in Table 7.2 are shown in Fig. 7.22. The horizontal axis denotes the number of MVPs for the iterative method

IDR(16) and the vertical axis represents the relative error of the approximate solution at each iteration. To make the relative error of the residual reach 1.0×10^{-5}, the 3×3 case requires 435 MVPs and the 9×9 case requires 2567 MVPs. The number of MVPs of the 9×9 case is reduced to 2114, after applying the NVF-BD preconditioner introduced in Chapter 6. The 13×13 case reaches a relative error of 1.0×10^{-4} after 5000 MVPs, and the 27×27 case only reaches a 1.2×10^{-1} relative error after 5000 MVPs. Clearly, more iterations are needed when more scatterers are taken into account.

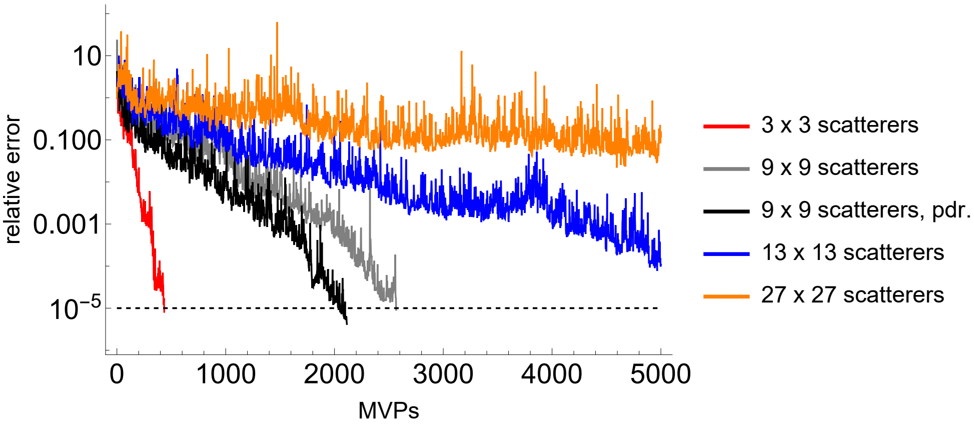


Figure 7.22: Iteration details of the 4 cases with different numbers of scatterers. The 9×9 scatterers case was also tested with the NVF-BD preconditioner proposed in Chapter 6, as indicated by pdr in the legend. The dashed line denotes the desired relative error level in the residual of 1.0×10^{-5}.

7.3.4 Further discussion

We have observed difficulties in the convergence of the Krylov subspace methods for larger cases in Section 7.3.3. To analyze which factor causes this difficulty, we perform the following two groups of simulations:

1. increasing the wavelength of the incident plane wave,
2. reducing the height of the scatterers.

In the first experiment, we increase the wavelength from 1000 nm to 2000 nm and use the discretization parameters specified in Table 7.2. In the sense of relative sizing in an electromagnetic scattering problem, increasing the wavelength of the incident electric field is equivalent to considering a smaller geometry with the original wavelength. We expect this metasurface problem with a larger wavelength to be easier to solve since the scattering effects tend to weaken as the incident waves interact with the geometry on a coarser scale. Fig. 7.22 displays the iteration details of all 4 cases listed in Table 7.2. It is apparent that fewer iterations are required as compared to Fig. 7.22.

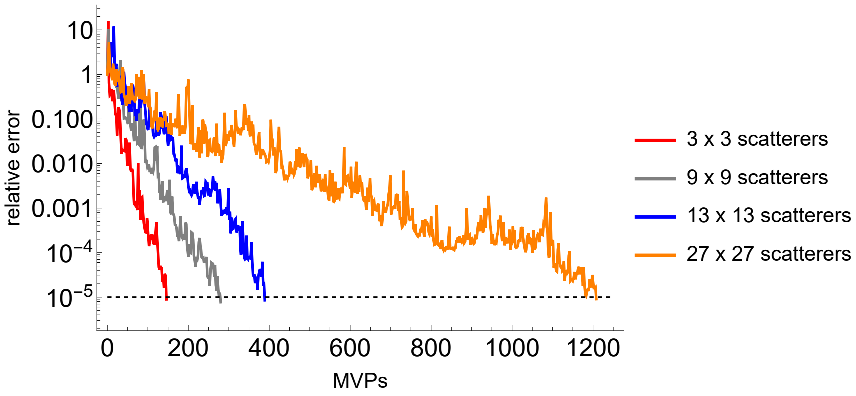
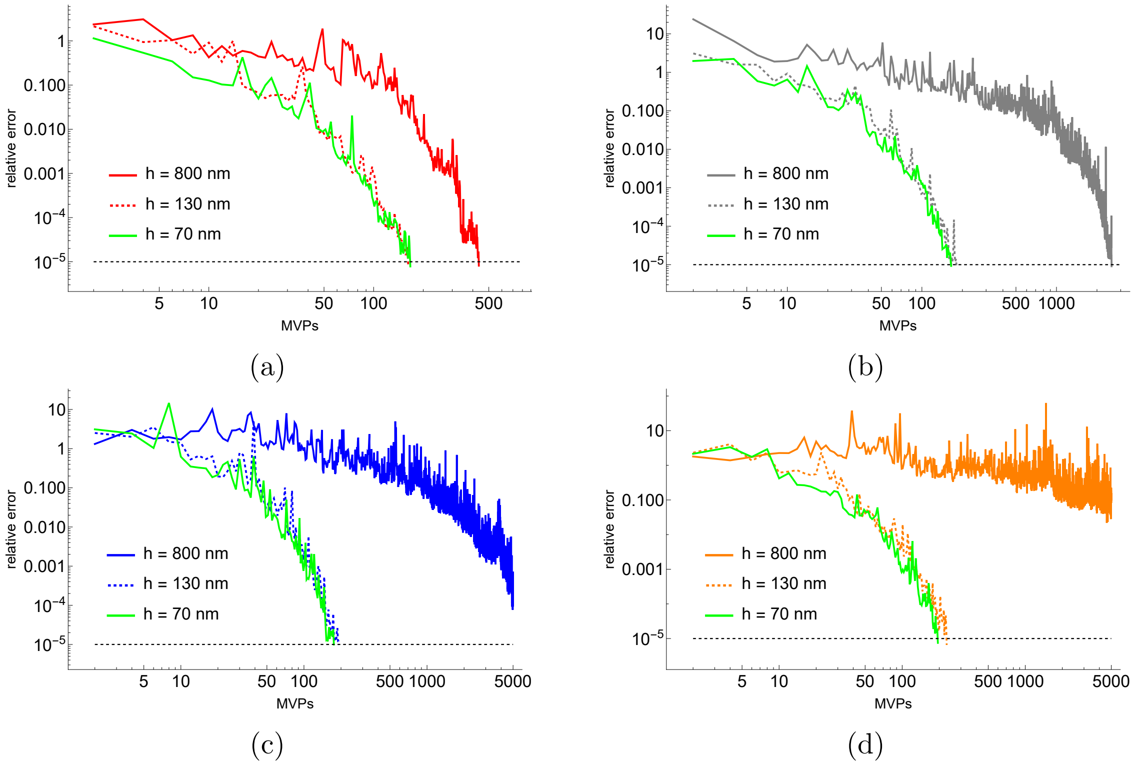


Figure 7.23: Iteration details of the 4 cases in Table 7.2 with an increased wavelength of 2000 nm. The dashed line denotes the desired relative residual error of 1.0×10^{-5}.



(a)

(b)

(c)

(d)

Figure 7.24: Iteration details of the 4 cases, as listed in Table 7.2, with reduced height of the scatterers from 800 nm to 130 nm and 70 nm. (a) 3×3 scatterers, (b) 9×9 scatterers, (c) 13×13 scatterers, and (d) 27×27 scatterers. The dashed line denotes the desired relative residual error level of 1.0×10^{-5}.

In the second experiment, we decrease the height of all the scatterers from 800 nm to 130 nm and 70 nm, respectively. The original wavelength inside the scatterers is 276.3 nm. Therefore, the reduced heights are around 1/2 and 1/4 of the original wavelength in the material of the scatterer[6]. The diameters of the cylindrical scatterers are kept the same. Fig. 7.24 shows the double-logarithmic plots of the iteration details of the 4 cases with the original scatterer height of 800 nm and the reduced scatterer heights.

We summarize the results concerning the number of iterations for all cases involved in the above two experiments in Table 7.3. The second column and the third column record the number of MVPs with the original wavelength of 1000 nm (marked by "org.") and the increased wavelength of 2000 nm, respectively. The recorded numbers of MVPs related to Experiment 2 are given in the last two columns of Table 7.3. Both Fig. 7.24 and Table 7.3 indicate that the scatterers' height is crucial to the number of iterations. A smaller height of the scatterers not only reduces the number of iterations, but it also yields an only moderate increase in the number of iterations when including more scatterers in the transverse direction.

Table 7.3: Number of MVPs comparison.

scatterers	org.	$\lambda_0 = 2000$ nm	$h = 130$ nm	$h = 70$ nm
3×3	435	146	162	167
9×9	2567	280	181	165
13×13	5000+	389	193	175
27×27	5000+	1208	230	195

[6] The second experiment only reduces the geometry of the scatterers in the longitudinal direction, while the first experiment essentially reduces the scatterers in all directions.

Chapter 8

Conclusions and outlook

8.1 Conclusions

The main results of this thesis are the flexible geometry representation, improved multiplication operators, and a preconditioner to reduce the number of iterations of Krylov subspace methods, for a spatial spectral Maxwell solver using Gabor frames.

