THE NUMERICAL ANALYSIS OF RCWA AND ITS USE IN SIMULATING THIN-FILM SOLAR CELLS

by

Benjamin J. Civiletti

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics

Summer 2020

THE NUMERICAL ANALYSIS OF RCWA AND ITS USE IN SIMULATING THIN-FILM SOLAR CELLS

by

Benjamin J. Civiletti

Approved: _____
Louis Rossi, Ph.D.
Chair of the Department of Mathematical Sciences

Approved: _____
John Pelesko, Ph.D.
Dean of the College of Arts and Sciences

Approved: _____
Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Peter Monk, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Constantin Bacuta, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Philippe Guyenne, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Daniel Weile, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor Prof. Peter Monk for his encouragement and support during my time as a graduate student. Without his countless useful suggestions, this thesis would not exist. I would also like to acknowledge Prof. Francisco-Javier Sayas who taught me functional analysis. Sadly, he passed away in April 2019.

I would like to acknowledge all the people at Villanova and TCNJ who contributed to my mathematical knowledge. In particular, I would like to thank Prof. Timothy Feeman for his inspiring lectures and kindness. For all the friends I met at UD, especially those in my cohort, I extend my gratitude. I would also like to acknowledge the staff at UD for keeping everything working, especially Deborah See.

I want to give special thanks to my family for all their support over the years: my parents for encouraging me to pursue academic endeavors, and my brother Matt for his friendship. Finally, I would like to acknowledge my girlfriend Megan for all her support and love.

This work was financially supported by the NSF under grant number DMS-1619904.

TABLE OF CONTENTS

Appendix

LIST OF FIGURES

ABSTRACT

The ability to rapidly compute the electromagnetic field inside a thin-film solar cell given a set of constitutive parameters of the cell is an important tool in the field of solar cell optimization. Many numerical methods exist to model diffraction of electromagnetic waves by periodic gratings. Among these, spectral methods are popular for their computational speed and easy application to many different grating geometries. The main theme of this thesis is to study the convergence properties of a quasi-spectral method called the Rigorous Coupled-Wave Approach (RCWA). To do this, we investigate the relevant scattering problems: scattering of an s- or p-polarized incident plane-wave by an inhomogenous, periodic medium. In each of these two cases, we establish convergence of the RCWA under appropriate assumptions on the constitutive parameters of the cell. In many cases we consider, the variational formulation is not coercive so we employ Rellich identities to establish convergence. We present numerical examples to test our prediction of the convergence rate, which suggests that our analytical results could be pessimistic. We also study a new spectral method that combines the RCWA with transformation optics, whereby the domain is first mapped to a simpler one using a coordinate transform. Our preliminary numerical tests demonstrate that this hybrid method could be more stable than the standard RCWA, and might extend to 3D to rapidly solve the full 3D Maxwell system in crossed gratings.

Chapter 1

INTRODUCTION

Direct electromagnetic scattering is concerned with determining the unknown field everywhere, given information about an object illuminated by a known incident field. The object may be, for example, a bounded scatterer, a rough layer that is infinite in extent but not perfectly periodic, or periodic. The incident field may be a plane wave (or the superposition of plane waves) or due to a point source; but in any case, questions arise about the *well-posedness* of the scattering problem. In the sense of Hadamard [50], a scattering problem is *well-posed* if

1. a solution to the scattering problem exists,

2. the solution is unique, and

3. the solution depends continuously on the data.

When the object is sufficiently complicated, a solution to the scattering problem cannot be found analytically. This is the case with many scientific and engineering applications, and so great care has to be taken to ensure that the resulting scattering problems are *well-posed*. Some basic examples of applications of these problems are a bounded obstacle placed in a homogeneous space, as with the case for radar technology. Or perhaps the scatterer is bounded in one direction, but has a periodic micro structure in the other, as with the case of photonic crystals and thin-film solar cells. We refer the interested reader to [23] for more information about applications of these problems.

Here, we consider the problem of time-harmonic electromagnetic scattering of a plane wave u^i incident on an isotropic medium that is bounded in one direction and

periodic in another. An example of such a domain is illustrated in Figure 1.1. Since the object is periodic (and infinite in extent), it is desirable to only consider a scattering problem in a domain containing only one period. It turns out that the true solutions to these problems are *quasi-periodic* (see Section 2.1.1 for a definition) in the same direction (and with the same period) as the medium. Therefore, knowing the solution in a unit cell is sufficient to knowing it in the whole domain, as the solution is periodic up to some *phase factors*.

The application of this mathematical problem that we are interested in is the modeling of thin-film solar cells. Typically, these photonic devices consist of many layers of differing materials, and their thickness is on the order of the wavelength of the incident light (around 1000 nm). In this regime, the full-wave Maxwell's equations must be solved to properly model and evaluate the device on a computer.

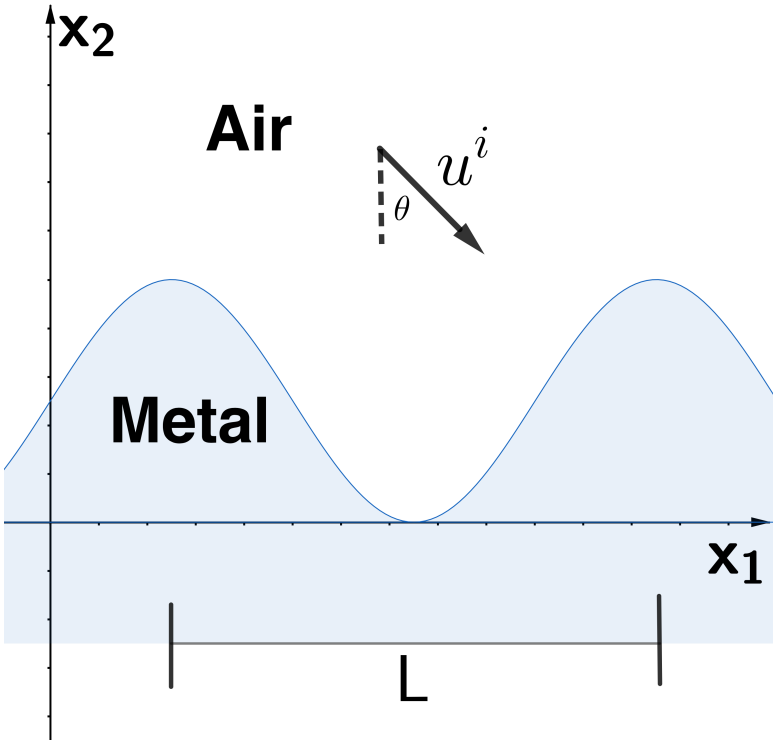


Figure 1.1: A grating structure (e.g., an interface) is illuminated by a monochromatic plane wave u^i with an angle of incidence θ. The metallic grating is x_1–periodic with period $L > 0$.

At the interface between two material layers, there is a jump in the electromagnetic properties (relative permittivity ε) of the materials. Therefore, we are interested in problems where the isotropic coefficients of the relevant scattering problems are piecewise smooth, but have jumps across the interfaces. In this regime, the true solutions to these problems often have quite low Sobolev regularity depending on the choice of ε and how smooth the interfaces are. Furthermore, it is common for thin-film solar cell designs to include a undulating periodic surface called a grating. The aim of such a design is to enhance the trapping of photons inside the solar cell over the usable solar spectrum, in the hope of increasing efficiency. Since the useful solar spectrum is quite complicated, there is a need for multifrequency results in solar cell modeling. In particular, the AM1.5G solar spectrum [55] is illustrated in Figure 1.2, is used as a standard in solar cell research to provide the relative amplitude of the incident wave at different frequencies.

Many numerical methods exist for these problems, such as the Finite Element Method (FEM), the Finite Difference Time Domain (FDTD) method and integral equation methods [56]. Usually optimal design of a cell requires the solution of many forward problems, so computational speed and easy application to all kinds of grating geometry is required. Therefore, any successful numerical method has to to resolve the solution in a reasonable amount of computational time, while using the available computational resources. For the FEM, for example, any change in geometry requires remeshing that can be computationally expensive. For the integral equation method, great care needs to be taken to find the fundamental solutions, but the Rigorous Coupled-Wave Approach (RCWA) method avoids this. Since many thousands of scattering problems are solved for the optimization problem, the RCWA method is often used to perform the optical calculations.

At its core, the RCWA is a meshless method that exploits the fact that the true solution is quasi-periodic and can be represented as Fourier series. The periodic relative permittivity can also be expressed in terms of a Fourier series, and so in this sense the RCWA is a spectral method since it solves for the Fourier coefficients

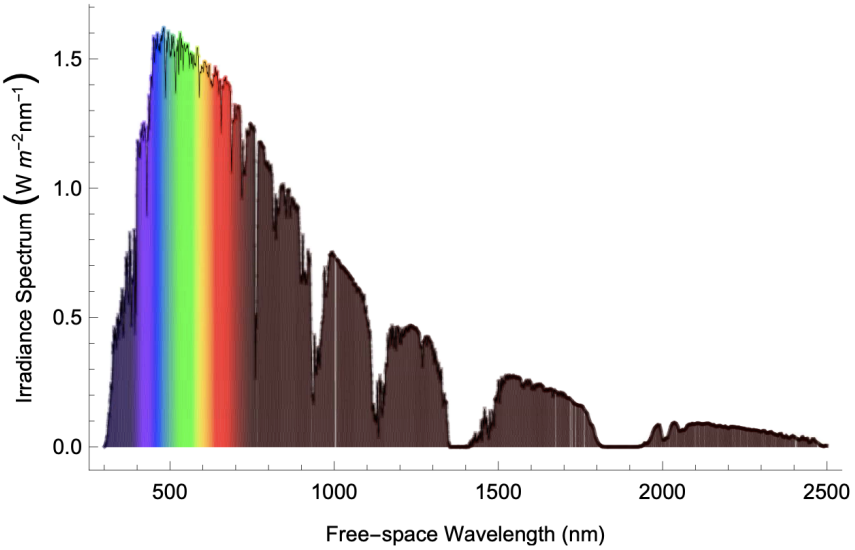


Figure 1.2: The AM1.5G solar spectrum [55], showing the spectral irradiance as a function of the free-space wavelength λ_0 (nm) of light.

of the solution. When these Fourier representations are substituted into Maxwell's equations, one obtains an infinite system of Ordinary Differential Equations (ODEs) and algebraic equations for the coefficients. Obviously, this system must be truncated in order to solve for some finite number of coefficients on a computer. However, even after truncation, this system is difficult to solve.

Therefore, the RCWA introduces another form of discretization by replacing the true relative permittivity by a stairstep approximation. The goal of this approximation is to make the finite system of ODEs and algebraic equations amenable to a fast solution algorithm. To do this, a domain decomposition scheme is performed whereby the domain is split into thin horizontal slices stacked in the direction perpendicular to the periodicity of the grating. In each slice, by design, the solution is just a superposition of upward and downward propagating modes. Continuity is enforced across the inter-slice boundaries and appropriate boundary conditions are satisfied to calculate the solution in the entire domain.

The RCWA has its roots in coupled-wave analysis for diffraction problems, e.g., in a single layer with a sinusoidal spatial variation of the relative permittivity [3].

The formal approach was proposed in the early 1980s by Moharam and Gaylord [4]. Around that time, rectangular grating profiles became more common due to advancement in manufacturing techniques such as binary optics and the ion-etching technique. Therefore, the need to approximate the electromagnetic field in such devices became an interest to many. Indeed, the first form of the RCWA assumed rectangular grating profiles. Peng, Tamir and Bertoni [53] along with Moharam and Gaylord thought it possible to extend the method to arbitrary grating profiles using a stairstep approximation. At the time, the validity of this approximation was not checked mathematically. Instead, it was assumed that the numerical solution would converge to the true solution as the size of the stairstep discretization decreased and number of Fourier modes increased.

Subsequently, the near-field convergence with respect to the number of retained Fourier modes was drastically improved by Li [5]. The RCWA approach is now a workhorse for obtaining rapid simulations in a grating. It has been used, for example, to study the excitation of surface plasmon-polariton waves for optical sensing [6] and in the design process of solar cells [7, 8, 1, 12]. Some open problems for the RCWA were discussed by Hench and Strakoš [9]. One open problem discussed is whether the discretized solution approximates the true solution, and if so, to what order. We address this open problem in this thesis.

In light of this discussion, we conclude this introduction by describing the goal of this thesis: to study the mathematical theory of relevant electromagnetic scattering problems in a periodic, bounded domain, and use this theory for the continuous problems to help prove that the RCWA solution converges. Since the application requires the use of many types of materials, we prove our theory for quite general piecewise smooth relative permittivity (the relative permeability is set to unity because solar cells are non magnetic). We try, whenever possible, to allow different signs for the real and imaginary parts of the relative permittivity. For example, certain metals (e.g., gold) have a relative permittivity where the real part is negative and the imaginary part is positive at some frequencies.

This thesis is organized as follows. In Chapter 2, we define the scattering problems that we will study, To do this, under the assumption of translation invariance of the domain, the time-harmonic Maxwell's equations can be expressed as a pair of scalar Helmholtz equations. In the infinite half-spaces above and below the domain, we show that the solution to either of the Helmholtz equations can be expressed using a Rayleigh expansion. This expansion, along with Dirichlet-to-Neumann (DtN) maps, will be used to enforce the appropriate radiation conditions in the variational problems. We conclude by recalling some properties of the DtN maps and the resulting sesquilinear form related to the variational problems.

In Chapter 3 we derive the RCWA method. To do this, we define the stairstep approximation of interfaces and give the relevant perturbed variational problems. We also show how substitution of the Fourier representations for the fields and relative permittivity into Maxwell's equations yields an infinite system of ODEs and algebraic equations. For the truncated system with truncation parameter M, we describe the RCWA solution and the boundary conditions that are enforced across the inter-slice boundaries, and also at the top and bottom boundaries. We prove that the RCWA method is actually a Galerkin scheme, i.e., the RCWA solution solves the appropriate variational problems.

In Chapter 4, we consider scattering of an incident plane wave that is s-polarized. We prove a Rellich identity for solutions to the variational problem, and show that under certain non-trapping conditions for the coefficients that an *a-priori* estimate holds for this problem for right hand sides in L^2. Our approach to this problem involves deriving an explicit continuity constant that can be shown to hold for the original problem as well as the perturbed problem. Hence, the perturbed problem depends continuously of the data uniformly in the slice thickness h. We then show how the a-priori estimates imply that the solution is unique. Finally, we prove that the RCWA solution converges for this case, and prove the order of convergence in parameters M and h. The results of this chapter have been published in [36].

In Chapter 5, we consider scattering of an incident plane wave that is p-polarized. This chapter is similar to Chapter 4, but the solutions to the Helmholtz problem in this case have, generally speaking, lower Sobolev regularity. Another difficulty is the right hand side of the Helmholtz equation for the scattered field is not in L^2. Therefore, to mitigate some of the difficulty we derive a Rellich identity for the case where the relative permittivity is C^∞ in $\mathbb{R}^2$ and the right hand side F is in L^2. We prove that an *a-priori* estimate holds in this case, and show how to extend the estimate to cases where the right hand side is in the dual space or where the coefficients are only L^∞. To derive these estimates, the relative permittivity ε must satisfy certain non-trapping conditions and a density argument is used to obtain the results. Finally, we prove that the RCWA method converges for this case and prove the order of convergence in parameters M and h. We also consider a case in which all the materials in the grating are absorbing where the problem is coercive

In Chapter 6 we investigate the C Method, a spectral method that uses a coordinate transformation to deal with the grating interface. At its core, this method exploits the fact that Maxwell's equations (and therefore the subsequent Helmholtz equations) are invariant to coordinate transforms up to the coefficients. For this problem, this means that a trade-off is made where the geometry is simpler but the isotropic coefficients in the original variables are anisotropic in the transformed variables. However, the benefit is that the grating interfaces are not approximated (other than sampling them with the Fast Fourier Transform) and so no additional error is introduced. We derive a hybrid C-RCWA method, and show preliminary numerical results.

After presenting some conclusions and directions for future research, we provide the RCWA codes used in the numerical experiments presented in this thesis.

Chapter 2

THE SCATTERING PROBLEM

2.1 Maxwell's Equations in Periodic Media

Throughout this thesis, the Euclidean coordinates $\boldsymbol{x} = (x_1, x_2, x_3)$ are used, along with the associated unit vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$, and $\boldsymbol{e}_3$. We assume that the domain is periodic in the x_1 direction with period $L > 0$, and invariant in the x_3 direction. Because of this choice of domain, Maxwell's equations can be simplified to a scalar Helmholtz equation, depending on the polarization of the incident field. We consider linear optics with an $\exp(-i\omega t)$ dependence on time t, where $i = \sqrt{-1}$ and ω is the angular frequency of light. Any matrix $\tilde{A}(\boldsymbol{x}, t)$ will be represented by an associated complex matrix $A(\boldsymbol{x})$, where $\tilde{A}(\boldsymbol{x}, t) = \Re[A(\boldsymbol{x}) \exp(-i\omega t)]$. We define the complex vector fields $\boldsymbol{E} = E_1\boldsymbol{e}_1 + E_2\boldsymbol{e}_2 + E_3\boldsymbol{e}_3$ and $\boldsymbol{H} = H_1\boldsymbol{e}_1 + H_2\boldsymbol{e}_2 + H_3\boldsymbol{e}_3$ to be the electric and magnetic field phasors, respectively. The time-harmonic Maxwell's equations then have the form

$$\nabla \times \boldsymbol{E} = i\omega\mu_0\mu\boldsymbol{H}, \tag{2.1}$$

$$\nabla \times \boldsymbol{H} = -i\omega\varepsilon_0\varepsilon\boldsymbol{E}, \tag{2.2}$$

where μ is the relative magnetic permeability, μ_0 is the permeability of free-space, ε is the relative permittivity and ε_0 is the permittivity of free-space, i.e. a vacuum. By assumption, $\varepsilon = \varepsilon(x_1, x_2)$ and ε is x_1–periodic with period L. In addition, for the

solar cell application we assume that $\mu = 1$ since no magnetic components are present. Component-wise, we obtain six coupled partial differential equations (PDEs), namely

$$\frac{\partial E_3}{\partial x_2} - \frac{\partial E_2}{\partial x_3} = i\omega\mu_0 H_1, \tag{2.3}$$

$$\frac{\partial E_1}{\partial x_3} - \frac{\partial E_3}{\partial x_1} = i\omega\mu_0 H_2, \tag{2.4}$$

$$\frac{\partial E_2}{\partial x_1} - \frac{\partial E_1}{\partial x_2} = i\omega\mu_0 H_3, \tag{2.5}$$

$$\frac{\partial H_3}{\partial x_2} - \frac{\partial H_2}{\partial x_3} = -i\omega\varepsilon_0\varepsilon E_1, \tag{2.6}$$

$$\frac{\partial H_1}{\partial x_3} - \frac{\partial H_3}{\partial x_1} = -i\omega\varepsilon_0\varepsilon E_2, \tag{2.7}$$

$$\frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2} = -i\omega\varepsilon_0\varepsilon E_3. \tag{2.8}$$

Throughout, we assume that there is a height $H > 0$ large enough so that $\varepsilon(\boldsymbol{x}) = \varepsilon_+$ whenever $x_2 > H$, and $\varepsilon(\boldsymbol{x}) = \varepsilon_-$ for $x_2 < -H$, where ε_+ and ε_- are real positive constants. The domain is illuminated with a downward propagating incident plane wave, where the incident electric and magnetic fields are given by

$$\boldsymbol{E}^{inc}(\boldsymbol{x}) = \boldsymbol{A}\exp\left[i\sqrt{\varepsilon_+}\kappa\left(x_1\sin\theta - x_2\cos\theta\right)\right], \tag{2.9}$$

$$\boldsymbol{H}^{inc}(\boldsymbol{x}) = \eta_0^{-1}\boldsymbol{B}\exp\left[i\sqrt{\varepsilon_+}\kappa\left(x_1\sin\theta - x_2\cos\theta\right)\right], \tag{2.10}$$

respectively. Here, $\boldsymbol{A}$ and $\boldsymbol{B}$ are polarization vectors, θ is the angle of incidence with respect to the x_2 axis, and the wavenumber in air is denoted by $\kappa = \omega/c_0$ where $c_0 = 1/\sqrt{\varepsilon_0\mu_0}$ is the speed of light in air and $\eta_0 = \sqrt{\mu_0/\varepsilon_0}$.