We proposed two methods to compute Gabor coefficients for 2D indicator functions supported on a polygonal domain. The first one utilized a recurrence relation after applying a finite Taylor-series expansion of the complex error function. Owing to the recurrence relation structure, the pertaining system can be solved efficiently with Olver's algorithm. However, this method requires a high working precision, which results in a limited use in practice. The most valuable lesson learned from this Taylor-based method is that expansion functions should be used that exhibit faster convergence. In a subsequent investigation, we found that the rational expansion of the Faddeeva function is a good candidate for this, since it yields uniform convergence on the whole complex plane, which constitutes the fundament of the second method. A recurrence relation was derived and solved by Olver's algorithm. This so-called RE-DE method surpasses direct numerical quadrature in both accuracy and computation time.

When simulating scattering behavior on high-contrast objects with the spatial spectral method, we encountered an extremely ill-conditioned system due to the originally implemented multiplication operator. The conditioning problem was resolved by an improved multiplication operator for Gabor coefficients, with a cutting procedure in the spectral domain. Equipped with the new multiplication operator, we solved a 3D scattering problem with the contrast $\chi = 16$ and obtained a 10^{-3} relative error in the far field.

To increase the rate of convergence of Krylov subspace solvers applied to the spatial spectral method, an NVF-BD preconditioner was designed and applied in cooperation with BiCGstab(ℓ) or IDR(s). The preconditioned system reduces the number of iterations and computation time in both 2D and 3D examples with high contrast or negative permittivity. In a 2D TM scattering problem, we observed the NVF-BD preconditioner saves up to 90% for high-contrast cases. For some 2D TM cases with negative permittivities and 3D cases with high contrast, the NVF-BD preconditioner makes the solver converge within a reasonable number of iterations while the unpreconditioned system does not converge at all. Analysis of the computation time demonstrates that the overall solution time can be decreased by applying the NVF-BD preconditioner, even though the reduction may be diminished when a large number of transverse basis functions is used. This is attributed to the additional MVPs required for the preconditioner, which introduces a quadratic computational complexity in the transverse directions. Nevertheless, the NVF-BD preconditioner still offers the potential for reducing the total solution time.

Finally, three scattering applications were studied in this thesis. In the first application, we modeled and solved the scattering behavior for a single-pad resist-only metrology target. After comparing the solution of the finite spatial spectral solver with a periodic Maxwell solver, we observed a 10^{-2} level of relative difference. Numerical results suggest that the spatial spectral Maxwell solver has the ability to solve similar types of scattering problems accurately and efficiently, and can hence contribute to an optical metrology sensor model for non-periodic metrology targets. In the second application, we simulated a computer-generated waveguide hologram (CGWH) scattering problem. The hologram area contains 10,126 bar-type scatterers in a 41.6×41.4 μm^2 area. We also simulated a metasurface problem in the third application. Numerical results show the height of the scatterers is a crucial parameter for the rate of convergence of the iterative solver.

8.2 Outlook

Acceleration on the Gabor coefficients computation with the RE-DE method

A rational-expansion-based method has been proposed in Chapter 4. In the sense of computing a single fundamental integral, numerical results show that this method significantly outperforms the direct numerical integration method in terms of computation time, by taking advantage of a recurrence relation.

A complex polygonal object usually requires many Gabor coefficients for reconstruction. Therefore, a reduction in the required number of Gabor coefficients is also important. The RE-DE method proposed in Chapter 4 should be incorporated with a scheme where the total number of Gabor coefficients is optimized. The cost is that the derived second-order difference equations in Section 4.3 might not hold anymore. However, the fundamental integrals can still be computed efficiently by applying the truncated rational expansion of the Faddeeva function by utilizing its fast convergence property on the whole complex plane. Overall, the total computation time should be reduced further after optimizing the required number of Gabor coefficients.

Geometry with curves

The two difference-equation-based methods in Chapter 3 and Chapter 4 are designed for computing the Gabor coefficients of a 2D indicator function supported on a polygonal domain. Even though the N-sided polygon has the advantage to approximate an arbitrary 2D shape by increasing N, the fundamental difference between a straight line and a curve might affect the accuracy and efficiency of the approximation with a polygon. For a curve with high curvature, a refined polygonal approximation might be required, at least locally in the area with high curvature. The analytical RE-DE-based method proposed in Chapter 4 could be extended to circles or ellipses if a similar difference equation can be derived.

Approximation to the NVF-BD preconditioner

The NVF-BD preconditioner proposed in Chapter 6 requires computing the inverse of matrix C, i.e., the matrix representation of the field-material interaction operator $\mathcal{C}_\varepsilon$. For 2D TM polarization, it is not hard to compute this preconditioner since the size of matrix C is usually not so large. However, in 3D cases, computing the inverse directly is a challenging task due to its large dimension. In the 3D example studied in Section 6.4.3, an extra linear system is solved iteratively per iteration to perform the MVP associated with the NVF-BD preconditioner. Even though in general it is not hard to find C^{-1}, owing to its block-diagonal structure, the requirement to perform the multiplication with the preconditioner iteratively in every iteration significantly limits its applications in practice.

One way to approach this problem is to find an approximation of the NVF-BD preconditioner. Analogous to the situation for circulant/Toeplitz preconditioners for Toeplitz systems [216, 217, 220], the ideal approximated matrix $\tilde{C}$ should have a similar eigenvalue distribution of C, and its inverse should be available at a lower computational cost. The starting point should be an explicit representation of the matrix C. This is a challenging task since it requires the explicit representation of the finite 2D Gabor expansion of the C_ε operator and the improved multiplication operators as introduced in Chapter 5.

New preconditioner design

The matrix equation of the spatial spectral method is $(C - G \cdot M) \cdot \mathbf{u} = \mathbf{f}$, and a strong clustering of eigenvalues of the system $C - G \cdot M$ tends to reduce the number of iterations. Several factors affect the distribution of eigenvalues and the skewness of the system $C - G \cdot M$: the normal vector field formulation, the contrast between the scatterers and the environment, and the size of the scattering object(s) with respect to the wavelength of the source.

The NVF-BD preconditioner proposed in Chapter 6 cures the spreading effect of the spectrum due to the NVF and yields a preconditioned system in the form of $I - G \cdot X$. However, for higher-contrast or negative-permittivity scattering problems, there might still be convergence issues for the preconditioned system. On one side, more powerful iterative algorithms with a faster rate of convergence are required to tackle these problems. However, we expect only moderate improvement from this direction. When such advanced iterative methods are not available yet, the other realistic option is to resort to a new preconditioner to reduce the number of iterations by further curing the eigenvalue distribution. There are several requirements for designing such a new preconditioner. Firstly, a clear structure of the system $C - G \cdot M$ of the spatial spectral system should be known a priori. Secondly, the matrix structure of the original system (e.g., symmetry, definiteness, multilevel block structure) should be preserved as much as possible. Thirdly, the new preconditioner should be constructed only once per problem, and its cost should not significantly increase with the number of unknowns.

Validation against experimental measurements

The results of the thesis at hand are mainly compared with external numerical references. In the future, it is also important to conduct experimental validation to verify the performance and reliability of this spatial spectral Maxwell solver. For instance, on the cases in Chapter 7. Studying a similar metrology case for which experimental measurement data is available would provide valuable validation information for and test capabilities of the spatial spectral Maxwell solver. The results of the cases with large structures should also be validated, e.g., the CGWH device with an induced Gaussian beam source. Current numerical results suggest a reasonable scattered beam focus, but a comparison of finer details of the focused beam in the far-field should be conducted.

Appendix A

Extension of an electric field from a plane to a volume

To facilitate real wafer metrology applications, we take a look into the incorporation of a general incident field that is composed of a large number of plane waves. The interest lies in consistently extending the electric field, defined on a horizontal plane at $z = z_d$ in the upper halfspace, to the entire simulation domain $\mathcal{D}$ via the spatial spectral method. The first assumption is that an incident field originates from the upper half-space, given the geometrical formulation of a planarly layered background medium as in [24, p.27], and the simulation domain $\mathcal{D}$ is contained in the n-th layer below the top interface. The second assumption concerns the incidence, i.e. we assume the propagation directions of the plane waves are specified on a uniform grid Λ in the spectral domain. To be precise, this incident field is described as

$$\mathbf{E}^i(\mathbf{k}_T, z_d) = [E^i_x(\mathbf{k}_T)\hat{\mathbf{x}} + E^i_y(\mathbf{k}_T)\hat{\mathbf{y}} + E^i_z(\mathbf{k}_T)\hat{\mathbf{z}}], \tag{A.1}$$

for all $\mathbf{k}_T \in \Lambda$, where z_d satisfies $z_d < z_1$.

We apply the following operations to achieve our goal. First, for each $\mathbf{k}_T \in \Lambda$, the incident field is multiplied by a factor $\exp(\gamma_0[z_1 - z_d])$ with the, generally complex, propagation coefficient

$$\gamma_0 = \sqrt{k_x^2 + k_y^2 - \varepsilon_{rb,0}k_0^2}. \tag{A.2}$$

In other words, each plane wave is propagated to the interface between the upper half-space and the first layer of the planarly layered background medium, i.e. $z = z_1$. Now, the incident field has to be propagated through the layered medium. Unfortunately, this is in its current form not immediately possible due to the branch cuts and poles of the layered medium. Therefore, we employ a complex-plane deformation for $\mathbf{k}_T$, as in [24], to evade these branch cuts and poles. As a result, the second step is to perform the following ordinary real-space Fourier transformation

$$\mathbf{E}^i(\mathbf{x}_T, z_1) = \iint_{\mathbf{k}_T \in \mathbb{R}^2} \mathbf{E}^i(\mathbf{k}_T, z_1)e^{-j\mathbf{k}_T\cdot\mathbf{x}_T}d\mathbf{k}_T, \tag{A.3}$$

where in practice this Fourier transformation is performed on Gabor coefficients together with the list-based representation introduced in [111]. From Eq. (A.3) we can subsequently apply the complex-plane deformed Fourier transformation operator $\mathcal{F}_{\mathbf{x}_T}$ as

$$\mathbf{E}^i[\tau(\mathbf{k}_T), z_1] = \mathcal{F}_{\mathbf{x}_T}[\mathbf{E}^i(\mathbf{x}_T, z_1)](\mathbf{k}_T, z_1). \tag{A.4}$$

The variable τ refers to the complex-plane deformation as formulated in [24, Ch. 8]. Third, we compute the tensorial transmission coefficients $\mathcal{T}^d_{0n}$ and $\mathcal{T}^u_{0n}$ on the complex $\mathbf{k}_T$-plane, see [24, Ch. 2.2.3]. These coefficients link the amplitude of an incident field in the upper half-space, namely layer 0, to its effective amplitude in layer n owing to all transmission and reflection coefficients of the layered medium. We use these coefficients in the fourth step to describe the effective downward-propagating response $\mathbf{W}^d_n(\mathbf{k}_T, z)$ in domain $\mathcal{D}$ as

$$\mathbf{W}^d_n[\tau(\mathbf{k}_T), z] = e^{-\gamma_n(z-z_n)}\mathcal{T}^d_{0n} \cdot \mathbf{E}^i[\tau(\mathbf{k}_T), z_1], \tag{A.5}$$

while the effective upward-propagating response $\mathbf{W}^u_n(\mathbf{k}_T, z)$ is written as

$$\mathbf{W}^u_n[\tau(\mathbf{k}_T), z] = e^{-\gamma_n(z_{n+1}-z)}\mathcal{T}^u_{0n} \cdot \mathbf{E}^i[\tau(\mathbf{k}_T), z_1]. \tag{A.6}$$

The variable γ_n represents the propagation coefficient in layer n. The last step is performing an inverse complex-plane-deformed Fourier transformation by the operator $\mathcal{F}^{-1}_{\mathbf{k}_T}[\ldots](\mathbf{x}_T, z)$, which provides the incident field in simulation domain $\mathcal{D}$ as

$$\mathbf{E}^i(\mathbf{x}_T, z) = \mathcal{F}^{-1}_{\mathbf{x}_T}\big\{\mathbf{W}^d_n[\tau(\mathbf{k}_T), z] + \mathbf{W}^u_n[\tau(\mathbf{k}_T), z]\big\}(\mathbf{x}_T, z). \tag{A.7}$$

Curriculum Vitae

Ligang Sun was born in Qingdao, China on March 23, 1993. After finishing his pre-university study at the Qingdao No. 17 high school, he started his undergraduate study in Computational Mathematics at Shanghai University in 2011. From September 2015 to December 2016, he continued his Master's study at the University of Texas Rio Grande Valley in the United States, with a major in Applied Mathematics. He became a research assistant at Gottfried Wilhelm Leibniz University Hannover in 2017, focusing on uncertainty quantification in discrete-time nonlinear systems.

Starting from November 2018, he has been working as a PhD student at the Eindhoven University of Technology, specifically in the Laboratory of Electromagnetic and Multi-Physics Modeling and Computation within the Electromagnetics group. His research project, which was funded by NWO and received further support from ASML and VSL, centered around Maxwell modeling and its applications in the fields of optical metrology and integrated photonics. The primary results of his PhD research are presented in the thesis at hand.

List of publications

Journal articles

- L. Sun, R. J. Dilz, and M. C. van Beurden. A rational-expansion-based method to compute Gabor coefficients of 2D indicator functions supported on polygonal domain. *Mathematics and Computers in Simulation*, 206: 487-502, 2023.

- L. Sun, R. J. Dilz, and M. C. van Beurden. A normal-vector-field-based preconditioner for a spatial spectral domain-integral equation method for multi-layered electromagnetic scattering problems. *Progress In Electromagnetics Research C*, 123: 1-16, 2022.

- S. Eijsvogel, L. Sun, F. Sepehripour, R. J. Dilz and M. C. van Beurden. Describing discontinuous finite 3D scattering objects in Gabor coefficients: fast and accurate methods. *JOSA A*, 39: 86-97, 2022.

Conference papers and abstracts

- L. Sun, M. C. van Beurden, and R. J. Dilz. Applications of the spatial spectral Maxwell solver. *The 44th Photonics and Electromagnetics Research Symposium (PIERS)*, Prague, Czech Republic. 2023. Accepted.

- L. Sun, D. de Vocht, R. J. Dilz, and M. C. van Beurden. Simulating a computer-generated waveguide hologram scattering problem with an artificial 2D Gaussian beam source. *2023 USNC-URSI National Radio Science Meeting*, Boulder, CO, the United States. 2023. pp. 178-179.

- L. Sun, R. J. Dilz, and M. C. van Beurden. A note on Gabor coefficient computing with Taylor series expansion. *2022 International Conference on Scientific Computing in Electrical Engineering*, Amsterdam, the Netherlands. 2022. pp. 134-135.

- L. Sun, S. Eijsvogel, F. Sepehripour, R. J. Dilz, and M. C. van Beurden. Computation of Gabor coefficients for objects with polygonal cross section. *2021 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, Honolulu, HI, USA. 2021. pp. 072-077.

- S. Eijsvogel, F. Sepehripour, L. Sun, R. J. Dilz and M. C. van Beurden. Local normal vector field formulation for polygonal building blocks in a Gabor representation. *2020 International Conference on Scientific Computing in Electrical Engineering*, Eindhoven, the Netherlands. 2020. pp. 61-62.

Acknowledgements

First of all, I would like to express my sincere gratitude towards my first promoter, Martijn van Beurden. Five years ago, Martijn gave me this opportunity to start my PhD journey, and I distinctly remember during the interview I was told that I would get what I was looking for: the challenge. Now, looking back, I can say without any guilt that I have gone through sufficient challenges and probably have reached the limits of my capabilities. I extend my deepest appreciation to Martijn for his patience in my slow growth when learning computational electromagnetics, particularly as a student coming from a different background. He consistently maintained a wide tolerance for my way of thinking, for which I am truly grateful. Furthermore, I value his commitment to maintaining high standards for the quality of my research work by means of highly critical comments and pursuit of perfectionism. What I learned from Martijn is much more than how to become a researcher. I also learned how to remain patient and persistent; remain sharp to distinguish what the essence is and what is irrelevant; remain rational by focusing on the results only, instead of being affected by emotion, and also remain professional and realistic when involved in cooperation and negotiation.

I am grateful to my co-promoter, Roeland Dilz. Roeland's PhD work is the foundation of the thesis at hand, and my progress would be slower without the help of the original author. For a long time, I have been impressed by his intuition of physics, imagination, and creativity, and I am grateful for his quality control from a physical perspective. This aspect has helped overcome my personal limitation since I usually prioritize mathematical formulations over the meaningful interpretation of results. I am also appreciative of Roeland's patience in demonstrating the physics of electromagnetics, and his support in enhancing my coding skills, particularly during the challenging circumstances posed by the COVID-19 pandemic. Additionally, I would like to thank Roeland for his efforts in organizing the biweekly seminar within our lab. This platform has fostered meaningful academic discussions and enhanced our collective knowledge.

The completion of my PhD work relies on the support of all the members of the MAX-META XT project over the past years. Specifically, I want to extend my heartfelt acknowledgments to Frank Buijnsters for setting up the project that led to Section 7.1 in this thesis and for providing support from the beginning till the end. I thank Filippo Alpeggiani in particular for his significant contributions to the second part of the optical metrology case. I also appreciate Artur Palha for his support in the first half of my PhD project. It was my pleasure to work with these scientists and the collaboration has been a pleasure owing to their high-quality work and professionalism. I am grateful to Henk-Jan Smilde from ASML for providing support from this prestigious company. I thank Giampiero Gerini and Tiberiu Ceccotti from TNO for their valuable support in analyzing a complex scattering problem. I am grateful to Jos van der Tol and David de Vocht from the Photonic Integration group, with whom we have established a collaboration on an application case. Also the contributions to the user committee by Omar El Gawhary and Richard Koops from VSL are gratefully acknowledged. I extend my thanks to Bas de Hon for his knowledgeable instructions and insightful discussions. Furthermore, I would like to thank all the defense committee members for their valuable suggestions and comments, which have greatly contributed to improving the quality of this thesis.

There are certainly unforgettable memories with my friends and colleagues during my PhD life. I appreciate Qiang Liu, Lu Huang and Teng Li for their help and care at the beginning of my joining of TU/e. I appreciate the discussions with Stefan Eijsvogel, who is always enthusiastic and helpful. I would also like to express my appreciation for the fruitful conversations I have had with Fahimeh Sepehripour, who was in the same research project. I feel lucky to have had Ellen van Wesel, Leroy Driessen, and Daan van den Hof as my office mates. The lively and enjoyable atmosphere created by our jokes made office life far from boring. I would like to express my gratitude to Thomas Bressner, Radovan Bojanic, Teun van den Biggelaar, Tom van Nunen, Elles Raaijmakers, Steven Beumer, and numerous other EM colleagues for the engaging and stimulating conversations we had on a wide range of topics. Furthermore, I would like to extend my thanks to Bart Smolders for his leadership of the EM group, and to Suzanne Kuijlaars for her efforts in creating a supportive atmosphere within the group.