Because the incident field is divergence-free, $\boldsymbol{A}$ must satisfy $\boldsymbol{A}\cdot\left(\boldsymbol{e}_1\sin\theta - \boldsymbol{e}_2\cos\theta\right) = 0$. Also, $\boldsymbol{B}$ must satisfy a similar equation. The polarization vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ are not independent because $\left(\boldsymbol{E}^{inc}, \boldsymbol{H}^{inc}\right)$ satisfies Maxwell's equations with $\varepsilon = \varepsilon_+$. When the polarization vector $\boldsymbol{A} = (0, 0, 1)$ the incident plane-wave is said to be s-polarized, and when the polarization vector $\boldsymbol{A} = (-\cos\theta, \sin\theta, 0)$ the incident field is p-polarized. The polarization vector $\boldsymbol{B} = (-\cos\theta, \sin\theta, 0)$ in the s-polarization case and $\boldsymbol{B} = (0, 0, -1)$ in the p-polarization case. Both the incident plane-wave and the geometry are invariant in the x_3 direction, so therefore all electromagnetic fields are

assumed to be invariant in the x_3 direction. Thus, the partial derivatives with respect to x_3 in the system (2.3)–(2.8) are all zero. We obtain the following system of PDEs:

$$\frac{\partial E_3}{\partial x_2} = i\omega\mu_0 H_1, \tag{2.11}$$

$$-\frac{\partial E_3}{\partial x_1} = i\omega\mu_0 H_2, \tag{2.12}$$

$$\frac{\partial E_2}{\partial x_1} - \frac{\partial E_1}{\partial x_2} = i\omega\mu_0 H_3, \tag{2.13}$$

$$\frac{1}{\varepsilon}\frac{\partial H_3}{\partial x_2} = -i\omega\varepsilon_0 E_1, \tag{2.14}$$

$$\frac{1}{\varepsilon}\frac{\partial H_3}{\partial x_1} = i\omega\varepsilon_0 E_2, \tag{2.15}$$

$$\frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2} = -i\omega\varepsilon_0\varepsilon E_3. \tag{2.16}$$

This leads us to study two sets of decoupled equations for the two independent polarization states. The first given by (2.11), (2.12) and (2.16) applies to the s-polarization state with the sole non-zero component of the electric field being E_3. The second state is the p-polarization state given by (2.13)–(2.15), with the sole non-zero component of the magnetic field being H_3. In the literature, these are sometimes referred to as transverse electric (TE) and transverse magnetic (TM) polarization state, respectively. We can further simplify these equations, by eliminating all but the x_3 component of the electric field E_3 in the s-polarization case, and by eliminating all but the x_3 component of the magnetic field H_3 in the p-polarization case. For s-polarization, we obtain the scalar Helmholtz equation

$$\Delta E_3 + \kappa^2\varepsilon E_3 = 0. \tag{2.17}$$

For the p-polarization case, a similar substitution is made, and we obtain the scalar Helmholtz equation

$$\nabla \cdot \left(\frac{1}{\varepsilon}\nabla H_3\right) + \kappa^2 H_3 = 0. \tag{2.18}$$

Thus, by looking for translation invariant solutions to Maxwell's equations, the problem is greatly simplified. Later in this thesis, we will discuss the uniqueness and existence of the solutions of (2.17) and (2.18) under suitable boundary and radiation conditions. For

the remainder of this thesis, we denote E_3 or H_3 by u^t, depending on the polarization state (here the superscript t denotes u^t to be the total physical field in the system). It is useful in our analysis to also consider the scattered field $u = u^t - u^i$, where u^i is the incident field, so therefore we summarize the results of this section using the scattered field. We may compute an approximation to the total or scattered field. For the scattered field we seek the solution u to the Helmholtz equation

$$\nabla \cdot \Big(A \nabla u \Big) + \kappa^2 a u = f. \tag{2.19}$$

In (2.19), the s-polarization state corresponds to $A = \boldsymbol{I}$, $a = \varepsilon$ and $f = \kappa^2 u^i(\varepsilon_+ - \varepsilon)$. The downward propagating plane wave $u^i = \boldsymbol{E}^{inc} \cdot \boldsymbol{e}_3$. Similarly, the p-polarization state corresponds to $A = \frac{1}{\varepsilon}\boldsymbol{I}$, $a = 1$ and $f = \nabla \cdot \left[(\varepsilon_+^{-1} - \varepsilon^{-1})\nabla u^i\right]$. In this case, u^i is taken to be $\boldsymbol{H}_3^{inc} \cdot \boldsymbol{e}_3$. We discuss boundary and radiation conditions next.

2.1.1 Quasi-periodicity of the Solutions

In this section we show that, to obtain physically relevant solutions to the Helmholtz equation (2.19), u has to have the property

$$u(x_1 + L, x_2) = \exp(i\alpha_0 L) u(x_1, x_2), \tag{2.20}$$

where $\alpha_0 = \kappa\sqrt{\varepsilon_+}\sin\theta$. Throughout, if (2.20) holds we say that u is *quasi-periodic* with period L. Mathematically, this condition can be derived using the Floquet-Bloch transform [57]. Here we use a heuristic argument from [23] to see that the condition is needed. To show this, we note that the incident field (in either polarization state) is quasi-periodic with period L, and also solves the Helmholtz equation

$$\Delta u^i + \kappa^2 \varepsilon_+ u^i = 0. \tag{2.21}$$

By considering the scattered field u, we see that for the s-polarization state

$$\Delta u + \kappa^2 \varepsilon u = \kappa^2 u^i(\varepsilon_+ - \varepsilon). \tag{2.22}$$

A translation by L in x_1 on the right hand side yields a factor of $\exp(i\alpha_0 L)$, but the left hand side is invariant to this translation because ε is periodic with period L. We

see that $\exp(-i\alpha_0 L)u$ is also a solution of (2.22). If u is not quasi-periodic, then this process will generate a new solution, and we have no hope of having unique solutions to the Helmholtz equation (2.17). The same argument can be made for the p-polarization state, and so throughout this thesis we assume that u is quasi-periodic with period L.

2.1.2 Rayleigh Expansion of the Solutions

From the previous discussion, since u is quasiperiodic, we know that $\exp(-i\alpha_0 L x_1)u$ is periodic in x_1 with period L. Therefore it can be written as a Fourier series, namely

$$\exp(-i\alpha_0 L x_1)u(x_1, x_2) = \sum_{n\in\mathbb{Z}} u_n(x_2) \exp\left(\frac{2\pi i n}{L} x_1\right), \tag{2.23}$$

where the $u_n(x_2)$ are the coefficient functions. But then any quasi-periodic solution u to the Helmholtz equation (2.19) to can be written as

$$u(x_1, x_2) = \sum_{n\in\mathbb{Z}} u_n(x_2) \exp(i\alpha_n x_1), \qquad \boldsymbol{x} \in \mathbb{R}^2, \tag{2.24}$$

where $\alpha_n = \alpha_0 + (2\pi n)/L$. In the two half-spaces $x_2 > H$ and $x_2 < H$, the Helmholtz equation (2.19) (in either polarization state) can be written as

$$\begin{aligned} \Delta u + \kappa^2 \varepsilon_+ u = 0, \qquad & x_2 > H, \\ \Delta u + \kappa^2 \varepsilon_- u = 0, \qquad & x_2 < -H. \end{aligned} \tag{2.25}$$

Plugging the representation (2.24) into each Helmholtz equation (2.25) we get

$$\sum_{n\in\mathbb{Z}} \left[\frac{\partial^2}{\partial x_2^2} + (\kappa^2\varepsilon_+ - \alpha_n^2)\right] u_n(x_2) \exp\left(\frac{2\pi i n}{L} x_1\right) = 0. \tag{2.26}$$

By Parseval's Theorem, each Fourier coefficient function $u_n(x_2)$ solves the ordinary differential equation (ODE)

$$\left[\frac{d^2}{dx_2^2} + (\kappa^2\varepsilon_+ - \alpha_n^2)\right] u_n(x_2) = 0, \tag{2.27}$$

as long as the square root of $\kappa^2\varepsilon_+ - \alpha_n^2$ is well defined. To this end, we set

$$\beta_n^{\pm} = \begin{cases} \sqrt{\kappa^2\varepsilon_+ - \alpha_n^2} & \alpha_n^2 < \kappa^2\varepsilon_+, \\ i\sqrt{\alpha_n^2 - \kappa^2\varepsilon_+} & \alpha_n^2 > \kappa^2\varepsilon_+. \end{cases} \tag{2.28}$$

Thus, in the half-spaces $x_2 > H$ and $x_2 < -H$, u can be expressed as

$$u(x_1, x_2) = \sum_{n\in\mathbb{Z}} A_n \exp(i\alpha_n x_1 - i\beta_n^{\pm} x_2) + \sum_{n\in\mathbb{Z}} B_n \exp(i\alpha_n x_1 + i\beta_n^{\pm} x_2), \tag{2.29}$$

where the B_n and A_n are integrating constants. The two expressions on the right hand side of (2.29) generate two different types of solutions. In the first term, the solution is generated by a finite number of downward propagating plane waves (for $\kappa^2\varepsilon^{\pm} - \alpha_n^2 > 0$), and an infinite number of evanescent waves (for $\kappa^2\varepsilon^{\pm} - \alpha_n^2 < 0$) that diverge as $x_2 \to +\infty$, and decay exponentially as $x_2 \to -\infty$. The second term is generated by a finite number of upward propagating plane waves (for $\kappa^2\varepsilon^{\pm} - \alpha_n^2 > 0$), and an infinite number of evanescent waves that diverge as $x_2 \to -\infty$, and decay exponentially as $x_2 \to \infty$. In the half-space $x_2 \geq H$, we choose that the modes need to be upward propagating or evanescent upwards, so that

$$u(x_1, x_2) = \sum_{n\in\mathbb{Z}} u_n(H) \exp[i(x_2 - H)\beta_n^{+}] \exp(i\alpha_n x_1), \tag{2.30}$$

for $x_2 \geq H$. In the half-space $x_2 \leq -H$, we choose that the modes are downward propagating or evanescent downwards, so that

$$u(x_1, x_2) = \sum_{n\in\mathbb{Z}} u_n(-H) \exp[-i(x_2 + H)\beta_n^{-}] \exp(i\alpha_n x_1), \tag{2.31}$$

valid for all $x_2 \leq -H$. The representations in (2.30) and (2.31) are called the Rayleigh expansions of u. In this scattering problem, the choice to make u satisfy the Rayleigh expansions above and below the periodic media is the *radiation condition* on the solution. Unlike with other scattering problems, such as scattering from bounded media, the Sommerfeld radiation condition [61] is not appropriate in this case.

2.2 Geometry of the Scattering Problem

In this section we define the domain to be used in the remainder of this thesis. The rectangular domain Ω is given by

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^2, 0 < x_1 < L, -H < x_2 < H\}, \tag{2.32}$$

that includes one period of ε. Because a solution u to the Helmholtz equation (2.19) is quasi-periodic with period L, we only need to include one period of ε, and require that u satisfies quasi-periodic boundary conditions on the right boundary $\Gamma_R = \{\boldsymbol{x} \in \mathbb{R}^2, x_1 = L, -H < x_2 < H\}$ and left boundary $\Gamma_L = \{\boldsymbol{x} \in \mathbb{R}^2, x_1 = 0, -H < x_2 < H\}$. Above and below Ω, we know that the solution can be written as a Rayleigh expansion via (2.30) and (2.31), respectively. By virtue of these representations, we note that the restriction of u to the upper boundary $\Gamma_H = \{\boldsymbol{x} \in \mathbb{R}^2, 0 < x_1 < L, x_2 = H\}$ completely defines the solution in the domain $\Omega^+ = \{\boldsymbol{x} \in \mathbb{R}^2, 0 < x_1 < L, x_2 > H\}$. Similarly, the restriction of u to the bottom boundary $\Gamma_{-H} = \{\boldsymbol{x} \in \mathbb{R}^2, 0 < x_1 < L, x_2 = -H\}$ defines the solution in $\Omega^- = \{\boldsymbol{x} \in \mathbb{R}^2, 0 < x_1 < L, x_2 < H\}$. The geometry of the scattering problem is shown in Fig. 2.1.

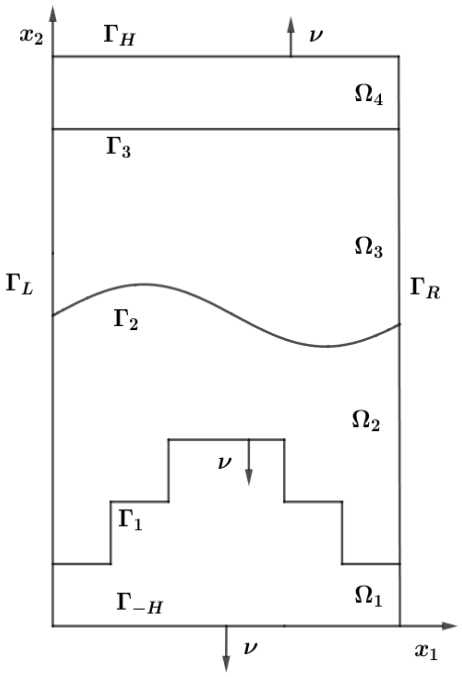


Figure 2.1: Geometry of the scattering problem, with $I = 3$ interfaces. The domain Ω lies between the two lines $\Gamma_H = \{\boldsymbol{x} \in \mathbb{R}^2, 0 < x_1 < L, x_2 = H\}$ and $\Gamma_{-H} = \{\boldsymbol{x} \in \mathbb{R}^2, 0 < x_1 < L, x_2 = -H\}$, so that $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2 \cup \cdots \cup \overline{\Omega}_{I+1}$. In each Ω_k the relative permittivity ε is assumed to be in C^2, but can jump over each interface Γ_k. The quasi-periodic boundaries are $\Gamma_R = \{\boldsymbol{x} \in \mathbb{R}^2, x_1 = L, -H < x_2 < H\}$ and $\Gamma_L = \{\boldsymbol{x} \in \mathbb{R}^2, x_1 = 0, -H < x_2 < H\}$. The interface Γ_1 is termed a stairstep.

In many applications, ε is assumed to be piecewise constant, but we are also

interested in the case where ε is piecewise smooth, in order to model graded materials that improve the efficiency of solar cells [12, 13]. Therefore, ε is assumed to be piecewise smooth in $\mathbb{R}^2$ and in general has jumps over some finite number of non-intersecting piecewise smooth interfaces.

Inside Ω, we assume that there are I interfaces $\hat{\Gamma}_k$ for $1 \leq k \leq I$. The interfaces are defined as

$$\hat{\Gamma}_k = \{\boldsymbol{x} \in \mathbb{R}^2, g_k(x_1) = x_2\}, \tag{2.33}$$

where $g_k : \mathbb{R} \to \mathbb{R}$ is a C^2 function except possibly at a finite number of values $x_{1k}, x_{2k}, \cdots, x_{N_k k}$. Let $\hat{H}(x)$ be the Heaviside function, and

$$\hat{\Pi}_{ab} = \hat{H}(x_1 - a) - \hat{H}(x_1 - b). \tag{2.34}$$

Then the g_k can be written as

$$g_k = \sum_{l=0}^{N_k} \hat{\Pi}_{x_{lk} x_{(l+1)k}} g_{lk}, \tag{2.35}$$

where the g_{lk} are C^2 functions with $x_{0k} = 0$ and $x_{(N_k+1)k} = L$. At the discontinuities, we require that

$$[g_k]_{x_{lk}} = g_k(x_{lk}^+) - g_k(x_{lk}^-) \neq 0, \tag{2.36}$$

for all $1 \leq k \leq I$ and $1 \leq l \leq N_k$, where $g_k(x_{lk}^+)$ is the limit taken from the right and $g_k(x_{lk}^-)$ is the limit taken from the left. We define the values $g_{lk}^+ = \max\{g_k(x_{lk}^+), g_k(x_{lk}^-)\}$ and $g_{lk}^- = \min\{g_k(x_{lk}^+), g_k(x_{lk}^-)\}$ along with the sets

$$W_{lk} = \{\boldsymbol{x} \in \mathbb{R}, x_1 = x_{lk}, g_{lk}^- \leq x_2 \leq g_{lk}^+\}, \tag{2.37}$$

for $1 \leq k \leq I$ and $1 \leq l \leq N_k$. We therefore define a stairstep interface to be

$$\Gamma_k = \hat{\Gamma}_k \cup \left(\bigcup_{l=1}^{N_k} W_{lk} \right). \tag{2.38}$$

An illustration of a suitable domain Ω with three interfaces is given in Fig. 2.1. We require that the interfaces do not intersect, so that for some $\delta > 0$, we have

$$\delta + \max_{0 \leq x_1 \leq L} g_{k-1}(x_1) < g_k(x_1) < -\delta + \min_{0 \leq x_1 \leq L} g_{k+1}(x_1) \tag{2.39}$$

for all $2 \leq k \leq I-1$, and the interfaces are bounded away from Γ_H and Γ_{-H}, namely

$$\delta - H < g_1(x_1) < -\delta + \min_{0 \leq x_1 \leq L} g_2(x_1), \tag{2.40}$$

$$\max_{0 \leq x_1 \leq L} g_{k-1}(x_1) < g_I(x_1) < -\delta + H. \tag{2.41}$$

Thus, the interfaces Γ_k separate Ω into $I+1$ subdomains, namely

$$\Omega_k = \{\boldsymbol{x} \in \Omega : \ g_{k-1}(x_1) < x_2 < g_k(x_1)\}, \tag{2.42}$$

for $1 \leq k \leq I+1$, where $g_0(x_1) = -H$ and $g_{I+1}(x_1) = H$.

2.3 Background Theory

In order to rigorously define the variational problems, we quickly review and define the Sobolev spaces and norms used in the remainder of this thesis. For a general reference see [51, 17]. We recall that the Sobolev space

$$H^1_{loc}(\mathbb{R}^2) := \{w \in \mathbb{R}^2 \to \mathbb{C}, w \in H^1(\mathcal{O}) \text{ for any bounded, open set } \mathcal{O} \subset \mathbb{R}^2\}, \tag{2.43}$$

where the Sobolev space $H^1(\mathcal{O})$ is defined as

$$H^1(\mathcal{O}) := \{w \in L^2(\mathcal{O}), \nabla w \in L^2(\mathcal{O})^2\}, \tag{2.44}$$

and ∇ is the distributional gradient. Since our problem is defined in a quasi-periodic setting, we define the strip $\Omega^* = \Omega \cup \Omega^+ \cup \Omega^-$ and

$$H^1_{qp,loc}(\Omega^*) := \{w \in H^1_{loc}(\Omega^*), w = W\big|_{\Omega^*} \text{ for some quasi-periodic } W \in H^1_{loc}(\mathbb{R}^2)\}. \tag{2.45}$$

Further, the variational formulation of the scattering problems should be defined on a bounded domain, so we seek solutions in

$$H^1_{qp}(\Omega) := \{w \in H^1(\Omega), w = W\big|_{\Omega} \text{ for some } W \in H^1_{qp,loc}(\Omega^*)\}, \tag{2.46}$$

endowed with the usual $H^1(\Omega)$ norm given as

$$\|w\|_{H^1(\Omega)} := \left(\int_\Omega |w|^2 + |\nabla w|^2 \right)^{1/2}. \tag{2.47}$$

We can also define $H^1_{qp}(\Omega)$ equivalently, as the completion of the quasi-periodic smooth functions $C^\infty_{qp} \cap H^1_{qp}(\Omega)$ in the $H^1(\Omega)$ norm. The proof of this uses mollifiers, and can be found in the standard literature (see [40]).

The space $C^{k,\alpha}(\Omega)$ consists of functions whose k-th order partial derivatives are uniformly Hölder continuous with exponent α, i.e., that

$$\sup_{\boldsymbol{x}\neq\boldsymbol{y}\in\Omega} \frac{\left|f(\boldsymbol{x}) - f(\boldsymbol{y})\right|}{\left|\boldsymbol{x}-\boldsymbol{y}\right|^\alpha} < +\infty, \tag{2.48}$$

for some $0 < \alpha \leq 1$.

The Sobolev spaces $W^{k,p}(\Omega)$ for $p \in [1,\infty)$ are defined in the usual way as

$$W^{k,p}(\Omega) := \{w \in W^k(\Omega), D^\alpha w \in L^p(\Omega) \text{ for all } |\alpha| \leq k\}, \tag{2.49}$$

where $W^k(\Omega)$ is the Banach space of k times weakly differentiable functions, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ is a multiindex of order $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2$. The norm on $W^{k,p}(\Omega)$ is

$$\|w\|_{W^{k,p}(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}|\leq k} \int_\Omega |D^{\boldsymbol{\alpha}} w|^p \right)^{1/p}. \tag{2.50}$$

For simplicity, we denote

$$H^k(\Omega) = W^{k,2}(\Omega), \tag{2.51}$$

since we are restricted to the case where $p = 2$, where H is used since these are Hilbert spaces. The spaces $H^k(\Omega)$ are endowed with the usual norm

$$\|w\|_{H^k(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}|\leq k} \int_\Omega |D^{\boldsymbol{\alpha}} w|^2 \right)^{1/2}. \tag{2.52}$$

We define the fractional Sobolev spaces, where k is not an integer, in the following way. For $k \in (0,1)$ we define $H^k(\Omega)$ as

$$H^k(\Omega) := \{w \in L^2(\Omega), \frac{\left|w(\boldsymbol{x}) - w(\boldsymbol{y})\right|}{|\boldsymbol{x}-\boldsymbol{y}|^{1+k}} \in L^2(\Omega\times\Omega)\}, \tag{2.53}$$

endowed with the norm

$$\|w\|_{H^k(\Omega)} := \left(\int_\Omega |w|^2 + \int_\Omega \int_\Omega \frac{\left|w(\boldsymbol{x}) - w(\boldsymbol{y})\right|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2+2k}} \right)^{1/2}. \tag{2.54}$$

Now for $k > 1$, we write $k = m + \sigma$ with m an integer, and define

$$H^k(\Omega) := \{w \in H^m(\Omega), D^{\boldsymbol{\alpha}} w \in H^\sigma(\Omega) \text{ for any } \boldsymbol{\alpha} \text{ s.t. } |\boldsymbol{\alpha}| = m\} : \tag{2.55}$$

a Hilbert space with the norm

$$\|w\|_{H^k(\Omega)} := \left(\|w\|^2_{H^m(\Omega)} + \sum_{|\boldsymbol{\alpha}|=m} \|D^{\boldsymbol{\alpha}} w\|^2_{H^\sigma(\Omega)} \right)^{1/2}. \tag{2.56}$$

We also let $(H^k_{qp}(\Omega))'$ be the dual space of $H^k_{qp}(\Omega)$ with the norm

$$\|F\|_{(H^k_{qp}(\Omega))'} := \sup_{0 \neq v \in H^k_{qp}(\Omega)} \frac{|F(v)|}{\|v\|_{H^1(\Omega)}}. \tag{2.57}$$

We let $H^k_{qp}(\Gamma_{\pm H})$ be the trace space of $H^{k+\frac{1}{2}}_{qp}(\Omega)$ in the usual way [52] . These Sobolev spaces on the boundary are usually endowed with the norm

$$\|w\|_{H^k(\Gamma_{\pm H})} := \left(\sum_{n \in \mathbb{Z}} (1 + n^2)^k |w_n^{\pm}|^2 \right)^{1/2}, \tag{2.58}$$

where

$$w_n^{\pm} := \frac{1}{L} \int_0^L w(\Gamma_{\pm H}) \exp(-i\alpha_n x_1)\, dx_1. \tag{2.59}$$

Finally, let $H^{-k}_{qp}(\Gamma_{\pm H})$ be the completion of $L^2_{qp}(\Gamma_{\pm H})$ in the norm

$$\|w\|_{H^{-k}_{qp}(\Gamma_{\pm H})} := \sup_{0 \neq v \in H^k_{qp}(\Gamma_{\pm H})} \frac{|(w, v)_{\Gamma_{\pm H}}|}{\|v\|_{H^k(\Gamma_{\pm H})}}. \tag{2.60}$$

We end this section by noting that the spaces $H^k_{qp}(\Gamma_{\pm H})$ can also be endowed with the equivalent norm

$$\|w\|_{H^k_{qp}(\Gamma_{\pm H})} := \left(\sum_{n \in \mathbb{Z}} |\kappa^2 - \alpha_n^2|^k |w_n^{\pm}|^2 \right)^{1/2}. \tag{2.61}$$

Also, we have chosen to avoid Rayleigh–Wood anomalies [24, 38] by assuming that $\alpha_n \neq \kappa\sqrt{\varepsilon_+}$ and $\alpha_n \neq \kappa\sqrt{\varepsilon_-}$ for any n. With these assumptions (2.61) is a norm equivalent to (2.58).

2.4 Variational Formulation and the Dirichlet-to-Neumann Map

We can now define the variational problems that will be studied at length in this thesis. From here on, we take $\varepsilon_- = \varepsilon_+$ for simplicity. We can include the case where $\varepsilon_- \neq \varepsilon_+$ as noted in remark 2 later in this section at the expense of notational complexity.

In light of the quasi-periodicity of the solutions and the radiation condition discussed earlier in Section 2.1.2, it is appropriate to formulate the variational problem in the domain Ω. On the right and left boundary Γ_R and Γ_L, we enforce quasi-periodic boundary conditions on a solution u. On the top and bottom boundaries Γ_H and Γ_{-H}, we enforce the radiation condition using the Dirichlet-to-Neumann (DtN) operators defined as follows. If $\phi \in H^{1/2}_{qp}(\Gamma_H)$, then

$$\phi(x_1) = \sum_{n\in\mathbb{Z}} \phi_n \exp(i\alpha_n x_1), \tag{2.62}$$

where

$$\alpha_n = \alpha_0 + n\,(2\pi/L)\,. \tag{2.63}$$

In the region Ω^+, let $v_\phi \in H^1_{qp,loc}(\Omega^+)$ satisfy

$$\Delta v_\phi + \kappa^2 \varepsilon_+ v_\phi = 0 \qquad \text{in } \Omega^+, \tag{2.64}$$

$$v_\phi = \phi \qquad \text{on } \Gamma_H, \tag{2.65}$$

together with the upward propagating radiation condition (2.30). Then the Rayleigh expansion

$$v_\phi(x_1, x_2) = \sum_{n\in\mathbb{Z}} \phi_n \exp[i(x_2 - H)\beta_n^+] \exp(i\alpha_n x_1) \tag{2.66}$$

holds. We define the Dirichlet-to-Neumann operators $T^\pm_{s,p} : H^{1/2}_{qp}(\Gamma_{\pm H}) \to H^{-1/2}_{qp}(\Gamma_{\pm H})$ on the top and bottom boundaries, respectively, as

$$(T^\pm_s \phi)(x_1) = \pm \frac{\partial v_\phi}{\partial x_2}\bigg|_{x_2 = \pm H} = i \sum_{n\in\mathbb{Z}} \phi_n \beta_n^+ \exp(i\alpha_n x_1). \tag{2.67}$$

For p-polarization the corresponding DtN operator is denoted $T^\pm_p$ and can be written as

$$(T^\pm_p \phi)(x_1) = \frac{1}{\varepsilon_+}(T^\pm_s \phi)(x_1). \tag{2.68}$$

Remark 1. *Later in this section we prove the claimed mapping property* (2.80)–(2.81). *We will prove later in this thesis that the solutions to the scattering problems are in* $H^{1+\delta}_{qp}(\Omega)$ *for some* $\delta > 0$ *(*(4.5.1.1) *and* (5.5.1.1)*). From the trace theorem [22] we have a constant* $c > 0$ *such that*

$$\|u\|_{H^{1/2+\delta}_{qp}(\partial\Omega)} \leq c\,\|u\|_{H^{1+\delta}_{qp}(\Omega)}\,. \tag{2.69}$$

Therefore, on the top boundary for example,

$$\sum_{n\in\mathbb{Z}} |u_n(H)|^2 |\kappa^2 - \alpha_n^2|^{1/2+\delta} < +\infty. \tag{2.70}$$

From the definition of α_n*, for* $|n|$ *large enough*

$$|\kappa^2 - \alpha_n^2|^{1/2+\delta} \geq \left(\frac{2\pi^2}{L}\right)^{1/2+\delta} |n|^{1+2\delta}. \tag{2.71}$$

Therefore, $|u_n(H)| \leq |n|^{-1-\delta}$ *for* n *large enough, since otherwise we would have a clear contradiction. By the Weierstrass M-test, the Rayleigh expansion* (2.30) *converges uniformly in the half-space* $x_2 > H$*, since* $\sum_{n\in\mathbb{R}} \frac{1}{|n|^{1+\delta}} < +\infty$*. The same argument shows that the Rayleigh expansion* (2.31) *also converges uniformly in the half-space* $x_2 < -H$*.*

In light of this remark, the definition of the DtN map makes sense, considering we have performed term-by-term differentiation in (2.66) to obtain (2.67). Now we can define the variational formulations for the scattering problems, depending on the polarization state. The scattered field u solves the Helmholtz equation

$$\nabla \cdot \Big(A\nabla u\Big) + \kappa^2 a u = f \qquad \text{in } \Omega, \tag{2.72}$$

$$\exp(-i\alpha_0 L)u(0,x_2) = u(L,x_2) \qquad \forall\, x_2, \tag{2.73}$$

$$\exp(-i\alpha_0 L)\frac{\partial}{\partial x_1}u(0,x_2) = \frac{\partial}{\partial x_1}u(L,x_2) \quad \forall\, x_2. \tag{2.74}$$

In the case where Ω is illuminated by an s-polarized plane wave, we have $A = \boldsymbol{I}$, $a = \varepsilon$ and $f = \kappa^2 u^i(\varepsilon_+ - \varepsilon)$. In the case where the incident plane wave is p-polarized, we take $A = \varepsilon^{-1}\boldsymbol{I}$, $a = 1$ and $f = \nabla\cdot\left[(\varepsilon_+^{-1} - \varepsilon^{-1})\nabla u^i\right]$. The multiplicative factors in

(2.73) and (2.74) embody the quasi-periodic boundary conditions. Multiplying (2.72) by a test function $v \in H^1_{qp}(\Omega)$ and using the divergence theorem in the usual way, we have

$$\int_\Omega \left(A\nabla u \nabla \overline{v} - \kappa^2 u \overline{v} \right) - \int_{\Gamma_H} \overline{v} A \nabla u \cdot \nu - \int_{\Gamma_{-H}} \overline{v} A \nabla u \cdot \nu = - \int_\Omega f \overline{v}. \tag{2.75}$$

The boundary integrals on Γ_R and Γ_L cancel because $\overline{v} A \nabla u$ is periodic for all $v \in H^1_{qp}(\Omega)$. By replacing the normal derivatives by the DtN operators, we obtain variational problems. It will be useful in our analysis to consider a general right hand side $F \in (H^1_{qp}(\Omega))'$, so we replace f by F for now. Given an $F \in (H^1_{qp}(\Omega))'$, we seek a $u \in H^1_{qp}(\Omega)$ such that

$$b_\varepsilon(u, v) = - \int_\Omega F \overline{v} \tag{2.76}$$

for all $v \in H^1_{qp}(\Omega)$, where the sesquilinear form $b_\varepsilon(\cdot, \cdot) : H^1_{qp}(\Omega) \times H^1_{qp}(\Omega) \to \mathbb{C}$ is given by

$$b_\varepsilon(u, v) = \int_\Omega \left(\nabla u \cdot \nabla \overline{v} - \kappa^2 \varepsilon u \overline{v} \right) - \int_{\Gamma_H} \overline{v} T_s^+(u) - \int_{\Gamma_{-H}} \overline{v} T_s^-(u). \tag{2.77}$$

Similarly, for the p-polarization state, given an $F \in (H^1_{qp}(\Omega))'$ we seek a $u \in H^1_{qp}(\Omega)$ such that

$$B_\varepsilon(u, v) = - \int_\Omega F \overline{v}, \tag{2.78}$$

where the the sesquilinear form $B_\varepsilon(\cdot, \cdot) : H^1_{qp}(\Omega) \times H^1_{qp}(\Omega) \to \mathbb{C}$ is given by

$$B_\varepsilon(u, v) = \int_\Omega \left(\frac{1}{\varepsilon} \nabla u \cdot \nabla \overline{v} - \kappa^2 u \overline{v} \right) - \int_{\Gamma_H} \overline{v} T_p^+(u) - \int_{\Gamma_{-H}} \overline{v} T_p^-(u). \tag{2.79}$$

We conclude this section with some useful properties of the DtN maps. The four DtN maps given in (2.67) and (2.68) are bounded from $H^{1/2}_{qp}(\Gamma_{\pm H})$ to $H^{-1/2}_{qp}(\Gamma_{\pm H})$. To see this, let $\phi \in H^{1/2}_{qp}(\Gamma_H)$, and then

$$\left\| T_s^+(\phi) \right\|^2_{H^{-1/2}_{qp}(\Gamma_H)} = \sum_{n \in \mathbb{Z}} |\beta_n^+|^2 |\phi_n|^2 |\kappa^2 - \alpha_n^2|^{-1/2} \tag{2.80}$$

$$= \sum_{n \in \mathbb{Z}} |\phi_n|^2 |\kappa^2 - \alpha_n^2|^{1/2} = \|\phi\|^2_{H^{1/2}_{qp}(\Gamma_H)}. \tag{2.81}$$

The same result holds for the three other DtN maps. Furthermore, the DtN boundary integrals found in the sesquilinear forms $b(\cdot, \cdot)$ and $B(\cdot, \cdot)$ are bounded.

Lemma 2.4.1. *$b_\varepsilon(\cdot,\cdot)$ and $B_\varepsilon(\cdot,\cdot)$ are bounded sesquilinear forms on $H^1_{qp}(\Omega)$.*

Proof. To show this, we notice that

$$\begin{aligned}\left|\int_{\Gamma_H}\overline{v}T_s^+(u)\right| &\le L\bigg(\sum_{n\in\mathbb{Z}}|v_n|^2|\kappa^2-\alpha_n^2|^{1/2}\bigg)^{1/2}\bigg(\sum_{n\in\mathbb{Z}}|\beta_n^+|^2|\alpha_n|^2|\kappa^2-\alpha_n^2|^{-1/2}\bigg)^{1/2}\\ &\le L\left\|v\right\|_{H^{1/2}_{qp}(\Gamma_H)}\left\|T_s^+(\phi)\right\|_{H^{-1/2}_{qp}(\Gamma_H)}\\ &\le C\left\|v\right\|_{H^1_{qp}(\Omega)}\left\|u\right\|_{H^1_{qp}(\Omega)},\end{aligned}\tag{2.82}$$

which follows from the trace theorem. The same argument holds for the three other DtN boundary integrals. □

Furthermore, we note that the signs of the real and imaginary parts of the DtN boundary integrals on $\Gamma_{\pm H}$ are known (when setting $v=u$). In particular, the following lemma is used many times throughout this thesis.

Lemma 2.4.2.

$$\left.\begin{aligned}\Re\textstyle\int_{\Gamma_{\pm H}}\overline{u}T_s^\pm(u) &= -L\sum_{\alpha_n^2>\kappa^2\varepsilon_+}\sqrt{\alpha_n^2-\kappa^2\varepsilon_+}\left|u_n^\pm(H)\right|^2\\ \Im\textstyle\int_{\Gamma_{\pm H}}\overline{u}T_s^\pm(u) &= L\sum_{\alpha_n^2<\kappa^2\varepsilon_+}\sqrt{\kappa^2\varepsilon_+-\alpha_n^2}\left|u_n^\pm(H)\right|^2\\ \Re\textstyle\int_{\Gamma_{\pm H}}\varepsilon_+\overline{u}T_p^\pm(u) &= -L\sum_{\alpha_n^2>\kappa^2\varepsilon_+}\sqrt{\alpha_n^2-\kappa^2\varepsilon_+}\left|u_n^\pm(H)\right|^2\\ \Im\textstyle\int_{\Gamma_{\pm H}}\varepsilon_+\overline{u}T_p^\pm(u) &= L\sum_{\alpha_n^2<\kappa^2\varepsilon_+}\sqrt{\kappa^2\varepsilon_+-\alpha_n^2}\left|u_n^\pm(H)\right|^2\end{aligned}\right\}.\tag{2.83}$$

Proof. These follow by applying Parseval's theorem. □

In some parts of our analysis, it is useful to consider the total field $u^t=u+u^i$, instead of the scattered field u. Therefore, we conclude this section by giving the variational problems for the total field, depending on the polarization state. For the s-polarization state, we seek a $u^t\in H^1_{qp}(\Omega)$ such that

$$b_\varepsilon(u^t,v)=\int_{\Gamma_H}\overline{v}\left(\frac{\partial u^i}{\partial x_2}-T_s^+(u^i)\right)\tag{2.84}$$

for all $v \in H^1_{qp}(\Omega)$. For the p-polarization state, we seek a $u^t \in H^1_{qp}(\Omega)$ such that

$$B_\varepsilon(u^t, v) = \int_{\Gamma_H} \overline{v}\left(\frac{1}{\varepsilon_+}\frac{\partial u^i}{\partial x_2} - T_p^+(u^i)\right) \tag{2.85}$$

for all $v \in H^1_{qp}(\Omega)$. We will prove later on in Chapters 4 and 5 that u and u^t are well defined.

Remark 2. *If $\varepsilon_- \neq \varepsilon_+$ then the foregoing variational problems would require modification since $f \neq 0$ in Ω^-. To get around this, we can construct a new problem where the solution w is related to u and u^t, and w solves the same p- or s-polarized problem but only with a modified right hand side. Thus, we could apply our analysis to w and use this to extend our results to the case where $\varepsilon_- \neq \varepsilon_+$. In particular, for some sufficiently small $\delta > 0$, let χ be a C^∞ cut-off function such that $\chi \equiv 1$ for $x_2 > H - \delta$ and $\chi \equiv 0$ for $x_2 < -H + \delta$. On setting $w = u^t - \chi u^i$, we can see that*

$$\Delta w + \kappa^2 \varepsilon w = -(\Delta\chi)u^i - 2\nabla\chi \cdot \nabla u^i + \chi\kappa^2(\varepsilon - \varepsilon_+)u^i, \tag{2.86}$$

in the s-polarization case. In the p-polarization case, we obtain

$$\nabla\cdot\left(\frac{1}{\varepsilon}\nabla w\right) + \kappa^2 w = -(\nabla\chi)\cdot\left(\frac{1}{\varepsilon}\nabla u^i\right) - \nabla\cdot\left(\frac{1}{\varepsilon}u^i\nabla\chi\right) + \chi\left[\nabla\cdot\left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon_+}\right)\nabla u^i\right]. \tag{2.87}$$

In both cases, w satisfies homogeneous boundary conditions on $\Gamma_{\pm H}$.

Chapter 3

THE RIGOROUS COUPLED WAVE APPROACH

3.1 Introduction

We are now ready to describe the RCWA to numerically approximate the solution u to the scattering problems (2.76) and (2.78). The efficient solution scheme used by the RCWA requires a stairstep approximation of the interfaces. To define this stairstep approximation, the domain Ω is decomposed into thin slices stacked along the x_2-axis, using a mesh $-H = h_1 < h_2 < \cdots < h_{S+1} = H$ for some $S > 0$. The slices specified as

$$S_j = \{\boldsymbol{x} \in \mathbb{R}^2,\ 0 < x_1 < L_x, h_j < x_2 < h_{j+1}\}, \quad j \in \{1, \cdots, S\}, \tag{3.1}$$

are stacked along the x_2-axis, where the slice thickness parameter $h = \max_j(h_{j+1} - h_j)$. In the j-th slice S_j, the true relative permittivity ε is sampled on the center line of the slice to yield

$$\varepsilon_h(x_1, x_2) = \varepsilon(x_1, h_{j+\frac{1}{2}}), \quad \boldsymbol{x} \in S_j, \tag{3.2}$$

where $h_{j+\frac{1}{2}} = \frac{1}{2}(h_{j+1} + h_j)$ for each j. Defined piecewise in Ω, ε_h amounts to a stairstep approximation of ε. On each slice ε_h is independent of x_2.

Because of this approximation we are also interested in two additional variational problems, that are a perturbation of the problems given in (2.76) and (2.78). To this end, given a general right hand side $F \in (H^1_{qp}(\Omega))'$, we seek a $u^h \in H^1_{qp}(\Omega)$ such that

$$b_{\varepsilon_h}(u^h, v) = -\int_\Omega F\overline{v} \tag{3.3}$$

for all $v \in H^1_{qp}(\Omega)$. Here, the sesquilinear form $b_{\varepsilon_h}(\cdot,\cdot)$ is the same as $b_\varepsilon(\cdot,\cdot)$, but ε is replaced with ε_h. Similarly, for the p-polarization state, given a general right hand side $F \in (H^1_{qp}(\Omega))'$, we seek a $u^h \in H^1_{qp}(\Omega)$ such that

$$B_{\varepsilon_h}(u^h, v) = -\int_\Omega F\overline{v}. \tag{3.4}$$

Again, the sesquilinear form $B_{\varepsilon_h}(\cdot,\cdot)$ is defined like in (2.79), but with ε replaced by the approximation ε_h. Of course, the actual right hand sides for the scattering problems are $f = \kappa^2 u^i(\varepsilon_+ - \varepsilon_h)$ for the s-polarization state and $f = \nabla\cdot\left[(\varepsilon_+^{-1} - \varepsilon_h^{-1})\nabla u^i\right]$ for the p-polarization state. It is useful in our analysis to consider a general right hand side in the dual space, so for now f is replaced by $F \in (H^1_{qp}(\Omega))'$. It is also useful to consider the variational problems for the total field $u^{h,t} = u^h + u^i$. These variational problems are given as follows. For the s-polarization state, we seek a $u^{h,t} \in H^1_{qp}(\Omega)$ such that

$$b_{\varepsilon_h}(u^{h,t}, v) = \int_{\Gamma_H} \overline{v}\left(\frac{\partial u^i}{\partial x_2} - T_s^+(u^i)\right) \tag{3.5}$$

for all $v \in H^1_{qp}(\Omega)$. For the p-polarization state, we seek a $u^{h,t} \in H^1_{qp}(\Omega)$ such that

$$B_{\varepsilon_h}(u^{h,t}, v) = \int_{\Gamma_H} \overline{v}\left(\frac{1}{\varepsilon_+}\frac{\partial u^i}{\partial x_2} - T_p^+(u^i)\right) \tag{3.6}$$

for all $v \in H^1_{qp}(\Omega)$.

3.2 Coupled Ordinary Differential Equations

In this section, we derive the system of ODESs that the RCWA solves in each slice. We follow [1]. The relative permittivity in the region $-H \le x_2 \le H$ can be expanded as a Fourier series with respect to x_1, as

$$\varepsilon(x_1, x_2) = \sum_{n\in\mathbb{Z}} \varepsilon_n(x_2)\exp\big(\frac{2\pi i n}{L}x_1\big), \tag{3.7}$$

where $k = 2\pi/L$. The field phasors can be written as the Rayleigh expansions

$$\boldsymbol{E}(\boldsymbol{x}) = \sum_{n\in\mathbb{Z}} \boldsymbol{E}_n(x_2)\exp(i\alpha_n x_1), \tag{3.8}$$

$$\boldsymbol{H}(\boldsymbol{x}) = \sum_{n\in\mathbb{Z}} \boldsymbol{H}_n(x_2)\exp(i\alpha_n x_1), \tag{3.9}$$

with the unknown coefficient functions $\boldsymbol{E}_n(x_2) = E_{1,n}(x_2)\boldsymbol{e}_1 + E_{2,n}(x_2)\boldsymbol{e}_2 + E_{3,n}(x_2)\boldsymbol{e}_3$ and $\boldsymbol{H}_n(x_2) = H_{1,n}(x_2)\boldsymbol{e}_1 + H_{2,n}(x_2)\boldsymbol{e}_2 + H_{3,n}(x_2)\boldsymbol{e}_3$. We now substitute these expressions into the Maxwell system (2.11)–(2.16), to obtain

$$\frac{\partial}{\partial x_2} E_{3,n}(x_2) = i\omega\mu_0 H_{1,n}(x_2), \tag{3.10}$$

$$\alpha_n E_{3,n}(x_2) = -\omega\mu_0 H_{2,n}(x_2), \tag{3.11}$$

$$i\alpha_n E_{2,n}(x_2) - \frac{\partial}{\partial x_2} E_{1,n}(x_2) = i\omega\mu_0 H_{3,n}(x_2), \tag{3.12}$$

$$\sum_{m\in\mathbb{Z}} \varepsilon^{-1}_{n-m} \frac{\partial}{\partial x_2} H_{3,m}(x_2) = -i\omega\varepsilon_0 E_{1,n}(x_2,), \tag{3.13}$$

$$\sum_{m\in\mathbb{Z}} \varepsilon^{-1}_{n-m}\alpha_m H_{3,m}(x_2) = \omega\varepsilon_0 E_{2,n}(x_2), \tag{3.14}$$

$$i\alpha_n H_{2,n}(x_2) - \frac{\partial}{\partial x_2} H_{1,n}(x_2) = -i\omega\varepsilon_0 \sum_{m\in\mathbb{Z}} \varepsilon_{n-m} E_{3,m}(x_2), \tag{3.15}$$

for all $x_2 \in (-H, H)$ and $n \in \mathbb{Z}$. The sums in (3.13)–(3.15) represent the discrete convolution. The coefficients ε_{n-m} are the $(m,n)-$th component of the Toeplitz matrix formed from the coefficients of ε. Similarly, the coefficients ε^{-1}_{n-m} are the $(m,n)-$th component of the Toeplitz matrix formed from the coefficients of ε^{-1}. To solve the system (3.10)–(3.15), we restrict the index $|n| \leq M$ for some $M > 0$, and define the $(2M+1)\times 1$ vectors

$$\boldsymbol{X}_\sigma(x_2) = \left[X_{\sigma,-M}, \cdots X_{\sigma,0}, \cdots X_{\sigma,M}\right]^T, \tag{3.16}$$

where $\boldsymbol{X} \in \{\boldsymbol{E}, \boldsymbol{H}\}$, $\sigma \in \{1,2,3\}$ and T denotes the transpose. Similarly, we define the $(2M+1)\times(2M+1)$ diagonal matrix

$$\boldsymbol{\alpha} = \mathrm{diag}(\alpha_n), \tag{3.17}$$

and the $(2M+1)\times(2M+1)$ Toeplitz matrix of a function ϕ is given as

$$\boldsymbol{\mathcal{T}}(\phi) = \begin{bmatrix} \phi_0 & \phi_{-1} & \phi_{-2} & \cdots & \phi_{-M} \\ \phi_1 & \phi_0 & \phi_1 & \cdots & \phi_{-M+1} \\ \phi_2 & \phi_2 & \phi_0 & \cdots & \phi_{-M+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_M & \phi_{M-1} & \phi_{M-2} & \cdots & \phi_0 \end{bmatrix}.$$