I would like to express my sincere gratitude to Prof. Dajun Zhang from Shanghai University, not only for his teachings in mathematical analysis but also for igniting a deep passion for the subject within me from an early age. I am also thankful for his support of my Master's study in the United States. In addition, I extend my appreciation to Prof. Anahit Galstyan from the University of Texas Rio Grande Valley for her significant help on my PhD application. I am also indebted to Prof. Vladik Kreinovich from the University of Texas at El Paso, whose guidance and influence greatly impacted a pivotal decision I had to make before pursuing my PhD. I would like to express my heartfelt thanks to my old friend Gongyue Huang, whose support has been a constant presence since our friendship began in 2008.

Last but certainly not least, I would like to deeply express gratitude to my partner, Xiaoshi Li, for her love and support throughout my PhD journey. During the most difficult times when my PhD absorbed most of my energy, it was she who stood by my side with patience and encouragement and navigate through the difficulties. Furthermore, I consider myself incredibly lucky to have encountered her in the Netherlands, and I deeply cherish the mental connection we share.

Bibliography

[1] I. M. Ross, "The invention of the transistor," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 7–28, 1998.

[2] J. Shalf, "The future of computing beyond Moore's law," *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2166, p. 20190061, 2020.

[3] ASML Holding N.V., "ASML annual report 2021," 2021. `https://www.asml.com/en/investors/annual-report/2021`. Online, Accessed September 2022.

[4] P. Feng, S.-C. Song, G. Nallapati, J. Zhu, J. Bao, V. Moroz, M. Choi, X.-W. Lin, Q. Lu, B. Colombeau, *et al.*, "Comparative analysis of semiconductor device architectures for 5-nm node and beyond," *IEEE Electron Device Letters*, vol. 38, no. 12, pp. 1657–1660, 2017.

[5] M. K. Gupta, P. Weckx, P. Schuddinck, D. Jang, B. Chehab, S. Cosemans, J. Ryckaert, and W. Dehaene, "A comprehensive study of nanosheet and forksheet SRAM for beyond N5 node," *IEEE Transactions on Electron Devices*, vol. 68, no. 8, pp. 3819–3825, 2021.

[6] ASML Holding N.V., "ASML annual report 2022," 2022. `https://www.asml.com/en/investors/annual-report/2022`. Online, Accessed February 2023.

[7] A. C. Diebold, *Handbook of Silicon Semiconductor Metrology.* CRC Press, 2001.

[8] C. A. Mack, "Fifty years of Moore's law," *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, no. 2, pp. 202–207, 2011.

[9] J. Whalen, "Three months, 700 steps: Why it takes so long to produce a computer chip," 2022. `https://www.washingtonpost.com/technology/2021/07/07/making-semiconductors-is-hard/`. Online, Accessed February 2023.

[10] M. G. M. M. van Kraaij, *Forward diffraction modelling: analysis and application to grating reconstruction.* PhD thesis, Eindhoven University of Technology, 2011.

[11] M. van de Kerkhof, T. van Empel, M. Lercel, C. Smeets, F. van de Wetering, A. Nikipelov, C. Cloin, A. Yakunin, and V. Banine, "Advanced particle contamination control in EUV scanners," in *Extreme Ultraviolet (EUV) Lithography X*, vol. 10957, pp. 191–203, SPIE, 2019.

[12] M. Lercel, C. Smeets, M. van de Kerkhof, A. Chen, T. van Empel, and V. Banine, "EUV reticle defectivity protection options," in *Photomask Technology 2019*, vol. 11148, pp. 180–190, SPIE, 2019.

[13] N. G. Orji, M. Badaroglu, B. M. Barnes, C. Beitia, B. D. Bunday, U. Celano, R. J. Kline, M. Neisser, Y. Obeng, and A. Vladar, "Metrology for the next generation of semiconductor devices," *Nature electronics*, vol. 1, no. 10, pp. 532–547, 2018.

[14] T. Yoshizawa, *Handbook of Optical Metrology: Principles and Applications.* CRC press, 2009.

[15] A. G. Marrugo, F. Gao, and S. Zhang, "State-of-the-art active optical techniques for three-dimensional surface metrology: a review," *JOSA A*, vol. 37, no. 9, pp. B60–B77, 2020.

[16] A. J. den Boef, "Optical metrology of semiconductor wafers in lithography," in *International Conference on Optics in Precision Engineering and Nanotechnology (icOPEN2013)*, vol. 8769, pp. 57–65, SPIE, 2013.

[17] A. J. den Boef, "Optical wafer metrology sensors for process-robust CD and overlay control in semiconductor device manufacturing," *Surface Topography: Metrology and Properties*, vol. 4, no. 2, p. 023001, 2016.

[18] J.-S. Kim, J.-M. Byun, R. Lancee, J.-H. Hwang, H.-J. Ha, K.-Y. Hu, S.-R. Jeon, W.-J. Jang, H.-S. Son, V. van der Meijden, *et al.*, "YieldStar uDBO overlay metrology in Samsung D1y DRAM volume production," in *Metrology, Inspection, and Process Control for Microlithography XXXIII*, vol. 10959, pp. 534–540, SPIE, 2019.

[19] Y. Shimizu, L.-C. Chen, D. W. Kim, X. Chen, X. Li, and H. Matsukuma, "An insight into optical metrology in manufacturing," *Measurement Science and Technology*, vol. 32, no. 4, p. 042003, 2021.

[20] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "Nvidia A100 tensor core GPU: performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, 2021.

[21] R. Meyer, Y. Fukuzumi, and Y. Dong, "3D NAND scaling in the next decade," in *2022 International Electron Devices Meeting (IEDM)*, pp. 26–1, IEEE, 2022.

[22] T. Li, Y. Liu, Y. Sun, E. Li, P. Wei, and Y. Li, "Multiple-field-point pupil wavefront optimization in computational lithography," *Applied Optics*, vol. 58, no. 30, pp. 8331–8338, 2019.

[23] L. F. van Rijswijk, F. J. Buijnsters, and M. C. van Beurden, "A linear-complexity layer-coupling algorithm for 1D-and 2D-periodic scattering in multilayered media," *Progress In Electromagnetics Research B*, vol. 96, p. 197, 2022.

[24] R. J. Dilz, *A Spatial Spectral Domain Integral Equation Solver for Electromagnetic Scattering in Dielectric Layered Media.* Technische Universiteit Eindhoven, 2017.

[25] K. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Transactions on Antennas and Propagation*, vol. 14, no. 3, pp. 302–307, 1966.

[26] A. Taflove and M. E. Brodwin, "Numerical solution of steady-state electromagnetic scattering problems using the time-dependent Maxwell's equations," *IEEE Transactions on Microwave Theory and Techniques*, vol. 23, no. 8, pp. 623–630, 1975.

[27] A. Taflove, "Review of the formulation and applications of the finite-difference time-domain method for numerical modeling of electromagnetic wave interactions with arbitrary structures," *Wave Motion*, vol. 10, no. 6, pp. 547–582, 1988.

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

[29] J.-P. Berenger, "Three-dimensional perfectly matched layer for the absorption of electromagnetic waves," *Journal of Computational Physics*, vol. 127, no. 2, pp. 363–379, 1996.

[30] K. S. Yee and J. S. Chen, "The finite-difference time-domain (FDTD) and the finite-volume time-domain (FVTD) methods in solving Maxwell's equations," *IEEE Transactions on Antennas and Propagation*, vol. 45, no. 3, pp. 354–363, 1997.

[31] A. Z. Elsherbeni and V. Demir, *The Finite-Difference Time-Domain Method for Electromagnetics with MATLAB® Simulations*, vol. 2. IET, 2015.

[32] D. M. Sullivan, *Electromagnetic Simulation Using the FDTD Method.* John Wiley & Sons, 2013.

[33] A. Taflove, S. C. Hagness, and M. Piket-May, "Computational electromagnetics: the finite-difference time-domain method," *The Electrical Engineering Handbook*, vol. 3, pp. 629–670, 2005.

[34] K. L. Shlager and J. B. Schneider, "A selective survey of the finite-difference time-domain literature," *IEEE Antennas and Propagation Magazine*, vol. 37, no. 4, pp. 39–57, 1995.

[35] J.-M. Jin, *Theory and Computation of Electromagnetic Fields.* John Wiley & Sons, 2011.

[36] K. T. Cheung, Y. Foo, C. H. To, and J. A. Zapien, "Towards FDTD modeling of spectroscopic ellipsometry data at large angles of incidence," *Applied Surface Science*, vol. 281, pp. 2–7, 2013.

[37] B. M. Barnes, H. Zhou, M.-A. Henn, M. Y. Sohn, and R. M. Silver, "Optimizing image-based patterned defect inspection through FDTD simulations at multiple ultraviolet wavelengths," in *Modeling Aspects in Optical Metrology VI*, vol. 10330, p. 103300W, International Society for Optics and Photonics, 2017.

[38] S. Kwon, J. Park, K. Kim, Y. Cho, and M. Lee, "Microsphere-assisted, nanospot, non-destructive metrology for semiconductor devices," *Light: Science & Applications*, vol. 11, no. 1, pp. 1–14, 2022.

[39] T. Karpisz, B. Salski, R. Buczynski, P. Kopyt, and A. Pacewicz, "Computationally-efficient FDTD modeling of supercontinuum generation in photonic crystal fibers," *Optical and Quantum Electronics*, vol. 48, no. 3, pp. 1–11, 2016.