Equations (3.11) and (3.14) yield

$$\boldsymbol{H}_2(x_2) = -\frac{1}{\omega\mu_0}\boldsymbol{\alpha}\boldsymbol{E}_3(x_2), \tag{3.18}$$

$$\boldsymbol{E}_2(x_2) = \frac{1}{\omega\varepsilon_0}\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon}\right)(x_2)\boldsymbol{\alpha}\boldsymbol{H}_3(x_2). \tag{3.19}$$

These equations can be used in the remaining four to eliminate the use of the vectors $\boldsymbol{E}_2$ and $\boldsymbol{H}_2$. We obtain the system

$$\frac{\partial}{\partial x_2}\boldsymbol{E}_1(x_2) = i\eta_0\left[\frac{1}{\kappa}\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon}\right)\boldsymbol{\alpha} - \kappa\boldsymbol{I}\right]\boldsymbol{H}_3(x_2), \tag{3.20}$$

$$\frac{\partial}{\partial x_2}\boldsymbol{E}_3(x_2) = i\kappa\eta_0\boldsymbol{H}_1(x_2), \tag{3.21}$$

$$\eta_0\frac{\partial}{\partial x_2}\boldsymbol{H}_1(x_2) = i\left[\kappa\boldsymbol{\mathcal{T}}(\varepsilon) - \frac{1}{\kappa}\boldsymbol{\alpha}^2\right]\boldsymbol{E}_3(x_2), \tag{3.22}$$

$$\eta_0\frac{\partial}{\partial x_2}\boldsymbol{H}_3(x_2) = i\left[-\kappa\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon}\right)^{-1}\right]\boldsymbol{E}_1(x_2). \tag{3.23}$$

Now by defining the $4(2M+1)\times 1$ vector

$$\boldsymbol{f}(x_2) = \left[\boldsymbol{E}_1^T, \boldsymbol{E}_3^T, \eta_0\boldsymbol{H}_1^T, \eta_0\boldsymbol{H}_3^T\right]^T, \tag{3.24}$$

we obtain the vector ODE

$$\frac{\partial}{\partial x_2}\boldsymbol{f}(x_2) = i\boldsymbol{P}(x_2)\boldsymbol{f}(x_2), \tag{3.25}$$

where the $4(2M+1)\times 4(2M+1)$ matrix

$$\boldsymbol{P}(x_2) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{14}(x_2) \\ \boldsymbol{0} & \boldsymbol{0} & \kappa\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{32}(x_2) & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{P}_{41}(x_2) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $\boldsymbol{0}$ is the $(2M+1)\times(2M+1)$ zero matrix and $\boldsymbol{I}$ is the $(2M+1)\times(2M+1)$ identity matrix. The $(2M+1)\times(2M+1)$ sub matrices of the system are defined as

$$\boldsymbol{P}_{14}(x_2) = \frac{1}{\kappa}\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon}\right)\boldsymbol{\alpha} - \kappa\boldsymbol{I}, \tag{3.26}$$

$$\boldsymbol{P}_{32}(x_2) = \kappa\boldsymbol{\mathcal{T}}(\varepsilon) - \frac{1}{\kappa}\boldsymbol{\alpha}^2, \tag{3.27}$$

$$\boldsymbol{P}_{41}(x_2) = -\kappa\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon}\right)^{-1}. \tag{3.28}$$

The vector ODE in (3.25) holds for all $x_2 \in (-H, H)$, and so even this truncated problem is difficult to solve. The approximated permittivity ε_h is chosen so that the matrix $\boldsymbol{P}(x_2)$ is constant in each slice S_j, since ε_h does not depend on x_2 in any slice S_j. We therefore replace the Toeplitz matrices $\boldsymbol{\mathcal{T}}(\varepsilon)$ and $\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon}\right)$ in (3.30)–(3.32) with $\boldsymbol{\mathcal{T}}(\varepsilon_h)$ and $\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon_h}\right)$, respectively. Since the true relative permittivity is sampled along the center line of each slice, we call the resulting matrices

$$\boldsymbol{P}_{j+\frac{1}{2}} = \boldsymbol{P}(h_{j+\frac{1}{2}}) = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{P}_{14}(h_{j+\frac{1}{2}}) \\ \mathbf{0} & \mathbf{0} & \kappa \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}_{32}(h_{j+\frac{1}{2}}) & \mathbf{0} & \mathbf{0} \\ \boldsymbol{P}_{41}(h_{j+\frac{1}{2}}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{3.29}$$

for all $j \in \{1, \cdots, S\}$. Here, the $(2M+1) \times (2M+1)$ sub matrices are

$$\boldsymbol{P}_{14}(h_{j+\frac{1}{2}}) = \frac{1}{\kappa} \boldsymbol{\alpha} \boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon_h}\right) \boldsymbol{\alpha} - \kappa \boldsymbol{I}, \tag{3.30}$$

$$\boldsymbol{P}_{32}(h_{j+\frac{1}{2}}) = \kappa \boldsymbol{\mathcal{T}}(\varepsilon_h) - \frac{1}{\kappa} \boldsymbol{\alpha}^2, \tag{3.31}$$

$$\boldsymbol{P}_{41}(h_{j+\frac{1}{2}}) = -\kappa \boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon_h}\right)^{-1}. \tag{3.32}$$

Therefore, we must solve the matrix ODE

$$\frac{\partial}{\partial x_2} \boldsymbol{f}(x_2) = i \boldsymbol{P}(h_{j+\frac{1}{2}}) \boldsymbol{f}(x_2) \tag{3.33}$$

for each $j \in \{1, \cdots, S\}$ and $x_2 \in S_j$. Once we enforce the appropriate continuity and boundary conditions, the RCWA solution can be computed in Ω as

$$u^{h,M,t}(x_1, x_2) = \sum_{n=-M}^{M} u_n^{h,M,t}(x_2) \exp(i \alpha_n x_1). \tag{3.34}$$

For the s-polarization state we recall that $u_n^{h,M,t} = E_{3,n}(x_2)$, and for the p-polarization state $u_n^{h,M,t}(x_2) = H_{3,n}(x_2)$. In each slice we assume that $\boldsymbol{P}(h_{j+\frac{1}{2}})$ can be diagonalized and written as

$$\boldsymbol{P}(h_{j+\frac{1}{2}}) = \boldsymbol{G}_j \boldsymbol{D}_j \boldsymbol{G}_j^{-1}, \tag{3.35}$$

where $\boldsymbol{G}_j$ is a $4(2M+1) \times 4(2M+1)$ matrix of eigenvectors, and $\boldsymbol{D}_j$ is a $4(2M+1) \times 4(2M+1)$ diagonal matrix of eigenvalues. This assumption, along with (3.33) implies that

$$\boldsymbol{f}(x_2) = \boldsymbol{G}_j \exp(i\boldsymbol{D}_j(x_2 - h_j))\boldsymbol{G}_j^{-1}\boldsymbol{f}(h_j) \tag{3.36}$$

for all $x_2 \in S_j$, $j \in \{1, \cdots, S\}$ and an unknown vector $\boldsymbol{f}(h_j)$. The eigenvalues are arranged in increasing order of the imaginary part, and the eigenvectors are arranged in the same order. This ordering is for the stability of the solution algorithm [62]. To enforce continuity across the interslice boundaries, we set

$$\boldsymbol{f}(h_{j+1}) = \boldsymbol{G}_j \exp(i\boldsymbol{D}_j\Delta_j)\boldsymbol{G}_j^{-1}\boldsymbol{f}(h_j), \tag{3.37}$$

for all $j \in \{1. \cdots, S\}$, where $\Delta_j = h_{j+1} - h_j$.

3.3 Boundary Conditions

As described in Section 2.1.2, we can express the total electric field $u^{h,t} = u^h + u^i$ as a Rayleigh expansion in both half-spaces $x_2 > H$ and $x_2 < -H$. For simplicity, we take $\varepsilon_+ = \varepsilon_- = 1$. To this end, we define the x_3–component of the reflected electric field as

$$u^{\text{ref}}(x_1, x_2) = \sum_{n=-M}^{M} u_n^{\text{ref}} \exp[i\beta_n^+(x_2 - H)] \exp(i\alpha_n x_1), \quad x_2 > H; \tag{3.38}$$

and the transmitted electric field as

$$u^{\text{tr}}(x_1, x_2) = \sum_{n=-M}^{M} u_n^{\text{tr}} \exp[-i\beta_n^+(x_2 + H)] \exp(i\alpha_n x_1), \quad x_2 < -H. \tag{3.39}$$

Similarly, we recall the incident electric field (2.9) and notice the x_3 component can be written as

$$u^i(x_1, x_2) = \exp(-\beta_0^+ H) \exp(-\beta_0^+(x_2 - H)) \exp(i\alpha_0 x_1). \tag{3.40}$$

The $2(2M+1)$ coefficients u_n^{ref} and u_n^i are unknown and must be solved for. To encode these coefficients we define the $(2M+1)\times 1$ vectors

$$\begin{aligned}\boldsymbol{u}^{inc} &= [u^i_{-M},\cdots,u^i_M]^T,\\ \boldsymbol{u}^{ref} &= [u^{ref}_{-M},\cdots,u^{ref}_M]^T,\\ \boldsymbol{u}^{tr} &= [u^{tr}_{-M},\cdots,u^{tr}_M]^T,\end{aligned}$$

along with the $(2M+1)\times(2M+1)$ diagonal matrix $\boldsymbol{\beta}^+$, where the diagonal entries are β_n^+ for all $n=-M,\cdots M$, respectively. For the s-polarization state, the boundary conditions are

$$\boldsymbol{E}_3(H) = \boldsymbol{u}^{inc}+\boldsymbol{u}^{ref}, \tag{3.41}$$

$$\frac{\partial}{\partial x_2}\boldsymbol{E}_3(H) = i\boldsymbol{\beta}^+(-\boldsymbol{u}^{inc}+\boldsymbol{u}^{ref}), \tag{3.42}$$

$$\boldsymbol{E}_3(-H) = \boldsymbol{u}^{tr}, \tag{3.43}$$

$$\frac{\partial}{\partial x_2}\boldsymbol{E}_3(-H) = -i\boldsymbol{\beta}^+\boldsymbol{u}^{tr}, \tag{3.44}$$

on the top and bottom boundaries. In this case, the only non-zero component of $\boldsymbol{u}^{inc}$ is $u_0^{inc}=\exp(-\beta_0^+H)$. We notice that (3.41) is equivalent to

$$\eta_0\boldsymbol{H}_1(H) = \frac{\boldsymbol{\beta}^+}{\kappa}(-\boldsymbol{u}^{inc}+\boldsymbol{u}^{ref}), \tag{3.45}$$

by virtue of (3.21). Also, (3.43) is equivalent to

$$\eta_0\boldsymbol{H}_1(-H) = \frac{-\boldsymbol{\beta}^+}{\kappa}\boldsymbol{u}^{tr}, \tag{3.46}$$

by applying (3.21). The boundary conditions for the p-polarization state are defined similarly. On the top and bottom boundaries, we enforce

$$\eta_0\boldsymbol{H}_3(H) = -(\boldsymbol{u}^{inc}+\boldsymbol{u}^{ref}), \tag{3.47}$$

$$\eta_0\frac{\partial}{\partial x_2}\boldsymbol{H}_3(H) = i\boldsymbol{\beta}^+(\boldsymbol{u}^{inc}-\boldsymbol{u}^{ref}), \tag{3.48}$$

$$\eta_0\boldsymbol{H}_3(-H) = -\boldsymbol{u}^{tr}, \tag{3.49}$$

$$\eta_0\frac{\partial}{\partial x_2}\boldsymbol{H}_3(-H) = i\boldsymbol{\beta}^+\boldsymbol{u}^{tr}. \tag{3.50}$$

The minus signs in the aforementioned boundary conditions follow from (2.13). By virtue of (3.23), we see that (3.48) and (3.50) are equivalent to

$$\boldsymbol{E}_1(H) = \frac{\boldsymbol{\beta}^+}{\kappa}(\boldsymbol{u}^{ref} - \boldsymbol{u}^{inc}), \tag{3.51}$$

$$\boldsymbol{E}_1(-H) = \frac{-\boldsymbol{\beta}^+}{\kappa}\boldsymbol{u}^{tr}, \tag{3.52}$$

respectively.

3.4 The Solution Algorithm

We now describe the RCWA stepping algorithm for the two polarization states [1, 26]. Starting from the bottom slice, the boundary conditions and the continuity of the tangential components over the interslice boundaries are enforced. To enforce the boundary conditions, we define the $2(2M+1)\times 1$ vectors $\boldsymbol{A}$, $\boldsymbol{R}$ and $\boldsymbol{T}$ as

$$\boldsymbol{A} = [a^s_{-M}, \cdots, a^s_M, a^p_{-M}, \cdots, a^p_M]^T, \tag{3.53}$$

$$\boldsymbol{R} = [r^s_{-M}, \cdots, r^s_M, r^p_{-M}, \cdots, r^p_M]^T, \tag{3.54}$$

$$\boldsymbol{T} = [t^s_{-M}, \cdots, t^s_M, t^p_{-M}, \cdots, t^p_M]^T, \tag{3.55}$$

respectively. We also define the $2(2M+1)\times 2(2M+1)$ matrices $\boldsymbol{Y}_e^{\pm}$ and $\boldsymbol{Y}_h^{\pm}$ as

$$\boldsymbol{Y}_e^{\pm} = \begin{bmatrix} \boldsymbol{0} & \mp\frac{\boldsymbol{\beta}^+}{\kappa} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix},$$

$$\boldsymbol{Y}_h^{\pm} = \begin{bmatrix} \mp\frac{\boldsymbol{\beta}^+}{\kappa} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I} \end{bmatrix},$$

respectively. Now, the boundary conditions are given in the matrix form

$$\boldsymbol{f}(H) = \begin{bmatrix} \boldsymbol{Y}_e^+ & \boldsymbol{Y}_e^- \\ \boldsymbol{Y}_h^+ & \boldsymbol{Y}_h^- \end{bmatrix} \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{R} \end{bmatrix},$$

$$\boldsymbol{f}(-H) = \begin{bmatrix} \boldsymbol{Y}_e^+ \\ \boldsymbol{Y}_h^+ \end{bmatrix} \boldsymbol{T}.$$

The s-polarization state now corresponds to $\boldsymbol{A} = [(\boldsymbol{u}^{inc})^T \mid \boldsymbol{0}^T]^T$, $\boldsymbol{R} = [(\boldsymbol{u}^{ref})^T \mid \boldsymbol{0}^T]^T$ and $\boldsymbol{T} = [(\boldsymbol{u}^{tr})^T \mid \boldsymbol{0}^T]^T$. Similarly, the p-polarization state corresponds to $\boldsymbol{A} =$

$[\mathbf{0}^T \mid (\boldsymbol{u}^{inc})^T]^T$, $\boldsymbol{R} = [\mathbf{0}^T \mid (\boldsymbol{u}^{ref})^T]^T$ and $\boldsymbol{T} = [\mathbf{0}^T \mid (\boldsymbol{u}^{tr})^T]^T$. Here, $\mathbf{0}$ is the $(2M+1)\times 1$ zero vector. To solve for the $\boldsymbol{f}(h_j)$ vectors, we assume that

$$\boldsymbol{f}(h_j) = \boldsymbol{Z}_j \boldsymbol{T}_j, \tag{3.56}$$

for all $j \in \{1, \cdots, S+1\}$, where

$$\boldsymbol{T}_1 = \boldsymbol{T}, \tag{3.57}$$

$$\boldsymbol{Z}_1 = \begin{bmatrix} \boldsymbol{Y}_e^+ \\ \boldsymbol{Y}_h^+ \end{bmatrix}, \tag{3.58}$$

in order to enforce the boundary conditions on Γ_{-H}. Using (3.56) in (3.37), we have

$$\boldsymbol{Z}_{j+1}\boldsymbol{T}_{j+1} = \boldsymbol{G}_j \begin{bmatrix} e^{i\Delta_j \boldsymbol{D}_j^u} & \mathbf{0} \\ \mathbf{0} & e^{i\Delta_j \boldsymbol{D}_j^l} \end{bmatrix} \boldsymbol{G}_j^{-1} \boldsymbol{Z}_j \boldsymbol{T}_j \tag{3.59}$$

for all $j \in \{1, \cdots, S\}$, where $\boldsymbol{D}_j^u$ and $\boldsymbol{D}_j^l$ are the upper and lower diagonal submatrices of $\boldsymbol{D}_j$. Continuity of $\boldsymbol{f}(x_2)$ across the interslice boundaries is enforced by (3.59). The recursion relation

$$\boldsymbol{T}_{j+1} = \exp(i\Delta_j \boldsymbol{D}_j^u)\boldsymbol{W}_j^u \boldsymbol{T}_j \tag{3.60}$$

is enforced where

$$\begin{bmatrix} \boldsymbol{W}_j^u \\ \boldsymbol{W}_j^l \end{bmatrix} = \boldsymbol{G}_j^{-1} \boldsymbol{Z}_j \tag{3.61}$$

for all $j \in \{1, \cdots, S\}$. Using this recursion relation in (3.59) yields

$$\boldsymbol{Z}_{j+1} = \boldsymbol{G}_j \begin{bmatrix} \boldsymbol{I} \\ e^{i\Delta_j \boldsymbol{D}_j^l} \boldsymbol{W}_j^l (\boldsymbol{W}_j^u)^{-1} e^{-i\Delta_j \boldsymbol{D}_j^u} \end{bmatrix}. \tag{3.62}$$

Starting from the bottom where $j = 1$, the matrix $\boldsymbol{P}_{j+\frac{1}{2}}$ is computed by sampling the ε_h at $x_2 = \frac{h_{j+1}+h_j}{2}$ and using the Fast Fourier Transform (FFT) in MATLAB® to compute the Toeplitz matrices. With $\boldsymbol{Z}_1$ known via (3.58), $\boldsymbol{Z}_j$ for all $j \in \{2, \cdots, S\}$ is solved using (3.61) and (3.62). Simultaneously $\boldsymbol{W}_j^u$ for $j \in \{1, \cdots, S\}$ is found using relation (3.61). Now using (3.56) again, we have in particular

$$\boldsymbol{f}(h_{S+1}) = \boldsymbol{Z}_{S+1}\boldsymbol{T}_{S+1} = \begin{bmatrix} \boldsymbol{Y}_e^+ & \boldsymbol{Y}_e^- \\ \boldsymbol{Y}_h^+ & \boldsymbol{Y}_h^- \end{bmatrix} \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{R} \end{bmatrix}, \tag{3.63}$$

by the boundary condition on Γ_H. The system above is equivalent to

$$\begin{bmatrix} \boldsymbol{T}_{S+1} \\ \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}^u_{S+1} & -\boldsymbol{Y}_e^- \\ \boldsymbol{Z}^l_{S+1} & -\boldsymbol{Y}_h^- \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{Y}_e^+ \boldsymbol{A} \\ \boldsymbol{Y}_h^+ \boldsymbol{A} \end{bmatrix}. \tag{3.64}$$

Therefore, $\boldsymbol{T}_{S+1}$ can be found and $\boldsymbol{T}_j$ for all $j \in \{1, \cdots, S\}$ can be solved by reversing the sense of the recursion relation (3.60).

Finally, $\boldsymbol{f}(h_j)$ can be found for all $j \in \{1, \cdots, S\}$ by using the relation (3.56). The RCWA solution can now be reconstructed in Ω from the Fourier coefficients in the slices, as

$$u^{h,M,t}(x_1, x_2) = \sum_{n=-M}^{M} u_n^{h,M,t}(x_2) \exp(i\alpha_n x_1). \tag{3.65}$$

For the s-polarization case, $u_n^{h,M,t}(x_2) = E_{3,n}(h_j)$ for $x_2 \in (h_j, h_{j+1})$. For the p-polarization case $u_n^{h,M,t}(x_2) = H_{3,n}(h_j)$ for $x_2 \in (h_j, h_{j+1})$.

3.5 The RCWA as a Galerkin Scheme

In this section, we show that the RCWA is a Galerkin scheme, i.e., solves an appropriate variational problem. We recall that the Fourier coefficients of the tangential components are continuous. Furthermore, in light of (3.21) and (3.23), we have

$$\boldsymbol{f}(h_j^-) = \boldsymbol{f}(h_j^+), \tag{3.66}$$

$$\frac{\partial}{\partial x_2} \boldsymbol{E}_3(h_j^-) = \frac{\partial}{\partial x_2} \boldsymbol{E}_3(h_j^+), \tag{3.67}$$

$$\mathcal{T}\left(\frac{1}{\varepsilon_h}\right) \frac{\partial}{\partial x_2} \boldsymbol{H}_3(h_j^-) = \mathcal{T}\left(\frac{1}{\varepsilon_h}\right) \frac{\partial}{\partial x_2} \boldsymbol{H}_3(h_j^+), \tag{3.68}$$

for all $j \in \{2, \cdots, S\}$. From the system of ODEs described by (3.33) and (3.29), we find that

$$\frac{\partial^2}{\partial x_2^2} \boldsymbol{E}_3(x_2) = \left[\boldsymbol{\alpha}^2 - \kappa^2 \boldsymbol{\mathcal{T}}(\varepsilon_h)\right] \boldsymbol{E}_3(x_2), \tag{3.69}$$

in slice S_j. This holds because $\boldsymbol{\mathcal{T}}(\varepsilon_h)$ does not depend on x_2 in a slice. Thus, for the s-polarization state, each Fourier coefficient function $u_n^{h,M,t}(x_2)$ satisfies the second-order ODE

$$\frac{\partial^2}{\partial x_2^2} u_n^{h,M,t}(x_2) = \alpha_n^2 u_n^{h,M,t}(x_2) - \kappa^2 \sum_{m=-M}^{M} \varepsilon_{h,n-m} u_m^{h,M,t}(x_2), \tag{3.70}$$

for all $n \in \{-M, \cdots, M\}$ in each slice S_j.