[40] C. Uluisik, G. Cakir, M. Cakir, and L. Sevgi, "Radar cross section (RCS) modeling and simulation, part 1: a tutorial review of definitions, strategies, and canonical examples," *IEEE Antennas and Propagation Magazine*, vol. 50, no. 1, pp. 115–126, 2008.

[41] G. Cakir, M. Cakir, and L. Sevgi, "Radar cross section (RCS) modeling and simulation, part 2: A novel fdtd-based rcs prediction virtual tool for the resonance regime," *IEEE Antennas and Propagation Magazine*, vol. 50, no. 2, pp. 81–94, 2008.

[42] R. Luebbers and H. Langdon, "A simple feed model that reduces time steps needed for FDTD antenna and microstrip calculations," *IEEE Transactions on Antennas and propagation*, vol. 44, no. 7, pp. 1000–1005, 1996.

[43] S. C. Hagness, A. Taflove, and J. E. Bridges, "Three-dimensional FDTD analysis of a pulsed microwave confocal system for breast cancer detection: Design of an antenna-array element," *IEEE Transactions on Antennas and Propagation*, vol. 47, no. 5, pp. 783–791, 1999.

[44] I. Giannakis, A. Giannopoulos, and C. Warren, "Realistic FDTD GPR antenna models optimized using a novel linear/nonlinear full-waveform inversion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1768–1778, 2018.

[45] R. Palandech, R. Mittra, *et al.*, "Modeling three-dimensional discontinuities in waveguides using nonorthogonal FDTD algorithm," *IEEE Transactions on Microwave Theory and Techniques*, vol. 40, no. 2, pp. 346–352, 1992.

[46] S. Piltyay, A. Bulashenko, Y. Herhil, and O. Bulashenko, "FDTD and FEM simulation of microwave waveguide polarizers," in *2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT)*, pp. 357–363, IEEE, 2020.

[47] X. Yuan, D. Borup, J. Wiskin, M. Berggren, and S. A. Johnson, "Simulation of acoustic wave propagation in dispersive media with relaxation losses by using FDTD method with PML absorbing boundary condition," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 46, no. 1, pp. 14–23, 1999.

[48] I. M. Hallaj and R. O. Cleveland, "FDTD simulation of finite-amplitude pressure and temperature fields for biomedical ultrasound," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. L7–L12, 1999.

[49] G. Lacanna and M. Ripepe, "Modeling the acoustic flux inside the magmatic conduit by 3D-FDTD simulation," *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 6, p. e2019JB018849, 2020.

[50] D. M. Sullivan, "A frequency-dependent FDTD method for biological applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 40, no. 3, pp. 532–539, 1992.

[51] J. Chakarothai, "Novel FDTD scheme for analysis of frequency-dependent medium using fast inverse Laplace transform and Prony's method," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 9, pp. 6076–6089, 2018.

[52] H. M. Yao and L. Jiang, "Machine-learning-based PML for the FDTD method," *IEEE Antennas and Wireless Propagation Letters*, vol. 18, no. 1, pp. 192–196, 2018.

[53] I. Giannakis, A. Giannopoulos, and C. Warren, "A machine learning-based fast-forward solver for ground penetrating radar with application to full-waveform inversion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4417–4426, 2019.

[54] T. Weiland, "A discretization method for the solution of Maxwell's equations for six-component fields," *Electronics and Communications (AEU)*, vol. 31, no. 3, pp. 116–120, 1977.

[55] T. Weiland, "Time domain electromagnetic field computation with finite difference methods," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 9, no. 4, pp. 295–319, 1996.

[56] R. Schuhmann and T. Weiland, "The nonorthogonal finite integration technique applied to 2D-and 3D-eigenvalue problems," *IEEE Transactions on Magnetics*, vol. 36, no. 4, pp. 897–901, 2000.

[57] M. Clemens and T. Weiland, "Discrete electromagnetism with the finite integration technique," *Progress In Electromagnetics Research*, vol. 32, pp. 65–87, 2001.

[58] 3DS, "CST Studio Suite," 2022. `https://www.3ds.com/products-services/simulia/products/cst-studio-suite/`. Online, Accessed January 2023.

[59] Z. Rahimi, *The Finite Integration Technique (FIT) and the Application in Lithography Simulations.* Friedrich-Alexander-Universitaet Erlangen-Nuernberg (Germany), 2011.

[60] A. Hrennikoff, "Solution of problems of elasticity by the framework method," 1941.

[61] R. Courant, "Variational methods for the solution of problems of equilibrium and vibrations," *Bulletin of the American Mathematical Society*, vol. 49, no. 1, pp. 1–23, 1943.

[62] J. L. Volakis, A. Chatterjee, and L. C. Kempel, "Review of the finite-element method for three-dimensional electromagnetic scattering," *JOSA A*, vol. 11, no. 4, pp. 1422–1433, 1994.

[63] P. Silvester, "Finite element solution of homogeneous waveguide problems," *Alta Frequenza*, vol. 38, no. 1, pp. 313–317, 1969.

[64] R. Coccioli, T. Itoh, G. Pelosi, and P. P. Silvester, "Finite-element methods in microwaves: a selected bibliography," *IEEE Antennas and Propagation Magazine*, vol. 38, no. 6, pp. 34–48, 1996.

[65] M. G. Larson and F. Bengzon, *The Finite Element Method: Theory, Implementation, and Applications*, vol. 10. Springer Science & Business Media, 2013.

[66] D. B. Davidson, *Computational Electromagnetics for RF and Microwave Engineering*. Cambridge University Press, 2010.

[67] F. Bréchet, J. Marcou, D. Pagnoux, and P. Roy, "Complete analysis of the characteristics of propagation into photonic crystal fibers, by the finite element method," *Optical Fiber Technology*, vol. 6, no. 2, pp. 181–191, 2000.

[68] Y. Tsuji and M. Koshiba, "Finite element method using port truncation by perfectly matched layer boundary conditions for optical waveguide discontinuity problems," *Journal of Lightwave Technology*, vol. 20, no. 3, pp. 463–468, 2002.

[69] N. Marais, *Efficient High-order Time Domain Finite Element Methods in Electromagnetics*. PhD thesis, Stellenbosch: University of Stellenbosch, 2009.

[70] A. Anees and L. Angermann, "Time domain finite element method for Maxwell's equations," *IEEE Access*, vol. 7, pp. 63852–63867, 2019.

[71] A. Bossavit and J.-C. Vérité, "A mixed FEM-BIEM method to solve 3-D eddy-current problems," *IEEE Transactions on Magnetics*, vol. 18, no. 2, pp. 431–435, 1982.

[72] M. Barton and Z. Cendes, "New vector finite elements for three-dimensional magnetic field computation," *Journal of Applied Physics*, vol. 61, no. 8, pp. 3919–3921, 1987.

[73] B. Zhou and D. Jiao, "Direct finite-element solver of linear complexity for large-scale 3-D electromagnetic analysis and circuit extraction," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 10, pp. 3066–3080, 2015.

[74] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, no. 6, p. 409, 1952.

[75] R. F. Harrington, "Matrix methods for field problems," *Proceedings of the IEEE*, vol. 55, no. 2, pp. 136–149, 1967.

[76] C. Müller, "Foundations of the mathematical theory of electromagnetic waves," 1969.

[77] A. J. Poggio and E. K. Miller, *Integral Equation Solutions of Three-Dimensional Scattering Problems.* MB Assoc., 1970.

[78] Y. Chang and R. Harrington, "A surface formulation for characteristic modes of material bodies," *IEEE Transactions on Antennas and Propagation*, vol. 25, no. 6, pp. 789–795, 1977.

[79] T.-K. Wu and L. L. Tsai, "Scattering from arbitrarily-shaped lossy dielectric bodies of revolution," *Radio Science*, vol. 12, no. 5, pp. 709–718, 1977.

[80] M. M. Ney, "Method of moments as applied to electromagnetic problems," *IEEE Transactions on Microwave Theory and Techniques*, vol. 33, no. 10, pp. 972–980, 1985.

[81] W. C. Gibson, *The Method of Moments in Electromagnetics.* Chapman and Hall/CRC, 2021.

[82] S. Rao, D. Wilton, and A. Glisson, "Electromagnetic scattering by surfaces of arbitrary shape," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 3, pp. 409–418, 1982.

[83] H. A. Van der Vorst, "Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems," *SIAM Journal on Scientific and Statistical Computing*, vol. 13, no. 2, pp. 631–644, 1992.

[84] M. H. Gutknecht, "Variants of BiCGStab for matrices with complex spectrum," *SIAM Journal on Scientific Computing*, vol. 14, no. 5, pp. 1020–1033, 1993.

[85] G. L. Sleijpen and D. R. Fokkema, "BiCGstab (ell) for linear equations involving unsymmetric matrices with complex spectrum," *Electronic Transactions on Numerical Analysis.*, vol. 1, pp. 11–32, 1993.

[86] P. Sonneveld and M. B. van Gijzen, "IDR (s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 1035–1062, 2009.

[87] A. F. Peterson, S. L. Ray, R. Mittra, I. of Electrical, and E. Engineers, *Computational Methods for Electromagnetics*, vol. 351. IEEE press New York, 1998.

[88] A. Bondeson, T. Rylander, and P. Ingelström, *Computational Electromagnetics*. Springer, 2012.

[89] V. Rokhlin, "Rapid solution of integral equations of scattering theory in two dimensions," *Journal of Computational Physics*, vol. 86, no. 2, pp. 414–439, 1990.

[90] R. Coifman, V. Rokhlin, and S. Wandzura, "The fast multipole method for the wave equation: A pedestrian prescription," *IEEE Antennas and Propagation Magazine*, vol. 35, no. 3, pp. 7–12, 1993.

[91] J. M. Song and W. C. Chew, "Multilevel fast-multipole algorithm for solving combined field integral equations of electromagnetic scattering," *Microwave and Optical Technology Letters*, vol. 10, no. 1, pp. 14–19, 1995.

[92] B. Dembart and E. Yip, "A 3D fast multipole method for electromagnetics with multiple levels," in *11th Annual Review of Progress in Applied Computational Electromagnetics*, vol. 1, pp. 621–628, 1995.

[93] J. Song, C.-C. Lu, and W. C. Chew, "Multilevel fast multipole algorithm for electromagnetic scattering by large complex objects," *IEEE Transactions on Antennas and Propagation*, vol. 45, no. 10, pp. 1488–1493, 1997.

[94] V. Prakash and R. Mittra, "Characteristic basis function method: A new technique for efficient solution of method of moments matrix equations," *Microwave and Optical Technology Letters*, vol. 36, no. 2, pp. 95–100, 2003.

[95] E. Garcia, C. Delgado, I. Gonzalez, and F. Catedra, "Efficient parallelization of a CBFM-MLFMA scheme for the computation of complex electromagnetic problems," in *2008 IEEE Antennas and Propagation Society International Symposium*, pp. 1–4, IEEE, 2008.

[96] D. Ludick and D. Davidson, "Investigating efficient parallelization techniques for the characteristic basis function method (CBFM)," in *2009 International Conference on Electromagnetics in Advanced Applications*, pp. 400–403, IEEE, 2009.

[97] M. Moharam, E. B. Grann, D. A. Pommet, and T. Gaylord, "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," *JOSA a*, vol. 12, no. 5, pp. 1068–1076, 1995.

[98] M. Moharam, D. A. Pommet, E. B. Grann, and T. Gaylord, "Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach," *JOSA A*, vol. 12, no. 5, pp. 1077–1086, 1995.

[99] M. Moharam and T. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," *JOSA*, vol. 71, no. 7, pp. 811–818, 1981.

[100] M. Moharam and T. K. Gaylord, "Diffraction analysis of dielectric surface-relief gratings," *JOSA*, vol. 72, no. 10, pp. 1385–1392, 1982.

[101] W. Baird, M. Moharam, and T. Gaylord, "Diffraction characteristics of planar absorption gratings," *Applied Physics B*, vol. 32, no. 1, pp. 15–20, 1983.

[102] M. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of metallic surface-relief gratings," *JOSA A*, vol. 3, no. 11, pp. 1780–1787, 1986.

[103] E. Popov and M. Nevière, "Grating theory: new equations in Fourier space leading to fast converging results for TM polarization," *JOSA A*, vol. 17, no. 10, pp. 1773–1784, 2000.

[104] E. Popov, M. Nevière, B. Gralak, and G. Tayeb, "Staircase approximation validity for arbitrary-shaped gratings," *JOSA A*, vol. 19, no. 1, pp. 33–42, 2002.

[105] E. Popov and M. Nevière, "Maxwell equations in Fourier space: fast-converging formulation for diffraction by arbitrary shaped, periodic, anisotropic media," *JOSA A*, vol. 18, no. 11, pp. 2886–2894, 2001.

[106] L. Li, "Use of Fourier series in the analysis of discontinuous periodic structures," *JOSA A*, vol. 13, no. 9, pp. 1870–1876, 1996.

[107] M. C. van Beurden and I. D. Setija, "Local normal vector field formulation for periodic scattering problems formulated in the spectral domain," *JOSA A*, vol. 34, no. 2, pp. 224–233, 2017.

[108] R. J. Dilz, M. G. van Kraaij, and M. C. van Beurden, "2D TM scattering problem for finite dielectric objects in a dielectric stratified medium employing Gabor frames in a domain integral equation," *JOSA A*, vol. 34, no. 8, pp. 1315–1321, 2017.

[109] R. J. Dilz and M. C. van Beurden, "A domain integral equation approach for simulating two dimensional transverse electric scattering in a layered medium with a Gabor frame discretization," *Journal of Computational Physics*, vol. 345, pp. 528–542, 2017.

[110] R. J. Dilz, M. G. van Kraaij, and M. C. van Beurden, "A 3D spatial spectral integral equation method for electromagnetic scattering from finite objects in a layered medium," *Optical and Quantum Electronics*, vol. 50, no. 5, pp. 1–22, 2018.

[111] R. J. Dilz and M. C. van Beurden, "Fast operations for a Gabor-frame-based integral equation with equidistant sampling," *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 1, pp. 82–85, 2017.

[112] S. Eijsvogel, R. J. Dilz, and M. C. van Beurden, "Inverse scattering with a parametrized spatial spectral volume integral equation for finite scatterers," *JOSA A*, 2023.

[113] J. A. C. Weideman, "Computation of the complex error function," *SIAM Journal on Numerical Analysis*, vol. 31, no. 5, pp. 1497–1518, 1994.

[114] D. S. Jones, *The Theory of Electromagnetism.* Pergamon Press, 1964.

[115] J. D. Jackson, *Classical Electrodynamics.* American Association of Physics Teachers, 1999.

[116] D. Colton and R. Kress, *Integral Equations in Scattering Theory.* Pure and Applied Mathematics, John Wiley and Sons, New York, 1983.

[117] S. Silver, *Microwave Antenna Theory and Design.* No. 19, Iet, 1949.

[118] C. Müller, *Grundprobleme der Mathematischen Theorie Elektromagnetischer Schwingungen*, vol. 88. Springer, 1957.

[119] K. A. Michalski and J. R. Mosig, "Multilayered media Green's functions in integral equation formulations," *IEEE Transactions on Antennas and Propagation*, vol. 45, no. 3, pp. 508–519, 1997.

[120] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*, vol. 31. John Wiley & Sons, 1994.

[121] J. R. Wait, *Electromagnetic Waves in Stratified Media: Revised Edition including Supplemented Material*, vol. 3. Elsevier, 2013.

[122] M. C. van Beurden, "Fast convergence with spectral volume integral equation for crossed block-shaped gratings with improved material interface conditions," *JOSA A*, vol. 28, no. 11, pp. 2269–2278, 2011.

[123] M. C. van Beurden, "A spectral volume integral equation method for arbitrary biperiodic gratings with explicit Fourier factorization," *Progress In Electromagnetics Research*, vol. 36, pp. 133–149, 2012.

[124] L. Li and C. W. Haggans, "Convergence of the coupled-wave method for metallic lamellar diffraction gratings," *JOSA A*, vol. 10, no. 6, pp. 1184–1189, 1993.

[125] G. Granet and B. Guizal, "Efficient implementation of the coupled-wave method for metallic lamellar gratings in TM polarization," *JOSA A*, vol. 13, no. 5, pp. 1019–1023, 1996.

[126] P. Lalanne and G. M. Morris, "Highly improved convergence of the coupled-wave method for tm polarization," *JOSA A*, vol. 13, no. 4, pp. 779–784, 1996.

[127] T. Schuster, J. Ruoff, N. Kerwien, S. Rafler, and W. Osten, "Normal vector method for convergence improvement using the RCWA for crossed gratings," *JOSA A*, vol. 24, no. 9, pp. 2880–2890, 2007.

[128] Y.-C. Chang, G. Li, H. Chu, and J. Opsal, “Efficient finite-element, Green’s function approach for critical-dimension metrology of three-dimensional gratings on multilayer films,” *JOSA A*, vol. 23, no. 3, pp. 638–645, 2006.

[129] K. Gröchenig, *Foundations of Time-Frequency Analysis.* Springer Science & Business Media, 2001.

[130] O. Christensen, H. G. Feichtinger, and S. Paukner, “Gabor analysis for imaging,” in *Handbook of Mathematical Methods in Imaging*, pp. 1717–1757, Springer, 2015.

[131] E. M. Stein and R. Shakarchi, *Real Analysis: Measure theory, Integration, and Hilbert Spaces*, vol. 3. Princeton University Press, 2009.

[132] H. G. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms: Theory and Applications.* Birkhäuser, Boston, 1998.

[133] H. G. Feichtinger and T. Strohmer, *Advances in Gabor Analysis.* Birkhäuser, Boston, 2003.

[134] M. J. Bastiaans, *Gabor’s expansion and the Zak transform for continuous-time and discrete-time signals: critical sampling and rational oversampling.* Citeseer, 1995.

[135] Y. Saad, *Iterative Methods for Sparse Linear Systems.* SIAM, 2003.

[136] V. Faber and T. Manteuffel, “Necessary and sufficient conditions for the existence of a conjugate gradient method,” *SIAM Journal on Numerical Analysis*, vol. 21, no. 2, pp. 352–362, 1984.

[137] C. C. Paige and M. A. Saunders, “Solution of sparse indefinite systems of linear equations,” *SIAM Journal on Numerical Analysis*, vol. 12, no. 4, pp. 617–629, 1975.

[138] Y. Saad and M. H. Schultz, “GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems,” *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 3, pp. 856–869, 1986.

[139] A. J. Wathen, “Preconditioning,” *Acta Numerica*, vol. 24, 2015.

[140] R. Fletcher, “Conjugate gradient methods for indefinite systems,” in *Numerical Analysis*, pp. 73–89, Springer, 1976.

[141] C. Lanczos, “Solution of systems of linear equations by minimized iterations,” *Journal of Research of the National Bureau of Standards*, vol. 49, no. 1, p. 33, 1952.

[142] P. Wesseling and P. Sonneveld, “Numerical experiments with a multiple grid and a preconditioned Lanczos type method,” in *Approximation Methods for Navier-Stokes Problems*, pp. 543–562, Springer, 1980.