We now prove the following theorem:

Theorem 3.5.1. *For the s-polarization state, the RCWA solution $u^{h,M,t}(x_1, x_2)$ solves the variational problem*

$$b_{\varepsilon_h}(u^{h,M,t}, v_M) = \int_{\Gamma_H} \overline{v_M}\left(\frac{\partial u^i}{\partial x_2} - T_s^+(u^i)\right) \tag{3.71}$$

for all $v_M \in V_M$, where $u^i = E_3^{inc}$.

Proof. Let $v_M \in V_M$, then $v_M = \sum_{n=-M}^{M} \xi_n \phi_n$ where $\xi_n \in H^1(-H, H)$ and $\phi_n \in S_M$ for all $n = -M, \cdots, M$. We multiply (3.70) by $\overline{\xi}_n$ and integrate in x_2 in each slice S_j. We obtain

$$\begin{aligned}\int_{h_j}^{h_{j+1}} \frac{\partial}{\partial x_2} u_n^{h,M,t}(x_2) \frac{\partial}{\partial x_2}\overline{\xi}_n(x_2)\, dx_2 - \overline{\xi}_n(h_{j+1}^-)\frac{\partial}{\partial x_2}u_n^{h,M,t}(h_{j+1}^-) + \overline{\xi}_n(h_j^+)\frac{\partial}{\partial x_2}u_n^{h,M,t}(h_j^+) \\ = \int_{h_j}^{h_{j+1}} \left[\kappa^2 \sum_{m=-M}^{M} \varepsilon_{h,n-m} u_m^{h,M,t}(x_2) - \alpha_n^2 u_n^{h,M,t}(x_2)\right]\overline{\xi}_n\, dx_2,\end{aligned} \tag{3.72}$$

for all $j \in \{1, \cdots, S\}$ after integrating by parts. We sum (3.72) over all j and use the continuity (3.67) to obtain

$$\begin{aligned}\int_{-H}^{H} \frac{\partial}{\partial x_2} u_n^{h,M,t}(x_2) \frac{\partial}{\partial x_2}\overline{\xi}_n(x_2)\, dx_2 - \overline{\xi}_n(H^-)\frac{\partial}{\partial x_2}u_n^{h,M,t}(H^-) + \overline{\xi}_n(-H^+)\frac{\partial}{\partial x_2}u_n^{h,M,t}(-H^+) \\ = \int_{-H}^{H} \left[\kappa^2 \sum_{m=-M}^{M} \varepsilon_{h,n-m}(x_2) u_m^{h,M,t}(x_2) - \alpha_n^2 u_n^{h,M,t}(x_2)\right]\overline{\xi}_n\, dx_2,\end{aligned} \tag{3.73}$$

Now we multiply both sides of (3.73) by $\phi_n\overline{\phi}_n$ and integrate in x_1 in $[0, L]$. Then summing the resulting equation over all $n \in \{-M, \cdots, M\}$ and using the orthogonality of the basis functions, we have

$$\int_\Omega \left(\nabla u^{h,M,t} \cdot \nabla \overline{v}_M - \kappa^2 \varepsilon_h u^{h,M,t}\overline{v}_M\right) - \int_{\Gamma_H} \overline{v}_M \frac{\partial}{\partial x_2} u^{h,M,t} + \int_{\Gamma_{-H}} \overline{v}_M \frac{\partial}{\partial x_2} u^{h,M,t} = 0. \tag{3.74}$$

After applying the boundary conditions (3.41)–(3.44), we have

$$\frac{\partial}{\partial x_2} u^{h,M,t}(H) = T_s^+(u^{h,M,t}) + \frac{\partial u^i}{\partial x_2} - T_s^+(u^i), \tag{3.75}$$

$$\frac{\partial}{\partial x_2} u^{h,M,t}(-H) = -T_s^-(u^{h,M,t}). \tag{3.76}$$

Substitution of these equations in (3.74) completes the proof. □

Similarly, from the system of ODEs described by (3.33) and (3.29), we also find

$$\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon_h}\right)\frac{\partial^2}{\partial x_2^2}H_3(x_2) = \left[\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\left(\frac{1}{\varepsilon_h}\right)\boldsymbol{\alpha} - \kappa^2\boldsymbol{I}\right]H_3(x_2), \tag{3.77}$$

in each S_j. Thus for the p-polarization state, each Fourier coefficient $u_n^{h,M,t}$ solves the second order ODE

$$\sum_{m=-M}^{M}\varepsilon_{h,n-m}^{-1}\frac{\partial^2}{\partial x_2^2}u_m^{h,M,t} = \alpha_n\sum_{m=-M}^{M}\varepsilon_{h,n-m}^{-1}u_m^{h,M,t}\alpha_m - \kappa^2 u_n^{h,M,t}, \tag{3.78}$$

for all $n \in \{-M, \cdots, M\}$ in each S_j.

We can now prove the following theorem:

Theorem 3.5.2. *For the p-polarization state, the RCWA solution $u^{h,M,t}(x_1, x_2)$ solves the variational problem*

$$B_{\varepsilon_h}(u^{h,M,t}, v_M) = \int_{\Gamma_H}\overline{v_M}\left(\frac{1}{\varepsilon_+}\frac{\partial u^i}{\partial x_2} - T_p^+(u^i)\right), \tag{3.79}$$

for all $v_M \in V_M$, where $u^i = -\eta_0^{-1}\boldsymbol{E}_3^{inc}\cdot\boldsymbol{e}_3$.

Proof. Let $v_M \in V_M$, then $v_M = \sum_{n=-M}^{M}\xi_n\phi_n$ where $\xi_n \in H^1(-H, H)$ and $\phi_n \in S_M$ for all $n \in \{-M, \cdots, M\}$. We multiply (3.78) by $\overline{\xi}_n$ and integrate over x_2 in each slice S_j. We obtain

$$\begin{aligned}
&\int_{h_j}^{h_{j+1}}\left(\sum_{m=-M}^{M}\varepsilon_{h,n-m}^{-1}\frac{\partial}{\partial x_2}u_m^{h,M,t}(x_2)\right)\frac{\partial}{\partial x_2}\overline{\xi}_n(x_2)\,dx_2 \\
&\quad - \overline{\xi}_n(h_{j+1}^-)\left(\sum_{m=-M}^{M}\varepsilon_{h,n-m}^{-1}\frac{\partial}{\partial x_2}u_m^{h,M,t}(h_{j+1}^-)\right) \\
&\quad + \overline{\xi}_n(h_j^+)\left(\sum_{m=-M}^{M}\varepsilon_{h,n-m}^{-1}\frac{\partial}{\partial x_2}u_m^{h,M,t}(h_j^+)\right) \\
&\quad = \int_{h_j}^{h_{j+1}}\left[\kappa^2 u_n^{h,M,t}(x_2) - \alpha_n\sum_{n=-M}^{M}\varepsilon_{h,n-m}^{-1}u_m^{h,M,t}(x_2)\alpha_m\right]\overline{\xi}_n(x_2)\,dx_2
\end{aligned} \tag{3.80}$$

for all $j \in \{1, \cdots, S\}$. We sum (3.80) over all $j \in \{1, \cdots, S\}$ and use the continuity (3.68) to obtain

$$\begin{aligned}
&\int_{-H}^{H} \left(\sum_{m=-M}^{M} \varepsilon_{h,n-m}^{-1}(x_2) \frac{\partial}{\partial x_2} u_m^{h,M,t}(x_2) \right) \frac{\partial}{\partial x_2} \overline{\xi}_n(x_2)\, dx_2 \\
&\quad - \overline{\xi}_n(H^-) \frac{1}{\varepsilon_+} \frac{\partial}{\partial x_2} u_n^{h,M,t}(H^-) \\
&\quad + \overline{\xi}_n(-H^+) \frac{1}{\varepsilon_+} \frac{\partial}{\partial x_2} u_n^{h,M,t}(-H^+) \\
&\quad = \int_{-H}^{H} \left[\kappa^2 u_n^{h,M,t}(x_2) - \alpha_n \sum_{n=-M}^{M} \varepsilon_{h,n-m}^{-1} u_m^{h,M,t}(x_2) \alpha_m \right] \overline{\xi}_n(x_2)\, dx_2.
\end{aligned} \tag{3.81}$$

Now we multiply both sides of (3.80) by $\phi_n \overline{\phi}_n$ and integrate over x_1 in $[0, L]$. We then sum the resulting equation over all $n \in \{-M, \cdots, M\}$ and use the orthogonality of the basis functions. We obtain

$$\int_\Omega \left(\frac{1}{\varepsilon_h} \nabla u^{h,M,t} \cdot \nabla \overline{v}_M - \kappa^2 u^{h,M,t} \overline{v}_M \right) - \int_{\Gamma_H} \overline{v}_M \frac{1}{\varepsilon_+} \frac{\partial}{\partial x_2} u^{h,M,t} + \int_{\Gamma_{-H}} \overline{v}_M \frac{1}{\varepsilon_+} \frac{\partial}{\partial x_2} u^{h,M,t} = 0. \tag{3.82}$$

By virtue of the boundary conditions (3.47)–(3.50), we have

$$\begin{aligned}
\frac{1}{\varepsilon_+} \frac{\partial}{\partial x_2} u^{h,M,t}(H) &= T_p^+(u^{h,M,t}) + \frac{1}{\varepsilon_+} \frac{\partial u^i}{\partial x_2} - T_p^+(u^i), \\
\frac{1}{\varepsilon_+} \frac{\partial}{\partial x_2} u^{h,M,t}(-H) &= -T_p^-(u^{h,M,t}).
\end{aligned}$$

Using these equations in (3.82) completes the proof. □

3.6 The Stairstep Approximation of Interfaces

In this section, we investigate the approximation of the relative permittivity ε using the stairstep approximation ε_h. We assume that any interface is approximated at most $\mathcal{P} \geq 2$ times in each slice S_j, for some fixed $\mathcal{P}$ independent of the slice. First, we assume that ε is piecewise constant and all interfaces are piecewise linear so that the support

$$\operatorname{supp}(\varepsilon - \varepsilon_h) = \bigcup_{\tau \in T_h} \tau, \tag{3.83}$$

where T_h is a set that contains at most $2\mathcal{P}S$ triangles. An illustration of the support $\operatorname{supp}(\varepsilon - \varepsilon_h)$ is given in Figure 3.1. Setting the area of a $\tau \in T_h$ as $\mathcal{A}(\tau) = \frac{1}{2}\mathcal{B}(\tau)\mathcal{H}(\tau)$, we notice that

$$\sum_{\tau \in T_h} \mathcal{B}(\tau) \leq L \tag{3.84}$$

for all $h > 0$. Furthermore, by definition of h it follows that $\mathcal{H}(\tau) \leq h/2$ for every $h > 0$ and $\tau \in T_h$. Then the area

$$\begin{aligned} \Big|\operatorname{supp}(\varepsilon - \varepsilon_h)\Big| &= \frac{1}{2} \sum_{\tau \in T_h} \mathcal{B}(\tau)\mathcal{H}(\tau) && (3.85) \\ &\leq \frac{h}{4} \sum_{\tau \in T_h} \mathcal{B}(\tau) && (3.86) \\ &\leq \frac{hL}{4}. && (3.87) \end{aligned}$$

Theorem 3.6.1. *Suppose that ε is piecewise constant, and there are $I \geq 1$ interfaces that are graphs of piecewise linear functions. Then there is a constant $C > 0$ independent of $h > 0$ such that*

$$\|\varepsilon - \varepsilon_h\|_{L^p(\Omega)} \leq C h^{1/p}. \tag{3.88}$$

Proof. From the previous discussion, we have

$$\begin{aligned} \|\varepsilon - \varepsilon_h\|_{L^p(\Omega)} &= \left(\int_{\operatorname{supp}(\varepsilon - \varepsilon_h)} |\varepsilon - \varepsilon_h|^p \right)^{1/p} \\ &\leq \Big(\sup_{\boldsymbol{x} \in \Omega} (|\varepsilon - \varepsilon_h|) \Big| \operatorname{supp}(\varepsilon - \varepsilon_h) \Big| \Big)^{1/p}, \end{aligned}$$

which completes the proof. □

Theorem 3.6.2. *Suppose that ε is piecewise constant, and there are $I \geq 1$ interfaces that are the graphs of C^2 functions. For $h > 0$ small enough, there is a constant $C > 0$ independent of $h > 0$ such that*

$$\|\varepsilon - \varepsilon_h\|_{L^p(\Omega)} \leq C h^{1/p}. \tag{3.89}$$

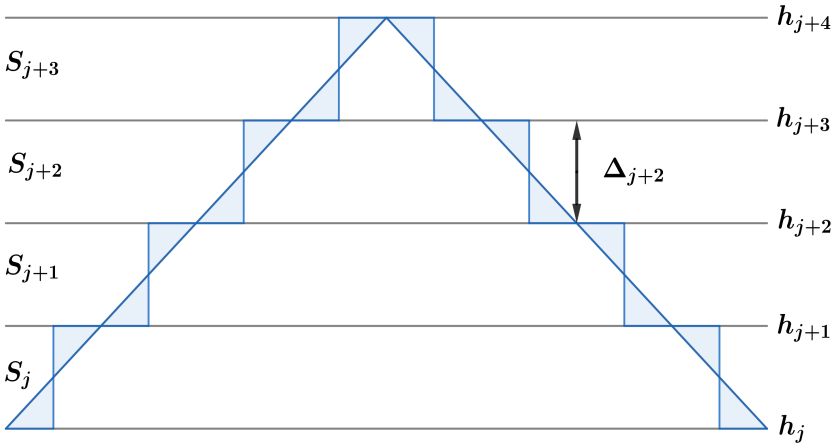


Figure 3.1: Illustration of the stairstep approximation of a piecewise linear interface, where ε is piecewise constant. The shaded triangular regions denote where ε differs from ε_h.

Proof. We approximate an interface by a piecewise linear approximation in the following way. Assume the interface is the graph of the C^2 function g. At the interslice boundaries and at the center line of each slice, we build a piecewise linear interpolation g^* of g. We let ε^* be the relative permittivity associated with the interface g^*. In the grating region we can write

$$\varepsilon = \begin{cases} \varepsilon_1 & x \geq g(x_1), \\ \varepsilon_2 & x_2 < g(x_1), \end{cases} \tag{3.90}$$

where ε_1 and ε_2 are constant. Therefore, the relative permittivity associated to g^* is

$$\varepsilon^* = \begin{cases} \varepsilon_1 & x \geq g^*(x_1), \\ \varepsilon_2 & x_2 < g^*(x_1). \end{cases} \tag{3.91}$$

By the previous result, we see that

$$\|\varepsilon - \varepsilon_h\|_{L^p(\Omega)} \leq Ch^{1/p} + \|\varepsilon - \varepsilon^*\|_{L^p(\Omega)}. \tag{3.92}$$

Standard approximation theory yields a constant $C > 0$ independent of $h > 0$ such that $|g - g^*| \leq Ch^2$. Since the function g is rectifiable, the arc length $\mathcal{A}(g)$ is well-defined and

$$\left|\text{supp}(\varepsilon - \varepsilon^*)\right| \leq \mathcal{A}(g)Ch^2, \tag{3.93}$$

where the measure is bounded by an approximating rectangle with base length $\mathcal{A}(g)$. It now follows from (3.92) that

$$\|\varepsilon - \varepsilon_h\|_{L^p(\Omega)} \leq C(h^{1/p} + h^{2/p}), \tag{3.94}$$

which completes the proof. □

Theorem 3.6.3. *Suppose that ε is piecewise C^2 and the interfaces are the graphs of piecewise C^2 functions. Then there is a constant $C > 0$ independent of $h > 0$ such that*

$$\|\varepsilon - \varepsilon_h\|_{L^p(\Omega)} \leq Ch^{1/p}, \tag{3.95}$$

for all $1 \leq p < \infty$ and h small enough.

Proof. Since we assume that any interface is approximated at most $\mathcal{P}$ times, we can separate each slice into a finite number of regions where ε is C^2 and a finite number of regions where ε has jumps. In any region where ε is smooth, it holds that

$$\frac{\varepsilon(x_1, x_2) - \varepsilon(x_1, h_{j-\frac{1}{2}})}{x_2 - h_{j-\frac{1}{2}}} \leq \|\varepsilon\|_{W^{1,\infty}(\Omega)} \tag{3.96}$$

by the mean value theorem. By the definition of the approximation ε_h, we see that

$$\varepsilon - \varepsilon_h \leq \|\varepsilon\|_{W^{1,\infty}(\Omega)}\, h. \tag{3.97}$$

Therefore, by applying the result for the regions where the ε has jumps, we get

$$\|\varepsilon - \varepsilon_h\|_{L^p(\Omega)} \leq C(h^{1/p} + h). \tag{3.98}$$

□

Remark 3. *If $\varepsilon \in C^2(\mathbb{R}^2)$, then for some constant $C > 0$ indepdendent of h, it holds that $\|\varepsilon - \varepsilon_h\|_{L^q(\Omega)} \leq Ch$.*

Chapter 4

ANALYSIS OF RCWA FOR S-POLARIZED LIGHT

4.1 Introduction

This chapter provides an error analysis of the two-dimensional RCWA for s-polarized incident light. The contribution of this chapter is that we use the fact that the RCWA is a Galerkin scheme to analyze its convergence properties. We generalize a Rellich identity and an *a-priori* estimate for two relevant continuous problems, and use them to show the existence and uniqueness of the relevant solutions. In particular, we apply these continuous results to the semi-discrete problem obtained by replacing ε by ε_h. We then show that under certain non-trapping conditions, the continuity constant for the *a-priori* estimate does not depend on slice thickness.

This chapter is organized as follows. In Section 4.2, we state the appropriate mathematical problem: an inhomogeneous Helmholtz equation with quasi-periodic boundary conditions, and give the variational formulation of our problem. In Section 4.3, we derive a Rellich identity for the Helmholtz equation and in Section 4.4, assuming the real part of the relative permittivity is positive, we give an *a-priori* estimate where the continuity constant is explicit. This explicit dependence is needed both for our analysis of stairstepping, as well as in a duality argument appearing in the analysis of convergence in the number of the retained Fourier modes. This restricts us to considering non-trapping domains, as discussed later in Section 4.4. The case where there is light trapping is not covered by our theory, although convergence is seen in practice [11, 12]. In Section 4.5 we show a similar *a-priori* estimate holds when the real part of the relative permittivity is negative. We then apply tools applicable to the Finite Element Method (FEM) in order to show that the RCWA converges with

respect to the number of retained Fourier modes in Section 4.5.1 and also with respect to the stairstep approximation of the grating interfaces in Section 4.5.2. These are the main results of the chapter. Finally, in Section 4.6, we compare the RCWA solution to a refined FEM solution to test our prediction of the order of convergence.

The content of this chapter has been published as a paper [36].

4.2 The Continuous Problem

An s-polarized plane wave propagating in the half-space $x_2 > H$ at an angle θ with respect to the x_2-axis is incident on the plane $x_2 = H$; the sole non-zero component of its electric field phasor is given by the third component of (2.9).

The scattered electric field phasor $u\boldsymbol{e}_3$ is given in terms of the total field by $u = u^t - u^{\text{inc}}$, where u satisfies the Helmholtz equation (2.72)–(2.74) with $A = \boldsymbol{I}$, $a = \varepsilon$ and $f = \kappa^2 u^i(\varepsilon_+ - \varepsilon)$.

We assume that ε is piecewise $C^{1,1}$ in $\mathbb{R}^2$ and that either

I. ε is real and $\Re(\varepsilon) > 0$, or

II. ε is complex, $\Im(\varepsilon) > c_1 > 0$, and $\Re(\varepsilon) > c_2 > 0$ in Ω and a positive real constant elsewhere, or

III. ε is complex, $\Im(\varepsilon) > c_1 > 0$, and $\Re(\varepsilon) \leq 0$ in Ω and a positive real constant elsewhere.

Case I encompasses insulators whereas Case II covers dissipative dielectric materials (but not metals), and Case III covers metals.

To show convergence of the RCWA in Case I, we prove a Rellich identity for the problem and show that an *a-priori* estimate holds when ε is piecewise $C^{1,1}$. We assume certain non-trapping conditions in order to ensure that the continuity constants in the *a-priori* estimates can be written explicitly in terms of κ and ε. Finally, we show that Case II and Case III follow from Case I.

4.2.1 Variational formulation

To prove convergence of the RCWA, we need to consider several different variational problems, because the approach replaces the true ε with an approximation ε_h. The approximation ε_h was given in (3.2).

As we showed in Chapter 2.4, after multiplying (2.72) by a test function and using the divergence theorem in the usual way, the resulting sesquilinear form $b_\varepsilon(\cdot,\cdot) : H^1_{qp}(\Omega) \times H^1_{qp}(\Omega) \to \mathbb{C}$ is

$$b_\varepsilon(u,v) = \int_\Omega \left(\nabla u \cdot \nabla \overline{v} - \kappa^2 \varepsilon u \overline{v} \right) - \int_{\Gamma_H} \overline{v} T_s^+(u) - \int_{\Gamma_{-H}} \overline{v} T_s^-(u). \tag{4.1}$$

for all $w \in H^1_{qp}(\Omega)$ and $v \in H^1_{qp}(\Omega)$. Given an $f \in L^2(\Omega)$, we seek a solution $u \in H^1_{qp}(\Omega)$ such that

$$b_\epsilon(u,v) = f(v) \tag{4.2}$$

for all $v \in H^1_{qp}(\Omega)$, where $f(v) = -\int_\Omega f\overline{v}$.

We are also interested in a perturbed problem in which ε is replaced by ε_h. Therefore, we define $b_{\varepsilon_h}(\cdot,\cdot)$ to be the same as $b_\varepsilon(\cdot,\cdot)$ but with ε_h instead of ε. Given an $f \in L^2(\Omega)$, we seek a solution $u^h \in H^1_{qp}(\Omega)$ such that

$$b_{\epsilon_h}(u^h,v) = f(v) \tag{4.3}$$

for all $v \in H^1_{qp}(\Omega)$.