[143] L. Sun, S. Eijsvogel, F. Sepehripour, R. Dilz, and M. C. van Beurden, "Computation of Gabor coefficients for objects with polygonal cross section," in *2021 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, pp. 072–077, IEEE, 2021.

[144] L. Sun, R. J. Dilz, and M. C. van Beurden, "A note on Gabor coefficient computing with Taylor series expansion," in *The 14th International Conference on Scientific Computing in Electrical Engineering*, 2022.

[145] S. Chevillard, "The functions erf and erfc computed with arbitrary precision," *Rapport de recherche RR2009-04, laboratoire de l'informatique du parallélisme (LIP)*, vol. 46, p. 29, 2009.

[146] W. Gautschi, "Computational aspects of three-term recurrence relations," *SIAM Review*, vol. 9, no. 1, pp. 24–82, 1967.

[147] J. Wimp, *Computation with Recurrence Relations.* Pitman, 1984.

[148] F. W. J. Olver, "Numerical solution of second-order linear difference equations," *Journal of Research of the National Bureau of Standards - B. Mathematics and Mathematical Physics*, vol. 71B, no. 2,3, 1967.

[149] J. Cash, "A note on Olver's algorithm for the solution of second-order linear difference equations," *Mathematics of Computation*, vol. 35, no. 151, pp. 767–772, 1980.

[150] J. Cash, "An extension of Olver's method for the numerical solution of linear recurrence relations," *Mathematics of Computation*, vol. 32, no. 142, pp. 497–510, 1978.

[151] G. P. Poppe and C. M. Wijers, "More efficient computation of the complex error function," *ACM Transactions on Mathematical Software (TOMS)*, vol. 16, no. 1, pp. 38–46, 1990.

[152] W. J. Cody, "Rational chebyshev approximations for the error function," *Mathematics of Computation*, vol. 23, no. 107, pp. 631–637, 1969.

[153] J. Schonfelder, "Chebyshev expansions for the error and related functions," *Mathematics of Computation*, vol. 32, no. 144, pp. 1232–1240, 1978.

[154] W. Gautschi, "Efficient computation of the complex error function," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 187–198, 1970.

[155] L. Sun, R. J. Dilz, and M. C. van Beurden, "A rational-expansion-based method to compute Gabor coefficients of 2D indicator functions supported on polygonal domain," *Mathematics and Computers in Simulation*, vol. 206, pp. 487–502, 2023.

[156] D. Gabor, "Theory of communication. Part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[157] S. Paukner, *Foundations of Gabor Analysis for Image Processing.* Master's thesis, 2007.

[158] H. G. Feichtinger, "Gabor expansions of signals: Computational aspects and open questions," in *Landscapes of Time-Frequency Analysis*, pp. 173–206, Springer, 2019.

[159] M. J. Bastiaans and M. C. Geilen, "On the discrete Gabor transform and the discrete Zak transform," *Signal Processing*, vol. 49, no. 3, pp. 151–166, 1996.

[160] T. Strohmer, "Approximation of dual Gabor frames, window decay, and wireless communications," *Applied and Computational Harmonic Analysis*, vol. 11, no. 2, pp. 243–262, 2001.

[161] T. Strohmer, "Pseudodifferential operators and banach algebras in mobile communications," *Applied and Computational Harmonic Analysis*, vol. 20, no. 2, pp. 237–249, 2006.

[162] J. J. Maciel and L. 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*, vol. 50, no. 12, pp. 1751–1759, 2002.

[163] J. J. Maciel and L. B. Felsen, "Gabor-based narrow-waisted Gaussian beam algorithm for transmission of aperture-excited 3D vector fields through arbitrarily shaped 3D dielectric layers," *Radio Science*, vol. 37, no. 2, pp. 1–9, 2002.

[164] D. Lugara and C. Letrou, "Printed antennas analysis by a Gabor frame-based method of moments," *IEEE Transactions on Antennas and Propagation*, vol. 50, no. 11, pp. 1588–1597, 2002.

[165] A. Shlivinski, E. Heyman, A. Boag, and C. Letrou, "A phase-space beam summation formulation for ultrawide-band radiation," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 8, pp. 2042–2056, 2004.

[166] A. Fluerasu and C. Letrou, "Gaussian beam launching for 3D physical modeling of propagation channels," *annals of telecommunications-annales des télécommunications*, vol. 64, no. 11, pp. 763–776, 2009.

[167] S. J. Floris and B. P. de Hon, "Wilson basis expansions of electromagnetic wavefields: a suitable framework for fiber optics," *Optical and Quantum Electronics*, vol. 50, no. 3, pp. 1–25, 2018.

[168] S. J. Floris and B. P. de Hon, "Electromagnetic reflection–transmission problems in a Wilson basis: fiber-optic mode-matching to homogeneous media," *Optical and Quantum Electronics*, vol. 50, no. 3, pp. 1–18, 2018.

[169] S. J. Floris, B. P. de Hon, M. C. van Beurden, and T. Bolhaar, “Electromagnetic mode matching in a Wilson basis: optical fiber connections with a gap,” *Optical and Quantum Electronics*, vol. 54, no. 11, pp. 1–36, 2022.

[170] S. Eijsvogel, L. Sun, F. Sepehripour, R. J. Dilz, and M. C. van Beurden, “Describing discontinuous finite 3D scattering objects in Gabor coefficients: fast and accurate methods,” *JOSA A*, vol. 39, no. 1, pp. 86–97, 2022.

[171] G. P. Lepage, “A new algorithm for adaptive multidimensional integration,” *Journal of Computational Physics*, vol. 27, no. 2, pp. 192–203, 1978.

[172] A. Klöckner, A. Barnett, L. Greengard, and M. O’Neil, “Quadrature by expansion: A new method for the evaluation of layer potentials,” *Journal of Computational Physics*, vol. 252, pp. 332–349, 2013.

[173] S. Chevillard, “The functions erf and erfc computed with arbitrary precision and explicit error bounds,” *Information and Computation*, vol. 216, pp. 72–95, 2012.

[174] W. J. Cody, “Performance evaluation of programs for the error and complementary error functions,” *ACM Transactions on Mathematical Software (TOMS)*, vol. 16, no. 1, pp. 29–37, 1990.

[175] D. Hunter and T. Regan, “A note on the evaluation of the complementary error function,” *Mathematics of Computation*, vol. 26, no. 118, pp. 539–541, 1972.

[176] J. Weideman, “Computing integrals of the complex error function,” in *Proceedings of Symposia in Applied Mathematics*, vol. 48, pp. 403–407, 1994.

[177] P. Van der Cruyssen, “A reformulation of Olver’s algorithm for the numerical solution of second-order linear difference equations,” *Numerische Mathematik*, vol. 32, no. 2, pp. 159–166, 1979.

[178] A. J. Janssen, “Some Weyl-Heisenberg frame bound calculations,” *Indagationes Mathematicae*, vol. 7, no. 2, pp. 165–183, 1996.

[179] T. Strohmer, “Rates of convergence for the approximation of dual shift-invariant systems in $\ell^2(\mathbb{Z})$,” *Journal of Fourier Analysis and Applications*, vol. 5, no. 6, pp. 599–615, 1999.

[180] N. Kaiblinger, “Approximation of the Fourier transform and the dual Gabor window,” *Journal of Fourier Analysis and Applications*, vol. 11, no. 1, pp. 25–42, 2005.

[181] H. G. Feichtinger, “On a new Segal algebra,” *Monatshefte für Mathematik*, vol. 92, no. 4, pp. 269–289, 1981.

[182] H. G. Feichtinger and N. Kaiblinger, “Quasi-interpolation in the Fourier algebra,” *Journal of Approximation Theory*, vol. 144, no. 1, pp. 103–118, 2007.

[183] T. Werther, E. Matusiak, Y. C. Eldar, and N. K. Subbana, "A unified approach to dual Gabor windows," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 1758–1768, 2007.

[184] P. L. Søndergaard, "Gabor frames by sampling and periodization," *Advances in Computational Mathematics*, vol. 27, no. 4, pp. 355–373, 2007.

[185] V. N. Faddeeva and N. M. Terent'ev, *Tables of Values of the Function* $W(z)$. Pergamon Press, 1961.

[186] B. D. Fried and S. D. Conte, *The Plasma Dispersion Function: the Hilbert Transform of the Gaussian*. Academic Press, 2015.

[187] D. B. Owen, "Tables for computing bivariate normal probabilities," *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 1075–1090, 1956.

[188] F. Olver and D. Sookne, "Note on backward recurrence algorithms," *Mathematics of Computation*, vol. 26, no. 120, pp. 941–947, 1972.

[189] J. W. Gibbs, "Letter to the Editor, Fourier's Series," *Nature*, vol. 59(1539), p. 606, 1899.

[190] D. Gottlieb, C.-W. Shu, A. Solomonoff, and H. Vandeven, "On the Gibbs phenomenon I: Recovering exponential accuracy from the Fourier partial sum of a nonperiodic analytic function," *Journal of Computational and Applied Mathematics*, vol. 43, no. 1-2, pp. 81–98, 1992.

[191] D. Gottlieb and C.-W. Shu, "Resolution properties of the Fourier method for discontinuous waves," *Computer methods in applied mechanics and engineering*, vol. 116, no. 1-4, pp. 27–37, 1994.

[192] D. Gottlieb and C.-W. Shu, "On the Gibbs phenomenon III: recovering exponential accuracy in a sub-interval from a spectral partial sum of a piecewise analytic function," *SIAM journal on numerical analysis*, vol. 33, no. 1, pp. 280–290, 1996.