To show that both of the foregoing problems have unique solutions, we show that either a Rellich identity holds for our problem and implies an *a-priori* estimate, or the problem is coercive depending on our assumptions on ε.

In some arguments throughout this chapter, it is useful to use the total field $u^t = u + u^i$, instead of the scattered field u. Therefore, we conclude this section by giving the variational problem for the total field. We seek a $u^t \in H^1_{qp}(\Omega)$ such that

$$b_\varepsilon(u^t,v) = \int_{\Gamma_H} \overline{v}\left(\frac{\partial u^i}{\partial x_2} - T_s^+(u^i) \right) \tag{4.4}$$

for all $v \in H^1_{qp}(\Omega)$.

4.3 A Rellich Identity

The main tool to prove convergence of the RCWA for the s-polarization state is a Rellich identity for the scattering problems (4.2) and (4.3). Later in this section, we use this identity to show convergence in the number of retained Fourier modes and also the slice thickness.

We now show that a Rellich identity for an unbounded layered-media problem [16] also holds in our quasi-periodic case. Following Lechleiter and Ritterbusch [16], we have the following lemma.

Lemma 4.3.1. *Assume that $\varepsilon \in C^{1,1}(\overline{\Omega_k})$ for all $k = 1, 2, \cdots, I+1$ is real in Ω and $\Re(\varepsilon) > 0$. If u is a solution to the variational problem (4.2) for $f \in L^2(\Omega)$, then the following Rellich identity holds:*

$$\begin{aligned}
&\int_\Omega \left[2\left|\frac{\partial u}{\partial x_2}\right|^2 + \kappa^2 (x_2 + H)\frac{\partial \varepsilon}{\partial x_2}|u|^2 \right] - \sum_{k=1}^{I} \kappa^2 \int_{\Gamma_k} (x_2 + H)\, [\![\varepsilon]\!]_{\Gamma_k} |u|^2 \nu_2 \\
&+ 2H \int_{\Gamma_H} \left(|\nabla u|^2 - 2\left|\frac{\partial u}{\partial x_2}\right|^2 - \kappa^2 \varepsilon |u|^2 \right) - \int_{\Gamma_H} \overline{u} T_s^+(u) - \int_{\Gamma_{-H}} \overline{u} T_s^-(u) \\
&= -2 \int_\Omega (x_2 + H) \Re(\overline{f} \frac{\partial u}{\partial x_2}) - \int_\Omega f \overline{u}.
\end{aligned}$$

Remark 4. *Here, ν_2 is the second component of the normal vector $\boldsymbol{\nu}$. For a stairstepped interface, the vertical sections do not appear in the sum in the first line of the Rellich identity, since $\nu_2 = 0$ there. Since the horizontal sections of a stairstep interface constitute a piecewise Lipschitz-continuous function at all but a finite number of x_1, we can control the L^2 norm of u.*

Proof. As in Lemma 3.1 (a) Ref. [16] , elliptic regularity implies that a solution $u \in H^1(\Omega)$ of (4.4) also belongs to $H^2(\Omega)$. Our proof follows [16], where we check that the same Rellich identity holds for quasi-periodic solutions. Choosing the test function $v = (x_2 + h)\frac{\partial u}{\partial x_2}$, we have

$$\begin{aligned}
\int_\Omega (x_2 + H)\frac{\partial u}{\partial x_2}\Delta \overline{u} &= -\int_\Omega \nabla\left[(x_2 + H)\frac{\partial u}{\partial x_2}\right] \cdot \nabla \overline{u} + \int_{\partial\Omega} (x_2 + H)\frac{\partial u}{\partial x_2}\frac{\partial \overline{u}}{\partial \boldsymbol{\nu}} \\
&= -\int_\Omega \left|\frac{\partial u}{\partial x_2}\right|^2 + (x_2 + H)\nabla\left(\frac{\partial u}{\partial x_2}\right) \cdot \nabla \overline{u} + 2H \int_{\Gamma_H} \left|\frac{\partial u}{\partial x_2}\right|^2.
\end{aligned}$$

Here, we used Green's first identity in the first step, and quasi-periodicity to cancel the left and right boundary integrals, since

$$\frac{\partial u_R}{\partial x_2}\nabla\overline{u}_R\cdot\nu_R=-\frac{\partial u_L}{\partial x_2}\nabla\overline{u}_L\cdot\nu_L. \tag{4.5}$$

By taking twice the real part of both sides, and using the identity

$$\frac{\partial}{\partial x_2}|\nabla u|^2=2\Re\left[\nabla u\cdot\nabla\left(\frac{\partial\overline{u}}{\partial x_2}\right)\right], \tag{4.6}$$

we obtain that

$$\begin{aligned}2\Re\int_\Omega(x_2+H)\frac{\partial u}{\partial x_2}\Delta\overline{u}&=-\int_\Omega\left[2\left|\frac{\partial u}{\partial x_2}\right|^2+(x_2+H)\frac{\partial}{\partial x_2}|\nabla u|^2\right]+2H\int_{\Gamma_H}2\left|\frac{\partial u}{\partial x_2}\right|^2.\\&=\int_\Omega\left(|\nabla u|^2-2\left|\frac{\partial u}{\partial x_2}\right|^2\right)+2H\int_{\Gamma_H}\left(-|\nabla u|^2+2\left|\frac{\partial u}{\partial x_2}\right|^2\right),\end{aligned} \tag{4.7}$$

where we used the Divergence theorem in the second step, that $x_2=-H$ on Γ_{-H}, and $\nu_2=0$ on Γ_L and Γ_R.

On the other hand, we have from (2.72) that $\Delta\overline{u}=\overline{f}-\kappa^2\varepsilon\overline{u}$ for ε real in Ω. Then,

$$\begin{aligned}&2\Re\int_\Omega(x_2+H)\frac{\partial u}{\partial x_2}\Delta\overline{u}\\&=2\int_\Omega(x_2+H)\Re\left(\frac{\partial u}{\partial x_2}\overline{f}\right)-\kappa^2\int_\Omega(x_2+H)\varepsilon2\Re\left(\frac{\partial u}{\partial x_2}\overline{u}\right)\\&=2\int_\Omega(x_2+H)\Re\left(\frac{\partial u}{\partial x_2}\overline{f}\right)-2H\int_{\Gamma_H}\kappa^2\varepsilon|u|^2+\kappa^2\int_\Omega\frac{\partial}{\partial x_2}\left[(x_2+H)\varepsilon\right]|u|^2\\&-\sum_{k=1}^{I}\kappa^2\int_{\Gamma_k}(x_2+H)\left[\!\left[\varepsilon\right]\!\right]_{\Gamma_k}|u|^2\nu_2.\end{aligned} \tag{4.8}$$

This follows from integrating by parts in the second step, from the identity

$$2\Re\left(\frac{\partial u}{\partial x_2}\overline{u}\right)=\frac{\partial}{\partial x_2}|u|^2, \tag{4.9}$$

and by the quasi-periodicity of u. The Rellich identity follows from (4.7) and (4.8).

□

4.4 *A-priori* Estimate

Using the Rellich identity along the lines of [16], we can prove an *a-priori* estimate for the solution with a continuity constant with explicit dependence on ε and h. This can be used to prove existence for all κ under the assumptions on ε given in the statement of the theorem in this section. The *a-priori* estimate holds for all such ε as described in Section 4.2 Case I, and so it holds for the stairstep approximation ε_h. We rely on the non-trapping conditions to ensure that the continuity constant is bounded independent of h. Assuming that $\nu_2 < 0$ and $[\![\varepsilon]\!]_{\Gamma_k} > 0$, we have the following lemma.

Lemma 4.4.1. *For all solutions $u \in H^1(\Omega)$ to the variational problem (4.2), there is a constant $C > 0$ such that*

$$\|u\|^2_{L^2(\Omega)} \leq C\left(2\left\|\frac{\partial u}{\partial x_2}\right\|^2_{L^2(\Omega)} - \kappa^2 \sum_{k=1}^{I} \int_{\Gamma_k} (x_2 + H)\,[\![\varepsilon]\!]_{\Gamma_k}\,|u|^2\nu_2\right), \tag{4.10}$$

where the constant

$$C = 2H\left(H + \frac{2}{\kappa^2 \min_{\hat{\Gamma}_k} |\nu_2| \min_k \inf_{\hat{\Gamma}_k} \left((x_2 + H)\,[\![\varepsilon]\!]_{\hat{\Gamma}_k}\right)}\right). \tag{4.11}$$

Proof. By the definition of the g_k, we can define the subsets of Ω by

$$V_{lk} = \{\boldsymbol{x} \in \Omega, x_{lk} \leq x_1 \leq x_{(l+1)k}, \min_{x_{lk} \leq x_1 \leq x_{(l+1)k}} g_k - \delta \leq x_2 \leq \min_{x_{lk} \leq x_1 \leq x_{(l+1)k}} g_{k+1} - \delta\}, \tag{4.12}$$

for all $k = 2, \cdots, I-1$ and all l. The upper bound on x_2 should be replaced with H when $k = I$, and similarly the lower bound on x_2 should be $-H$ when $k = 1$. By construction, we have

$$\Omega = \bigcup_{lk} V_{lk}. \tag{4.13}$$

Since each g_k is Lipschitz-continuous in V_{kl}, we apply [16] Lemma 4.3 to each V_{lk}, so that

$$\|u\|^2_{L^2(V_{lk})} \leq 4H\,\|u\|^2_{L^2(\hat{\Gamma}_{lk})} + 4H^2\left\|\frac{\partial u}{\partial x_2}\right\|^2_{L^2(V_{lk})}. \tag{4.14}$$

Now we sum over all k and j, and use that $\nu_2 \neq 0$ on any $\hat{\Gamma}_k$,

$$\begin{aligned}\|u\|^2_{L^2(\Omega)} &\leq \frac{4H}{\min_{\hat{\Gamma}_k}|\nu_2|}\sum_{k=1}^{I}\left(\left\||\nu_2|^{1/2}u\right\|^2_{L^2(\hat{\Gamma}_k)}\right) + 4H^2\left\|\frac{\partial u}{\partial x_2}\right\|^2_{L^2(\Omega)} \\ &\leq \frac{4H}{\kappa^2\min_{\hat{\Gamma}_k}|\nu_2|\min_k\inf_{\hat{\Gamma}_k}\left((x_2+H)[\![\varepsilon]\!]_{\hat{\Gamma}_k}\right)}\kappa^2\sum_{k=1}^{I}\int_{\Gamma_k}(x_2+H)[\![\varepsilon]\!]_{\Gamma_k}|u|^2|\nu_2| \\ &\quad + 4H^2\left\|\frac{\partial u}{\partial x_2}\right\|^2_{L^2(\Omega)}, \end{aligned} \tag{4.15}$$

where in the last line we used that $\nu_2 = 0$ on the vertical sections of the Γ_k. To complete the proof, by construction we have $-\nu_2 = |\nu_2|$ on Γ_k. □

Under the assumption $\Re(\varepsilon) > 0$ and $\Im(\varepsilon) = 0$, we prove the following theorem.

Theorem 4.4.2. *Assume that $\varepsilon \in C^{(1,1)}(\overline{\Omega_k})$ for all $k = 1, 2, \cdots, I+1$, and the non-trapping conditions*

$$\frac{\partial\varepsilon}{\partial x_2} \geq 0 \text{ in } \Omega_k, \quad [\![\varepsilon]\!]_{\Gamma_k} > 0, \text{ and } \Re(\varepsilon^+ - \varepsilon) \geq 0 \text{ on } \Gamma_H, \tag{4.16}$$

hold for all $k = 1, 2, \cdots, I+1$. Further, assume that ε is real in Ω. Then for $f \in L^2(\Omega)$ there exists a unique solution $u \in H^1(\Omega)$ of the variational problem (4.2). Also there is an explicit constant

$$C(\kappa,\varepsilon) = C(1+\kappa^2)\|\varepsilon\|_{L^\infty(\Omega)}(4\kappa H\sqrt{\varepsilon_+} + 4H + 1) + 1$$

with C defined as in (4.11), such that

$$\|u\|_{H^1(\Omega)} \leq C(\kappa,\varepsilon)\|f\|_{L^2(\Omega)}.$$

Proof. The proof follows the same procedure as in [16], but we use different Dirichlet-to-Neumann operators. For all $a \geq H$, we use the representation (2.30) to compute the coefficients

$$u_n(a) = \exp\left[i(a-H)\sqrt{\kappa^2\varepsilon_+ - \alpha_n^2}\right]u_n(H), \tag{4.17}$$

$$(\partial_2 u)_n(a) = i\sqrt{\kappa^2\varepsilon_+ - \alpha_n^2}\exp\left[i(a-H)\sqrt{\kappa^2\varepsilon_+ - \alpha_n^2}\right]u_n(H), \tag{4.18}$$

$$(\partial_1 u)_n(a) = i\alpha_n\exp\left[i(a-H)\sqrt{\kappa^2\varepsilon_+ - \alpha_n^2}\right]u_n(H). \tag{4.19}$$

Furthermore, using (4.17)–(4.19) we can bound the boundary integral on the second line of the Rellich identity,

$$\begin{aligned}
&\int_{\Gamma_H}\left(-|\nabla u|^2+2\left|\frac{\partial u}{\partial x_2}\right|^2+\kappa^2\varepsilon|u|^2\right)\\
&\quad=\sum_{n\in\mathbb{Z}}\left(\left|\kappa^2\varepsilon_+-\alpha_n^2\right|-\alpha_n^2+\kappa^2\varepsilon_+\right)\left|\exp\left[2i(a-H)\sqrt{\kappa^2\varepsilon_+-\alpha_n^2}\right]\right||u(H)|^2\\
&\quad=2\sum_{\alpha_n^2<\kappa^2\varepsilon_+}\left(\kappa^2-\alpha_n^2\right)|u_n(H)|^2\\
&\quad\le 2k\sqrt{\varepsilon_+}\Im\int_{\Gamma_H}\overline{u}T_s^+(u).
\end{aligned}\tag{4.20}$$

Now using the test function $v=u$ in the variational problem (4.2), and taking the imaginary part, we have

$$\begin{aligned}
\Im\int_{\Gamma_H}\overline{u}T_s^+(u)&=\Im\int_\Omega f\overline{u}-\Im\int_{\Gamma_{-H}}\overline{u}T_s^-(u)\\
&\le\Im\int_\Omega f\overline{u}.
\end{aligned}\tag{4.21}$$

From the non-trapping assumptions (4.16) for ε and using the estimates derived above, we get

$$\begin{aligned}
&\int_\Omega 2\left|\frac{\partial u}{\partial x_2}\right|^2-\sum_{k=1}^I\kappa^2\int_{\Gamma_k}(x_2+H)\left[\!\left[\varepsilon\right]\!\right]_{\Gamma_k}|u|^2\nu_2\\
&\le 4kH\sqrt{\varepsilon_+}\Im\int_{\Gamma_H}f\overline{u}-2\int_\Omega(x_2+H)\Re\big(\overline{f}\frac{\partial u}{\partial x_2}\big)-\Re\int_\Omega f\overline{u}.
\end{aligned}\tag{4.22}$$

Now we combine (4.22) and lemma 4.4.1 to obtain

$$\begin{aligned}
\|u\|_{L^2(\Omega)}^2&\le C\left[2\left\|\frac{\partial u}{\partial x_2}\right\|_{L^2(\Omega)}^2-\kappa^2\sum_k^I\int_{\Gamma_k}(x_2+H)\left[\!\left[\varepsilon\right]\!\right]_{\Gamma_k}\nu_2\right]\\
&\le C\left[4kH\sqrt{\varepsilon_+}\Im\int_{\Gamma_H}f\overline{u}-2\int_\Omega(x_2+H)\Re(\overline{f}\frac{\partial u}{\partial x_2})-\Re\int_\Omega f\overline{u}\right]\\
&\le C\left[(4kH\sqrt{\varepsilon_+}+4H+1)\,\|f\|_{L^2(\Omega)}\,\|u\|_{H^1(\Omega)}\right].
\end{aligned}\tag{4.23}$$

We note that for ε with $\Re(\varepsilon_\pm)>0$ and $\Im(\varepsilon_\pm)\ge 0$ the term $4k\sqrt{\varepsilon^+}H$ in the above L^2 estimate can be replaced by $2\rho\kappa H$, where ρ is defined as in [16] Lemma 4.2. Taking $u=v$ in the variational problem (4.2) and taking the real part, we have

$$\|u\|_{H^1(\Omega)}^2\le(1+\kappa^2)\,\|\varepsilon\|_{L^\infty(\Omega)}\,\|u\|_{L^2(\Omega)}^2+\|f\|_{L^2(\Omega)}\,\|u\|_{L^2(\Omega)}\,.$$

Consequently, for all $\kappa \geq \kappa_0 > 0$, we have an explicit constant $C(\kappa_0, \varepsilon) > 0$ such that $\|u\|_{H^1(\Omega)} \leq C(\kappa_0, \varepsilon)(1+\kappa^3)\|f\|_{L^2(\Omega)}$. Therefore we obtain existence, uniqueness and boundedness of the solution u to (4.2) and the solution u^h to (4.3). This follows because the *a-priori* estimate implies an inf-sup condition for $b_\varepsilon(u,v)$ and $b_{\varepsilon_h}(u,v)$ [15]. To show this, we define the coercive sesquilinear form $b_\varepsilon^+(\cdot,\cdot) : H^1_{qp}(\Omega) \times H^1_{qp}(\Omega) \to \mathbb{C}$ as

$$b_\varepsilon^+(w,v) = \int_\Omega \left(\nabla w \cdot \nabla \overline{v} + \kappa^2 w \overline{v}\right) - \int_{\Gamma_H} \overline{v} T_s^+(w) - \int_{\Gamma_{-H}} \overline{v} T_s^-(w). \tag{4.24}$$

It is clear by the definition that

$$b_\varepsilon^+(w,w) \geq \min(1,\kappa^2)\|w\|^2_{H^1(\Omega)} \tag{4.25}$$

for all $w \in H^1_{qp}(\Omega)$. Let $0 \neq u \in H^1_{qp}(\Omega)$. We can therefore choose an $F \in (H^1_{qp}(\Omega))'$ such that $F(v) = b_\varepsilon(u,v)$ for all $v \in H^1_{qp}(\Omega)$. By virtue of the Lax-Milgram Lemma, we let $u^+ \in H^1_{qp}(\Omega)$ be the solution to

$$b_\varepsilon^+(u^+,v) = \int_\Omega F\overline{v} \tag{4.26}$$

for all $v \in H^1_{qp}(\Omega)$. Furthermore, the *a-priori* estimate

$$\left\|u^+\right\|_{H^1(\Omega)} \leq \min(1,\kappa^2)^{-1}\|F\|_{(H^1_{qp}(\Omega))'} \tag{4.27}$$

holds for this problem. By construction, we notice that

$$b_\varepsilon(u-u^+,v) = \int_\Omega \kappa^2(1+\varepsilon)u^+\overline{v}$$

for all $v \in H^1_{qp}(\Omega)$. As $\kappa^2(1+\varepsilon)u^+ \in L^2(\Omega)$, we apply our *a-priori* estimate for non-trapping domains to obtain

$$\|u\|_{H^1(\Omega)} \leq \left(\kappa^2 C(\kappa,\varepsilon)\|1+\varepsilon\|_{L^\infty(\Omega)} + 1\right)\min(1,\kappa^2)^{-1}\|F\|_{(H^1_{qp}(\Omega))'}, \tag{4.28}$$

which follows by applying the triangle inequality. Using the definition of the dual norm, we see that

$$\sup_{0\neq v\in H^1_{qp}(\Omega)} \frac{\left|b_\varepsilon(u,v)\right|}{\|v\|_{H^1(\Omega)}} \geq \gamma^{-1}\|u\|_{H^1(\Omega)} \tag{4.29}$$

for all $0 \neq u, v \in H^1_{qp}(\Omega)$. And we obtain an inf-sup constant

$$\beta = \inf_{0\neq u\in H^1_{qp}(\Omega)} \sup_{0\neq v\in H^1_{qp}(\Omega)} \frac{\left|b_\varepsilon(u,v)\right|}{\|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}} > 0. \tag{4.30}$$

To show the transposed inf-sup condition holds, we let $\phi, \psi \in H^{1/2}(\Gamma_H)$, and notice

$$\int_{\Gamma_H} \phi T_s^+(\psi) = \int_{\Gamma_H} \psi T_s^+(\phi), \tag{4.31}$$

which follows immediately from Parseval's theorem, and the definition of the DtN operator T_s^+. The same identity holds for the DtN integral on the bottom boundary. It now follows that $b_\varepsilon(v,w) = b_\varepsilon(\overline{w}, \overline{v})$ for all $v, w \in H^1_{qp}(\Omega)$. Finally, we obtain the transposed inf-sup condition

$$\begin{aligned} \sup_{0\neq u\in H^1_{qp}(\Omega)} \frac{\left|b_\varepsilon(u,v)\right|}{\|u\|_{H^1(\Omega)}} &= \sup_{0\neq u\in H^1_{qp}(\Omega)} \frac{\left|b_\varepsilon(\overline{v},u)\right|}{\|u\|_{H^1(\Omega)}} \\ &\geq \beta \|v\|_{H^1(\Omega)} \end{aligned} \tag{4.32}$$

for all $0 \neq v \in H^1_{qp}(\Omega)$. □

Remark 5. *The non-trapping conditions (4.16) can be altered so that the signs of the conditions are all reversed. Under those assumptions, along with $\Re(\varepsilon_- - \varepsilon) \geq 0$ on Γ_{-H}, the same a-priori estimate holds.*

Corollary 4.4.2.1. *Assume $\Re(\varepsilon) > 0$ satisfies the non-trapping conditions (4.16) and $\Im(\varepsilon) \in L^\infty(\Omega)$ with $\Im(\varepsilon) > 0$. Then the solution $u \in H^1_{qp}(\Omega)$ to the variational problem (4.2) exists and is unique; furthermore there is a constant $C_1(\kappa, \varepsilon) > 0$ such that*

$$\|u\|_{H^1(\Omega)} \leq C_1(\kappa,\varepsilon) \|f\|_{L^2(\Omega)},$$

with the explicit constant defined as

$$C_1(\kappa,\varepsilon) = C(\kappa,\varepsilon)\left[2 + \kappa^2 \|\Im(\varepsilon)\|_{L^\infty(\Omega)} C(\kappa,\varepsilon)\right]. \tag{4.33}$$

Proof. We rewrite the Helmholtz equation (2.72) as

$$\Delta u + \kappa^2 \Re(\varepsilon) u = f - \kappa^2 i \Im(\varepsilon) u \tag{4.34}$$

in Ω. Since the $\Re(\varepsilon)$ satisfies the non-trapping conditions, we apply the *a-priori* estimate 4.4.2 to obtain

$$\|u\|_{H^1(\Omega)} \leq C(\kappa,\varepsilon)\left[\|f\|_{L^2(\Omega)} + \left\|k^2\Im(\varepsilon)u\right\|_{L^2(\Omega)}\right]. \tag{4.35}$$

By setting $v = u$ in the variational problem (4.2) and then taking the real part, we see that

$$\begin{aligned}\left\|\kappa^2\Im(\varepsilon)u\right\|^2_{L^2(\Omega)} &\leq \kappa^2 \|\Im(\varepsilon)\|_{L^\infty(\Omega)} \Im\int_\Omega f\overline{u} \\ &\leq \kappa^2 \|\Im(\varepsilon)\|_{L^\infty(\Omega)} \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)}.\end{aligned} \tag{4.36}$$

Thus, for all $\delta > 0$, it follows that

$$\left\|\kappa^2\Im(\varepsilon)u\right\|_{L^2(\Omega)} \leq \kappa \|\Im(\varepsilon)\|^{1/2}_{L^\infty(\Omega)} \left(\frac{\delta\|f\|_{L^2(\Omega)}}{2} + \frac{\|u\|_{L^2(\Omega)}}{2\delta}\right), \tag{4.37}$$

by applying the arithmetic geometric mean inequality with (4.36). Now, on setting $\delta = \kappa \|\Im(\varepsilon)\|^{1/2}_{L^\infty(\Omega)} C(\kappa,\varepsilon)$ we have

$$\left\|\kappa^2\Im(\varepsilon)u\right\|_{L^2(\Omega)} \leq \frac{\kappa^2\|\Im(\varepsilon)\|_{L^\infty(\Omega)} C(\kappa,\varepsilon)}{2}\|f\|_{L^2(\Omega)} + \frac{\|u\|_{L^2(\Omega)}}{2C(\kappa,\varepsilon)}. \tag{4.38}$$

The corollary follows from using (4.38) in (4.35). □

In the previous corollary we provided an *a-priori* estimate for the case where $\Re(\varepsilon) > 0$ and $\Im(\varepsilon) > 0$. Now we prove an *a-priori* estimate for the case where $\Re(\varepsilon) \leq 0$ and $\Im(\varepsilon) > c_1 > 0$. This case is necessary to allow, for example, metallic gratings.

Corollary 4.4.2.2. *Assume $\Re(\varepsilon) \leq 0$ and $\Im(\varepsilon) \geq c_1 > 0$, and both satisfy the non-trapping conditions (4.16). Then the solution $u \in H^1_{qp}(\Omega)$ to the variational problem (4.2) exists and is unique; furthermore, there is a constant $C_2(\kappa,\varepsilon) > 0$ such that*

$$\|u\|_{H^1(\Omega)} \leq C_2(\kappa,\varepsilon)\|f\|_{L^2(\Omega)},$$

with the explicit constant defined as

$$C_2(\kappa,\varepsilon) = C_1(\kappa,\varepsilon)\left[2 + \frac{\|\Re(\varepsilon)\|_{L^\infty(\Omega)} + 1}{c_1}\kappa^2\|\Im(\varepsilon)\|_{L^\infty(\Omega)} C_1(\kappa,\varepsilon)\right]. \tag{4.39}$$

Proof. We rewrite the Helmholtz equation (2.19) as

$$\Delta u + \kappa^2 \big(\tilde{\varepsilon}\big) u = f + \frac{\|\Re(\varepsilon)\|_{L^\infty(\Omega)} + 1}{c_1} \kappa^2 u \Im(\varepsilon) \tag{4.40}$$

in Ω, where $\tilde{\varepsilon} = \frac{\|\Re(\varepsilon)\|_{L^\infty(\Omega)}+1}{c_1}\Im(\varepsilon) + \Re(\varepsilon) + i\Im(\varepsilon)$. We notice that $\Re(\tilde{\varepsilon}) > 0$ and satisfies the non-trapping conditions (4.16). Therefore, by the *a-priori* estimate 4.4.2.1, we have

$$\|u\|_{H^1(\Omega)} \leq C_1(\kappa, \varepsilon) \left[\|f\|_{L^2(\Omega)} + \frac{\|\Re(\varepsilon)\|_{L^\infty(\Omega)} + 1}{c_1} \|\kappa \Im(\varepsilon) u\|_{L^2(\Omega)} \right]. \tag{4.41}$$

We apply the same technique as in 4.4.2.1, by choosing $\delta = \kappa \|\Im(\varepsilon)\|^{1/2}_{L^\infty(\Omega)} C_1(\kappa, \varepsilon)$ to obtain (4.38) with $C_1(\kappa, \varepsilon)$ instead of $C(\kappa, \varepsilon)$. The obtained estimate can then be used in (4.41). □

Lemma 4.4.3. *Suppose ε satisfies the non-trapping conditions (4.16). Then ε_h also satisfies them, and*

$$C(\kappa, \varepsilon_h) \leq C(\kappa, \varepsilon) \tag{4.42}$$

for all $h > 0$.

Proof. In the definition of the constant $C(\kappa, \varepsilon)$, the $\hat{\Gamma}_k$ only include the piecewise C^2 sections of the interfaces, and do not include any vertical sections. Thus, for the ε_h problem, $\hat{\Gamma}_{h,k}$ only constitute flat sections. Therefore, $\min_{\hat{\Gamma}_{h,k}} |\gamma_2| = 1$ for any $h > 0$. Furthermore,

$$[\![\varepsilon_h]\!]_{\hat{\Gamma}_{h,k}} = \varepsilon(x_1, h_{j+\frac{3}{2}}) - \varepsilon(x_1, h_{j-\frac{1}{2}}) > [\![\varepsilon]\!]_{\hat{\Gamma}_k}. \tag{4.43}$$

Since we have assumed that ε satisfies the non-trapping conditions, it holds that

$$\inf_{\hat{\Gamma}_{h,k}} \left((x_2 + H) [\![\varepsilon_h]\!]_{\hat{\Gamma}_{h,k}} \right)^{-1} \leq \inf_{\hat{\Gamma}_k} \left((x_2 + H) [\![\varepsilon]\!]_{\hat{\Gamma}_k} \right)^{-1}. \tag{4.44}$$

Finally, in each slice ε_h does not depend on x_2, so that $\frac{\partial \varepsilon_h}{\partial x_2} = 0$ in each S_j. This shows that ε_h satisfies the non-trapping conditions. □

Remark 6. *(1) Since the constants $C_1(\kappa, \varepsilon)$ and $C_2(\kappa, \varepsilon)$ are defined in terms of $C(\kappa, \varepsilon)$, it also holds that $C_1(\kappa, \varepsilon_h) \leq C_1(\kappa, \varepsilon)$ and $C_2(\kappa, \varepsilon_h) \leq C_2(\kappa, \varepsilon)$.*

(2) If the non-trapping conditions are not satisfied, we cannot assert that the $C(\kappa, \epsilon_h)$ is bounded independent of h. Indeed, if $\Im(\epsilon) = 0$, it may be that κ^2 is an exceptional frequency for the ϵ_h problem. Then $C(\kappa, \epsilon_h)$ would not be bounded. Even if $\Im(\epsilon) > 0$, it may be that $C(\kappa, \epsilon_h)$ depends poorly on h. In most problems this will not be the case, so we expect RCWA to converge even for trapping domains.

4.5 An Adjoint Problem

We now study an adjoint problem, related to (4.2). Given an $f \in L^2_{qp}(\Omega)$, let $z_f \in H^1_{qp}(\Omega)$ be the unique solution to the adjoint problem

$$\overline{b_\varepsilon(\xi, z_f)} = -\int_\Omega f\overline{\xi} \tag{4.45}$$

for all $\xi \in H^1_{qp}(\Omega)$. The solution z_f exists and is unique because it solves the same problem as (4.2) with $\overline{f}$ on the right hand side, and the same *a-priori* estimates hold. We extend the domain Ω by ℓ periods on the left and right, and then above and below by including the infinite half-spaces where $x_2 > H$ and $x_2 < -H$. This extended domain is then defined as

$$\Omega^E = \{\boldsymbol{x} \in \mathbb{R}^2,\ -\ell L_x < x_1 < (\ell+1)L\}. \tag{4.46}$$

As it is also useful to define a circular restricted domain, we choose an $R > 0$ such that

$$\Omega_R = \{\boldsymbol{x} \in \mathbb{R}^2, \left|\boldsymbol{x} - (L/2, 0)\right| < R\} \tag{4.47}$$

satisfies the set inclusion $\Omega \subset \Omega_R \subset \Omega^E$. The right hand side f is extended to Ω_R by quasi-periodicity in x_1 and by zero above and below. We can also extend the solution z_f to the domain Ω^E by quasi-periodicity to the left and right in x_1, and using the Rayleigh expansions (2.30) and (2.31) above and below, to obtain $z_f^E \in H^1_{qp}(\Omega^E)$.

Let χ be a smooth cut-off function such that $\chi = 1$ in Ω, $\chi = 0$ on $\partial\Omega_R$. We consider $w = \chi z_f^E$, and notice immediately that $w = z_f$ in Ω and $w = 0$ on $\partial\Omega_R$. Then $w \in H^1(\Omega_R)$ solves the Poisson problem

$$\left.\begin{aligned} \Delta w &= (\kappa^2 \varepsilon z_f^E - \overline{f^E})\chi + 2\nabla z_f^E \cdot \nabla\chi + z_f^E \Delta\chi && \text{in } \Omega_R \\ w &= 0 && \text{on } \partial\Omega_R \end{aligned}\right\}. \tag{4.48}$$

To show this is true, let $v \in H^1_{qp}(\Omega_R)$ and consider

$$\int_{\Omega_R} \Delta w \overline{v} = \int_{\Omega_R} \Delta\big(\chi z_f^E\big)\overline{v} = \int_{\Omega_R} \big(\Delta(z_f^E)\chi + 2\nabla\chi \cdot \nabla z_f^E + z_f^E \Delta\chi\big)\overline{v}. \tag{4.49}$$

We recall that by construction, the extended solution is the variational solution to $\Delta z_f^E = f^E - \kappa^2 \varepsilon z_f^E$ in Ω_R, and therefore by Green's first identity

$$\int_{\Omega_R} \Delta\big(z_f^E\big)\chi\overline{v} = \int_{\Omega_R} \left(\kappa^2 \varepsilon z_f^E - \overline{f^E}\right)\overline{v}, \tag{4.50}$$

where we used that $\chi = 0$ on $\partial\Omega_R$. The use of (4.50) in (4.49) shows that w solves the Poisson problem (4.48). We define the sets

$$\left.\begin{aligned} \Omega_E^+ &:= \{-\ell L < x_1 < (\ell+1)L, H \le x_2 < R\} \\ \Omega_E^- &:= \{-\ell L < x_1 < (\ell+1)L, -R < x_2 < -H\} \end{aligned}\right\}, \tag{4.51}$$

and prove the following theorem.

Theorem 4.5.1. *Let $z_f \in H^1_{qp}(\Omega)$ be the solution to the adjoint problem (4.45). Then there exists a constant $c > 0$ independent of ε and κ such that*

$$\big\|z_f^E\big\|_{H^1(\Omega_E^\pm)} \le c \, \|z_f\|_{H^1(\Omega)} \, . \tag{4.52}$$

Proof. The Rayleigh expansion

$$z_f(\boldsymbol{x}) = \sum_{n\in\mathbb{Z}} (z_f)_n(H) \exp(i(x_2 - H)\beta_n^+) \exp(i\alpha_n x_1) \tag{4.53}$$

is valid in Ω_E^+. After using Parseval's Theorem, it follows that

$$\big\|z_f^E\big\|^2_{H^1(\Omega_E^+)} \le RL(2\ell+1)(1+\kappa\varepsilon_+) \big\|z_f^E\big\|^2_{H^{1/2}(\Gamma_H)} \, .$$

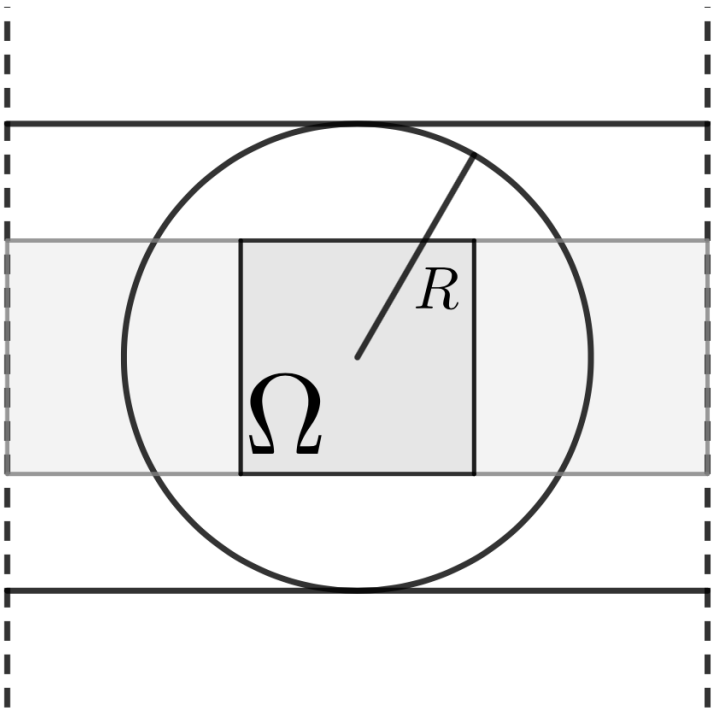


Figure 4.1: An illustration of the extended domain Ω^E, with $\ell = 1$.

By the trace theorem [22] there is a constant $c > 0$ such that

$$\left\| z_f^E \right\|_{H^1(\Omega_E^+)} \leq c \left\| z_f \right\|_{H^1(\Omega)} . \tag{4.54}$$

The same inequality holds for the H^1 norm in Ω_E^- using the Rayleigh expansion (2.31) below Ω. This completes the proof. □

Corollary 4.5.1.1. *Suppose ε satisfies the conditions of Theorem 4.4.2 or any of its corollaries. Given $f \in L^2(\Omega)$, let $z_f \in H^2(\Omega)$ be the unique solution to the adjoint problem (4.45). Then there exists a constant $C_3(\kappa, \varepsilon) > 0$ such that*

$$\left\| z_f \right\|_{H^2(\Omega)} \leq C_3(\kappa, \varepsilon) \left\| f \right\|_{L^2(\Omega)} . \tag{4.55}$$

Proof. Since the extended solution z_f^E solves the Poisson problem (4.48), by Gilbarg and Trudinger [17] there is a constant $C > 0$ independent of κ and ε such that

$$\begin{aligned} \left\| w \right\|_{H^2(\Omega_R)} &\leq C \left\| (\kappa^2 \varepsilon z_f^E - \overline{f^E})\chi + 2\nabla z_f^E \cdot \nabla \chi + z_f^E \Delta \chi \right\|_{L^2(\Omega_R)} \\ &\leq C(\chi) \left[\kappa^2 \left\| \varepsilon \right\|_{L^\infty(\Omega)} + 3 \right] \left[2\ell + 1 + 2c \right] \left\| z_f \right\|_{H^1(\Omega)} . \end{aligned}$$

We complete the proof by using the *a-priori* estimate for z_f and recalling that

$$\|z_f\|_{H^2(\Omega)} \leq \|w\|_{H^2(\Omega_R)}. \tag{4.56}$$

□

Remark 7. *Based on our assumptions on ε, the constant $C_3(\kappa, \varepsilon)$ will depend on $C_1(\kappa, \varepsilon)$ or $C_2(\kappa, \varepsilon)$. It also follows that, in any case, $C_3(\kappa, \varepsilon_h) \leq C_3(\kappa, \varepsilon)$. So in particular, the solution $u^h \in H^2(\Omega)$ to the variational problem (4.45) satisfies an a-priori estimate where the continuity constant is independent of h.*

4.5.1 Convergence in Number of Retained Fourier Modes

To show convergence with increasing number $2M+1$ of retained Fourier modes, we first consider an associated adjoint problem. Since $u^h \in H^2(\Omega)$, we show $O(M^{-2})$ convergence in the L^2 norm. To this end, for $f \in L^2(\Omega)$ we seek a $z_f^h \in H^1_{qp}(\Omega)$ such that

$$\overline{b_{\varepsilon_h}(\xi, z_f^h)} = -\int_\Omega f\bar{\xi} \tag{4.57}$$

for all $\xi \in H^1_{qp}(\Omega)$. Suppose ε is piecewise $C^{1,1}$ and satisfies the conditions of Theorem 4.4.2 or any of its corollaries. The adjoint problem has has a unique solution in $H^1_{qp}(\Omega)$ because ε_h is piecewise $C^{1,1}$ and also satisfies the non-trapping conditions. The Galerkin orthogonality

$$b_{\varepsilon_h}(u^h - u^{h,M}, v_M) = 0 \tag{4.58}$$

holds for all $v_M \in V_M$ because $u^{h,t} = u^h + u^i$ solves problem (4.4) with ε_h instead of ε and the RCWA solution $u^{h,M,t} = u^{h,M} + u^i$ solves (3.71). Taking $\xi = u^h - u^{h,M}$ in the adjoint problem, we have

$$\begin{aligned}\left\|u^h - u^{h,M}\right\|_{L^2(\Omega)} &\leq \gamma \left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} \sup_{0\neq f\in L^2(\Omega)} \left(\frac{1}{\|f\|_{L^2(\Omega)}} \inf_{v_M \in V_M} \left\|z_f^h - v_M\right\|_{H^1(\Omega)}\right) \\ &\leq C_3(\kappa,\varepsilon)\gamma \left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} M^{-1}. \end{aligned} \tag{4.59}$$

This follows because $\left\|z_f^h - \mathcal{F}_M z_f^h\right\|_{H^1(\Omega)} \leq M^{-1} \left\|z_f^h\right\|_{H^2(\Omega)}$, and by 4.5.1.1.

Theorem 4.5.2. *Suppose ε is piecewise $C^{1,1}$ in $\mathbb{R}^2$ and satisfies the conditions of Theorem 4.4.2 or any of its corollaries. Let $u^h \in H^1_{qp}(\Omega)$ be the solution to problem (4.3) with $f = \kappa^2 u^i(\varepsilon_+ - \varepsilon)$ on the right hand side, and $u^{h,M,t}$ be the RCWA solution. Then there is a constant $C > 0$ independent of h and M such that*

$$\left\|u^{h,t} - u^{h,M,t}\right\|_s \leq CM^{s-2}, \tag{4.60}$$

where $s \in \{0, 1\}$ and M is large enough.

Proof. Using Galerkin orthogonality again, it follows that

$$b_{\varepsilon_h}(u^h - u^{h,M}, u^h - u^{h,M}) = b_{\varepsilon_h}(u^h - u^{h,M}, u^h - \mathcal{F}_M u^h), \tag{4.61}$$

where $\mathcal{F}_M$ is the Fourier truncation operator. We also have that the sesquilinear form $b_{\varepsilon_h}(\cdot, \cdot)$ satisfies the Gärding inequality

$$\left|b_{\varepsilon_h}(w, w)\right| \geq \|w\|^2_{H^1(\Omega)} - \kappa^2 \max_{\boldsymbol{x}\in\Omega} \left(\Re(\varepsilon) + 1\right) \|w\|^2_{L^2(\Omega)} \tag{4.62}$$

for all $w \in H^1_{qp}(\Omega)$.

By using the Galerkin orthogonality (4.61), the Gärding inequality and the boundedness of $b_{\varepsilon_h}(\cdot, \cdot)$, we have

$$\begin{aligned} \left\|u^h - u^{h,M}\right\|^2_{H^1(\Omega)} - \kappa^2 \max_{\boldsymbol{x}\in\Omega} \left(\Re(\varepsilon) + 1\right) \left\|u^h - u^{h,M}\right\|^2_{L^2(\Omega)} \\ \leq \gamma \left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} \left\|u^h - \mathcal{F}_M u^h\right\|_{L^2(\Omega)}. \end{aligned} \tag{4.63}$$

Now dividing (4.63) through by $\left\|u^h - u^{h,M}\right\|_{H^1(\Omega)}$ and using the inequality (4.59), we have

$$\begin{aligned} \left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} \leq \gamma \left\|u^h - \mathcal{F}_M u^h\right\|_{H^1(\Omega)} \\ + \kappa^2 \max_{\boldsymbol{x}\in\Omega} \left(\Re(\varepsilon) + 1\right) C_3(\kappa, \varepsilon)\gamma \left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} M^{-1}. \end{aligned} \tag{4.64}$$

Now we take $M \geq 2\kappa^2 \max_{\boldsymbol{x}\in\Omega} \left(\Re(\varepsilon) + 1\right) C_3(\kappa, \varepsilon)\gamma$ in (4.64) to obtain

$$\left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} \leq 2\gamma \left\|u^h - \mathcal{F}_M u^h\right\|_{H^1(\Omega)}. \tag{4.65}$$

Using standard properties of Fourier series, it follows that

$$\left\|u^h - \mathcal{F}_M u^h\right\|_{H^1(\Omega)} \leq \|u^h\|_{H^2(\Omega)} M^{-1} \tag{4.66}$$

$$\leq \kappa^2 C_3(\kappa, \varepsilon) \left\|u^i(\varepsilon_+ - \varepsilon_h)\right\|_{L^2(\Omega)} M^{-1}, \tag{4.67}$$

where we have used that $u^h \in H^2(\Omega)$ and satisfies the *a-priori* estimate given in 4.5.1.1. To obtain a constant independent of $h > 0$, we see that

$$\left\|u^i(\varepsilon_+ - \varepsilon_h)\right\|_{L^2(\Omega)} \leq \left\|u^i\right\|_{L^2(\Omega)} \left(\varepsilon_+ + 3\|\varepsilon\|_{L^\infty(\Omega)}\right). \tag{4.68}$$

To complete the proof, using (4.59) again we see the extra order of convergence in the L^2 norm:

$$\left\|u^h - u^{h,M}\right\|_{L^2(\Omega)} \leq 2\gamma^2\kappa^2 C_3(\kappa, \varepsilon)^2 \left(\varepsilon_+ + 3\|\varepsilon\|_{L^\infty(\Omega)}\right) \left\|u^i\right\|_{L^2(\Omega)} M^{-2}. \tag{4.69}$$

□

4.5.2 Convergence in Slice Thickness

This section concerns the approximation theory of the RCWA with respect to slice thickness.