[193] D. Gottlieb and C.-W. Shu, "On the gibbs phenomenon IV: Recovering exponential accuracy in a subinterval from a gegenbauer partial sum of a piecewise analytic function," *Mathematics of Computation*, vol. 64, no. 211, pp. 1081–1095, 1995.

[194] D. Gottlieb and C.-W. Shu, "On the gibbs phenomenon V: Recovering exponential accuracy from collocation point values of a piecewise analytic function," *Numerische Mathematik*, vol. 71, no. 4, pp. 511–526, 1995.

[195] A. Zygmund, *Trigonometric Series*, vol. 1. Cambridge university press, 1977.

[196] L. Li, "New formulation of the Fourier modal method for crossed surface-relief gratings," *JOSA A*, vol. 14, no. 10, pp. 2758–2767, 1997.

[197] 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*, vol. 69, pp. 117–136, 2016.

[198] S. Burger, L. Zschiedrich, J. Pomplun, and F. Schmidt, "Finite-element based electromagnetic field simulations: benchmark results for isolated structures," *arXiv preprint arXiv:1310.2732*, 2013.

[199] L. Sun, R. J. Dilz, and M. C. van Beurden, "A normal-vector-field-based preconditioner for a spatial spectral domain-integral equation method for multi-layered electromagnetic scattering problems," *Progress In Electromagnetics Research C*, vol. 123, pp. 1–16, 2022.

[200] V. Domnenko, B. Küchler, W. Hoppe, J. Preuninger, U. Klostermann, W. Demmerle, M. Bohn, D. Krüger, R. R. H. Kim, and L. E. Tan, "EUV computational lithography using accelerated topographic mask simulation," in *Design-Process-Technology Co-optimization for Manufacturability XIII*, vol. 10962, p. 109620O, International Society for Optics and Photonics, 2019.

[201] Y.-S. Ku, H.-L. Pang, W.-T. Hsu, and D.-M. Shyu, "Accuracy of diffraction-based overlay metrology using a single array target," *Optical Engineering*, vol. 48, no. 12, p. 123601, 2009.

[202] L. Wang, Y. Wang, and X. Zhang, "Embedded metallic focus grating for silicon nitride waveguide with enhanced coupling and directive radiation," *Optics Express*, vol. 20, no. 16, pp. 17509–17521, 2012.

[203] D. O. Dzibrou, J. J. van der Tol, and M. K. Smit, "Tolerant polarization converter for ingaasp-inp photonic integrated circuits," *Optics Letters*, vol. 38, no. 18, pp. 3482–3484, 2013.

[204] N. Yu, P. Genevet, M. A. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, and Z. Gaburro, "Light propagation with phase discontinuities: generalized laws of reflection and refraction," *Science*, vol. 334, no. 6054, pp. 333–337, 2011.

[205] S. Jahani and Z. Jacob, "All-dielectric metamaterials," *Nature Nanotechnology*, vol. 11, no. 1, pp. 23–36, 2016.

[206] M. I. Sancer, K. Sertel, J. L. Volakis, and P. Van Alstine, "On volume integral equations," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 5, pp. 1488–1495, 2006.

[207] M. M. Botha, "Solving the volume integral equations of electromagnetic scattering," *Journal of Computational Physics*, vol. 218, no. 1, pp. 141–158, 2006.

[208] P. Ylä-Oijala, M. Taskinen, and S. Järvenpää, "Surface integral equation formulations for solving electromagnetic scattering problems with iterative methods," *Radio Science*, vol. 40, no. 6, 2005.

[209] H. van der Vorst, "A fast and smoothly convergent variant of BI-CG for the solution of nonsymmetrical linear systems," *SIAM Journal on Scientific and Statistical Computing*, vol. 13, pp. 631–644, 1992.

[210] R. Remis, "Circulant preconditioners for domain integral equations in electromagnetics," in *2012 International Conference on Electromagnetics in Advanced Applications*, pp. 337–340, IEEE, 2012.

[211] R. Remis, "Preconditioning techniques for domain integral equations," in *2013 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, pp. 235–238, IEEE, 2013.

[212] S. P. Groth, A. G. Polimeridis, A. Tambova, and J. K. White, "Circulant preconditioning in the volume integral equation method for silicon photonics," *JOSA A*, vol. 36, no. 6, pp. 1079–1088, 2019.

[213] F. Schneider, "Approximation of inverses of BTTB matrices," Master's thesis, Eindhoven University of Technology, 2016.

[214] V. I. Morgenshtern and H. Bölcskei, "A short course on frame theory," *arXiv preprint arXiv:1104.4300*, 2011.

[215] O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems: Theory and Computation.* SIAM, 2001.

[216] T. F. Chan, "An optimal circulant preconditioner for Toeplitz systems," *SIAM Journal on Scientific and Statistical Computing*, vol. 9, no. 4, pp. 766–771, 1988.

[217] R. H. Chan, "Circulant preconditioners for Hermitian Toeplitz systems," *SIAM Journal on Matrix Analysis and Applications*, vol. 10, no. 4, pp. 542–550, 1989.

[218] E. E. Tyrtyshnikov, "Optimal and superoptimal circulant preconditioners," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 2, pp. 459–473, 1992.

[219] R. H. Chan, "Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions," *IMA Journal of Numerical Analysis*, vol. 11, no. 3, pp. 333–345, 1991.

[220] R. H. Chan and K.-P. Ng, "Toeplitz preconditioners for Hermitian Toeplitz systems," *Linear Algebra and Its Applications*, vol. 190, pp. 181–208, 1993.

[221] D. Noutsos and P. Vassalos, "New band Toeplitz preconditioners for ill-conditioned symmetric positive definite Toeplitz systems," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 3, pp. 728–743, 2002.

[222] F.-R. Lin, "Preconditioners for block Toeplitz systems based on circulant preconditioners," *Numerical Algorithms*, vol. 26, no. 4, pp. 365–379, 2001.

[223] L. Sun, D. De Vocht, R. J. Dilz, and M. C. van Beurden, "Simulating a computer-generated waveguide hologram scattering problem with an artificial 2D Gaussian beam source," in *2023 United States National Committee of URSI National Radio Science Meeting (USNC-URSI NRSM)*, pp. 178–179, IEEE, 2023.

[224] H.-J. H. Smilde, A. den Boef, M. Kubis, M. Jak, M. van Schijndel, A. Fuchs, M. van der Schaar, S. Meyer, S. Morgan, J. Wu, *et al.*, "Evaluation of a novel ultra small target technology supporting on-product overlay measurements," in *Metrology, Inspection, and Process Control for Microlithography XXVI*, vol. 8324, pp. 431–438, SPIE, 2012.

[225] J. Maas, M. Ebert, K. Bhattacharyya, H. Cramer, A. Becht, S. Keij, R. Plug, A. Fuchs, M. Kubis, T. Hoogenboom, *et al.*, "YieldStar: a new metrology platform for advanced lithography control," in *27th European Mask and Lithography Conference*, vol. 7985, pp. 146–155, SPIE, 2011.

[226] T. Liu, Y. Jiao, and E. Bente, "Design of a computer-generated waveguide hologram for integrated free space sensing," in *2021 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD)*, pp. 95–96, IEEE, 2021.

[227] D. De Vocht, T. Liu, Y. Jiao, and E. Bente, "Integrated computer generated waveguide hologram for versatile free-space beam projection," in *23rd European Conference on Integrated Optics*, pp. 134–136, 2022.

[228] M. De Reus, *On Electromagnetically Induced Fluid Flows in Water.* Technische Universiteit Delft, 2013.

[229] L. Zou, W. Withayachumnankul, C. M. Shah, A. Mitchell, M. Bhaskaran, S. Sriram, and C. Fumeaux, "Dielectric resonator nanoantennas at visible frequencies," *Optics Express*, vol. 21, no. 1, pp. 1344–1352, 2013.

[230] P. R. West, J. L. Stewart, A. V. Kildishev, V. M. Shalaev, V. V. Shkunov, F. Strohkendl, Y. A. Zakharenkov, R. K. Dodds, and R. Byren, "All-dielectric subwavelength metasurface focusing lens," *Optics Express*, vol. 22, no. 21, pp. 26212–26221, 2014.

[231] A. Arbabi, Y. Horie, A. J. Ball, M. Bagheri, and A. Faraon, "Subwavelength-thick lenses with high numerical apertures and large efficiency based on high-contrast transmitarrays," *Nature Communications*, vol. 6, no. 1, p. 7069, 2015.

[232] B. C. Kress and P. Meyrueis, *Applied Digital Optics: From Micro-Optics to Nanophotonics.* John Wiley & Sons, 2009.

[233] F. Silvestri, G. Gerini, E. Pisano, and V. Galdi, “High numerical aperture all-dielectric metasurface micro-lenses,” in *2015 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting*, pp. 1030–1031, IEEE, 2015.

[234] F. Silvestri, G. Gerini, S. M. Bäumer, and E. J. Van Zwet, “Robust design procedure for dielectric resonator metasurface lens array,” *Optics Express*, vol. 24, no. 25, pp. 29153–29169, 2016.

[235] L. P. Stoevelaar, J. Berzinš, F. Silvestri, S. Fasold, K. Z. Kamali, H. Knopf, F. Eilenberger, F. Setzpfandt, T. Pertsch, S. M. Bäumer, *et al.*, “Nanostructure-modulated planar high spectral resolution spectro-polarimeter,” *Optics Express*, vol. 28, no. 14, pp. 19818–19836, 2020.