Theorem 4.5.3. *Suppose ε is piecewise $C^{1,1}$ in $\mathbb{R}^2$ and satisfies the conditions of 4.4.2 or any of its corollaries. Suppose that all interfaces are the graphs of piecewise C^2 functions. Let $u^h \in H^1_{qp}(\Omega)$ be the solution to problem (4.3) with $f = \kappa^2 u^i(\varepsilon_+ - \varepsilon_h)$ on the right hand side, and u be the solution of (4.2) with $f = \kappa^2 u^i(\varepsilon_+ - \varepsilon)$ on the right hand side. Then there is a constant $C > 0$ independent of h and M such that*

$$\left\|u - u^h\right\|_{H^1(\Omega)} \leq Ch^{1/2}.$$

Proof. We notice that

$$b_{\varepsilon_h}(u - u^h, v) = \int_\Omega \kappa^2(\varepsilon - \varepsilon_h) u^t \overline{v} \tag{4.70}$$

for all $v \in H^1_{qp}(\Omega)$. The right hand side $\kappa^2(\varepsilon - \varepsilon_h)u^t \in L^2(\Omega)$ and ε_h satisfies the conditions of 4.4.2 or any of its corollaries, depending on the assumption on ε. Therefore, there is a constant C independent of $h > 0$ such that

$$\begin{aligned}
\left\|u - u^h\right\|_{H^1(\Omega)} &\leq C\kappa^2 \left\|u^t\right\|_{L^\infty(\Omega)} \|\varepsilon - \varepsilon_h\|_{L^2(\Omega)} \\
&\leq C\kappa^2 \left\|u^t\right\|_{L^\infty(\Omega)} h^{1/2} \\
&\leq C\kappa^2 \|u^t\|_{H^2(\Omega)} h^{1/2} \\
&\leq C\kappa^2 \left(\|u^i\|_{H^2(\Omega)} + C_3(\kappa, \varepsilon) \left\|\kappa^2 u^i(\varepsilon_+ - \varepsilon)\right\|_{L^2(\Omega)} \right) h^{1/2}.
\end{aligned} \tag{4.71}$$

□

We now summarize the convergence results of this chapter:

Theorem 4.5.4. *Suppose ε is piecewise $C^{1,1}$ in $\mathbb{R}^2$ and satisfies the conditions of 4.4.2 or any of its corollaries. Let $u^t \in H^1_{qp}(\Omega)$ is the solution to problem (4.4), and $u^{h,M,t}$ be the RCWA solution. For M large enough, there is a constant $C > 0$ independent of h and M such that*

$$\left\|u^t - u^{h,M,t}\right\|_s \leq C\left(h^{1/2} + M^{s-2}\right), \tag{4.72}$$

with $s \in \{0, 1\}$.

Proof. These estimates follow from Theorems 4.5.2 and 4.5.3, since

$$\begin{aligned}
\left\|u^t - u^{h,M,t}\right\|_s &\leq \left\|u^t - u^{h,t}\right\|_s + \left\|u^{h,t} - u^{h,M,t}\right\|_s \\
&\leq \left\|u - u^h\right\|_{H^1(\Omega)} + \left\|u^{h,t} - u^{h,M,t}\right\|_s.
\end{aligned}$$

□

4.6 Numerical Examples

In this section we test Theorems 4.5.2 and 4.5.3 numerically by comparing the RCWA solution to a highly refined FEM solution. In order to avoid possible convergence enhancements due to symmetry, we study a non-symmetric grating profile. The example is shown in Figure 4.2a. We also show results for a symmetric grating, but the

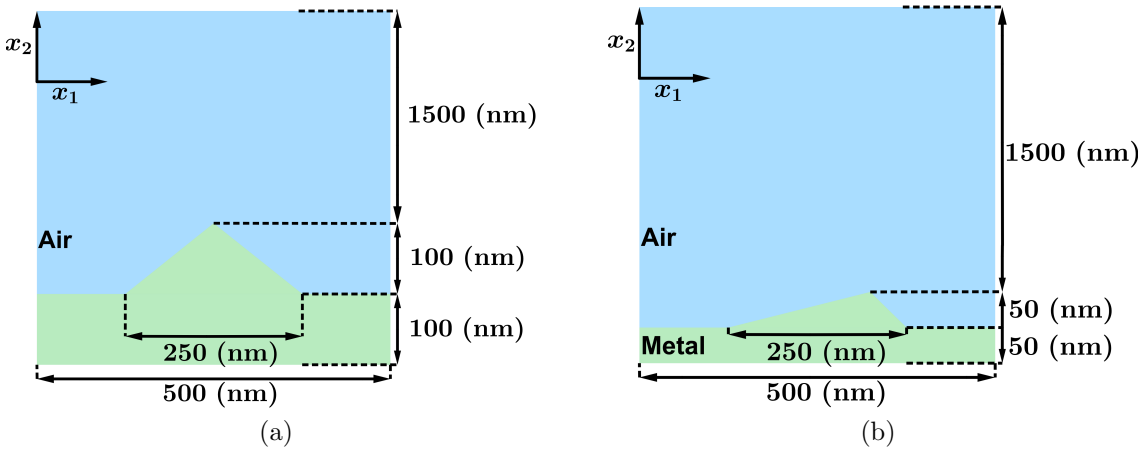


Figure 4.2: (a) Symmetric grating of maximum height 100 nm. (b) Asymmetric grating of maximum height 50 nm. The peak of the asymmetric grating is off center to the right by 62.5 nm. The thickness of the air layer is not to scale.

grating is taller to determine if the grating height effects the convergence with respect to the slice thickness h. In both of our examples, the relative permittivity of the fictitious metallic material is given as $\varepsilon_m = -15 + 4i$, while the relative permittivity of air is $\varepsilon_a = 1$. The thickness of the air layer is 1500 nm and the period $L = 500$ nm along the x_1 direction. In the first example, the non-symmetric grating of maximum height 50 nm is backed by a 50-nm-thick metallic layer beneath it. The symmetric grating has a maximum height of 100 nm. A plane wave in both examples is normally incident (i.e., $\theta = 0$) and the free-space wavelength $\lambda_0 = 2\pi/\kappa = 600$ nm. These parameters, while not describing any particular physical problem, are typical for practical applications.

Since the true solution to these problems cannot be computed analytically, we compare the RCWA solution to a highly refined FEM solution. The FEM solution u_{FE} in each example was computed using an adaptive method implemented in NGSolve [47]. The simulated domain is sandwiched between two perfectly matched layers (PMLs). Both of the PMLs are one wavelength thick and have a constant PML parameter of $1.5 + 2.5i$ [19]. This gives a reflection coefficient of 3×10^{-12}. The FEM solution was computed using 5th-order continuous finite elements. The adaptive algorithm uses mesh bisection and the Zienkiewicz–Zhu *a-posteriori* error estimator [20]. Mesh

adaptivity terminates when the algorithm reaches 100,000 degrees of freedom. We define the relative L^2 error between an RCWA solution and the FEM solution to be

$$\frac{\left\|u^{h,M,t} - u_{\mathrm{FE}}\right\|_{L^2(\Omega)}}{\left\|u_{\mathrm{FE}}\right\|_{L^2(\Omega)}}.$$

Figures 4.3a and 4.3b show the convergence of the non-symmetric example with respect to M and h, respectively. Figures 4.3c and 4.3d show the convergence of the symmetric example, similarly in Figs. 4.3b and 4.3d, the number of retained Fourier modes was fixed as $2M + 1 = 101$. Slice thickness h was allowed to change, where $h \in \{1/2, 1, 1.25, 2, 5, 10, 25, 50\}$ nm. In Figs. 4.3a and 4.3c, the slice thickness $h = 1$ nm was fixed but the number $2M + 1$ of retained Fourier modes was allowed to change with $M = 1, 2, \cdots, 50$.

We see that the rate of convergence is $O(h^{1.7})$ for the symmetric grating, and $O(h^{1.56})$ for the non-symmetric grating. In general, we can only prove at least $O(h^{1/2})$ in Theorem 4.5.3, so in some cases the convergence due to stairstepping error is better than predicted. The rate of convergence for the number of retained Fourier modes is, as predicted, $O(M^{-2})$ for both examples.

For results in a complicated grating motivated by solar cell applications see [34]. Convergence in h was not considered, but $O(M^{-2})$ convergence is seen.

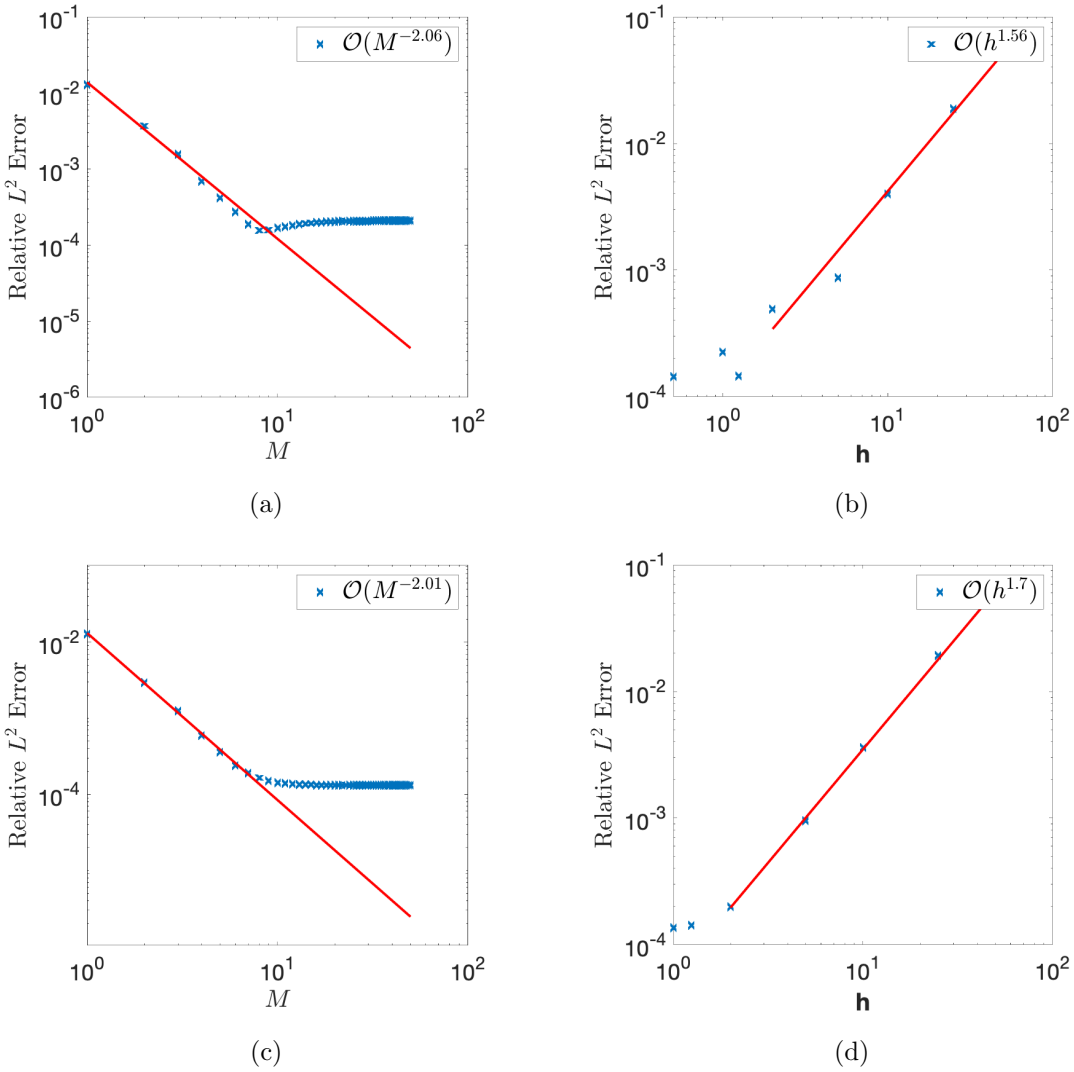


Figure 4.3: Convergence plots comparing the RCWA solution to a highly refined FEM solution. In (b) and (d), the number of retained Fourier modes was fixed as $2M + 1 = 101$. Slice thickness h was allowed to change, where $h \in \{1/2, 1, 1.25, 2, 5, 10, 25, 50\}$ nm. In (a) and (c), the slice thickness $h = 1$ nm was fixed and the number $2M + 1$ of retained Fourier modes was allowed to change with $M = 1, 2, \cdots, 50$. In all cases the error saturates around 10^{-4}.

4.7 Conclusion

In this section we studied the convergence properties of the 2D RCWA for s-polarized incident light. Our analysis relies on the fact that the RCWA solution solves the appropriate variational problem, and therefore we borrowed techniques from the analysis of the FEM. Since the RCWA discretizes the solution in two different ways, we provided theorems for the convergence of the method in terms of the number of retained Fourier modes and slice thickness. Our analysis assumes a non-trapping domain, which is not always true for many common RCWA applications. As we commented earlier in the section, our theory also predicts convergence in the trapping case, as long as both continuity constants in the *a-priori* estimates for problems (4.2) and (4.3) are bounded independent of h.

Chapter 5

ANALYSIS OF RCWA FOR P-POLARIZED LIGHT

5.1 Introduction

The case we study in this chapter is the one in which the domain is illuminated by a monochromatic p-polarized plane wave. Then the RCWA is equivalent to solving a system of second-order ODEs relating the Fourier modes of the magnetic field phasor [5]. However, as even this truncated system is difficult to solve, following the strategy adopted for the s-polarized case, the true relative permittivity is replaced by an approximation whereby the domain is discretized into thin slices in the direction perpendicular to the periodicity of the grating. In each slice, the true relative permittivity is approximated so that the solution in each slice can be computed without further approximation. The fast linear-algebraic algorithm described in Section 3.4 can then be used for this problem, by enforcing the continuity of the solution and the its normal derivative across the inter-slice boundaries [28, 29, 2]. The approximate solution in the entire domain is then formed by stitching together the solutions for the slices.

This chapter is organized as follows. In Section 5.2, we state the mathematical problem: an inhomogeneous Helmholtz equation with quasi-periodic boundary conditions. We state a variational formulation for our problem in Section 5.2.1. In Section 5.3 we derive a Rellich identity for solutions of the Helmholtz problem, assuming that ε is real and C^∞ smooth. Then we show that an *a-priori* estimate holds, where the continuity constant has explicit dependence on ε as long as certain non-trapping conditions are fulfilled. These non-trapping conditions ensure, roughly speaking, that a quantum of electromagnetic energy entering the domain leaves it after a finite time. We extend these *a-priori* estimates to more general ε in Section 5.4, using generalized

non-trapping conditions [35]. In particular, we use these results to show that if ε is piecewise smooth and satisfies the generalized non-trapping conditions, then the perturbed ε problem also satisfies them. Convergence in slice thickness is shown in Section 5.6 to follow from the foregoing conclusion. We further extend the *a-priori* estimates obtained in Section 5.5. Convergence in retained Fourier modes is shown in Section 5.7. The case where ε is everywhere complex is considered in Section 5.8. Finally, we test our prediction of convergence order with some numerical examples in Section 5.9 by comparing the RCWA solution with the solution delivered by a highly refined Finite Element Method.

The content of this chapter has been submitted for publication as a paper [54].

5.2 The Continuous Problem

A p-polarized plane wave propagating in the half-space $x_2 > H$ at an angle θ with respect to the x_2-axis is incident on the plane $x_2 = H$; the sole non-zero component of its magnetic field phasor is denoted as $u^i = H_3^{inc}$.

The scattered magnetic field phasor $u\boldsymbol{e}_3$ is given in terms of the total field u^t by $u = u^t - u^i$, where u satisfies the Helmholtz equation (2.72)–(2.74), $A = \varepsilon^{-1}\boldsymbol{I}$, $a = 1$, and

$$f = \nabla \cdot \left[(\varepsilon_+^{-1} - \varepsilon^{-1})\nabla u^i\right]. \tag{5.1}$$

As in the previous chapter, we assume $\varepsilon_- = \varepsilon_+$ so $f \equiv 0$ outside Ω. In Sections 5.3 and 5.4 we assume that $\varepsilon \in C^\infty(\mathbb{R}^2)$ so that $f \in L^2(\Omega)$. However, we are interested in the case where ε is only piecewise smooth so that $f \notin L^2(\Omega)$ in general. We discuss the regularity of f in more detail later in this chapter. In addition, we note that u is quasi-periodic in Ω, accounting for the multiplicative factors in equations (2.73) and (2.74).

We also assume that ε is piecewise C^2 in $\mathbb{R}^2$ and that either

I. ε is real and $\Re(\varepsilon) > 0$, or

II. ε is complex, $\Im(\varepsilon) > c_1 > 0$, and $\Re(\varepsilon) > c_2 > 0$ in Ω and a positive real constant elsewhere.

Case I encompasses insulators whereas Case II covers dissipative dielectric materials (but not metals).

To show convergence of the RCWA in Case I, we prove a Rellich identity for the problem and show that an *a-priori* estimate holds when $\varepsilon \in C^\infty(\mathbb{R}^2)$. Using a technique of Graham *et al.* [35], we then show that several different *a-priori* bounds hold for the chosen problem, even if $\varepsilon \in L^\infty(\Omega)$. Under the assumption of certain non-trapping conditions, the continuity constants in the *a-priori* estimates can be written explicitly in terms of κ and ε. In Case II, the problem is coercive and we employ the Strang lemmas [37] to prove convergence; hence, non-trapping conditions are unnecessary.

5.2.1 Variational Formulation

To prove convergence of the RCWA, we need to consider several different variational problems, because the approach replaces the true ε with an approximation ε_h. The approximation ε_h was defined as (3.2).

Next, let us replace the source function f defined in (5.1) by a more general source function denoted by F. We now state the variational problems that we study in this chapter.

Like we showed in Chapter 2.4, after multiplying (2.72) by a test function and using the divergence theorem in the usual way, the resulting sesquilinear form $B_\epsilon(\cdot,\cdot) : H^1_{qp}(\Omega) \times H^1_{qp}(\Omega) \to \mathbb{C}$ is obtained as

$$B_\varepsilon(w,v) = \int_\Omega \left(\frac{1}{\varepsilon} \nabla w \cdot \nabla \overline{v} - \kappa^2 w \overline{v} \right) - \int_{\Gamma_{-H}} \overline{v} T_p^-(w) - \int_{\Gamma_H} \overline{v} T_p^+(w) \tag{5.2}$$

for all $w \in H^1_{qp}(\Omega)$ and $v \in H^1_{qp}(\Omega)$. Given an $F \in (H^1_{qp}(\Omega))'$, we seek a solution $u \in H^1_{qp}(\Omega)$ such that

$$B_\epsilon(u,v) = F(v) \tag{5.3}$$

for all $v \in H^1_{qp}(\Omega)$, where $F(v) = -\int_\Omega F\overline{v}$.

We are also interested in a perturbed problem in which ε is replaced by ε_h. Therefore, we define $B_{\varepsilon_h}(\cdot,\cdot)$ to be the same as $B_\varepsilon(\cdot,\cdot)$ but with ε_h instead of ε. Given an $F \in (H^1_{qp}(\Omega))'$, we seek a solution $u^h \in H^1_{qp}(\Omega)$ such that

$$B_{\epsilon_h}(u^h, v) = F(v) \tag{5.4}$$

for all $v \in H^1_{qp}(\Omega)$.

To show that both of the foregoing problems have unique solutions, we have to show that either a Rellich identity holds for our problem and implies an *a-priori* estimate, or the problem is coercive depending on our assumptions on ε.

We now recall some properties of the Dirichlet-to-Neumann boundary integrals appearing in the sequilinear form (5.2). The signs of the real and imaginary parts of the Dirichlet-to-Neumann integral on Γ_H are known by virtue of Lemma 2.4.2. These facts are used many times throughout this chapter.

In some arguments throughout this chapter, it is useful to use the total magnetic field $u^t = u + u^i$, instead of the scattered field u. Therefore, we conclude this section by giving the variational problem for the total magnetic field. We seek a $u^t \in H^1_{qp}(\Omega)$ such that

$$B_\varepsilon(u^t, v) = \int_{\Gamma_H} \overline{v}\left(\frac{1}{\varepsilon_+}\frac{\partial u^i}{\partial x_2} - T_p^+(u^i)\right) \tag{5.5}$$

for all $v \in H^1_{qp}(\Omega)$.

5.3 A Rellich Identity for Quasi-periodic Solutions

In this section, we apply techniques developed by Lechleiter & Ritterbusch [16] for scattering by an arbitrarily rough surface to our quasi-periodic case. We assume that $\varepsilon \in C^\infty(\mathbb{R}^2)$ here, and later on we will show that similar *a-priori* estimates hold even when $\varepsilon \in L^\infty(\mathbb{R}^2)$. These estimates are used to address Case I of Section 5.2.

Theorem 5.3.1. *Assume that $\varepsilon \in C^\infty(\mathbb{R}^2)$ is real where $\Re(\varepsilon) > 0$, and $\varepsilon = \varepsilon_+$ for $|x_2| > H$. If $u \in H^1_{qp}(\Omega)$ is a solution to the variational problem (5.3) for a source $F \in L^2(\Omega)$, then the following Rellich identity holds:*

$$\begin{aligned}
&\int_\Omega \left[\frac{2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 - (x_2+H)\frac{\partial}{\partial x_2}\left(\frac{1}{\varepsilon}\right)\left|\nabla u\right|^2\right] + 2H\int_{\Gamma_H}\left(\frac{-2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 + \frac{1}{\varepsilon}\left|\nabla u\right|^2 - \kappa^2\left|u\right|^2\right)\\
&- \int_{\Gamma_H}\overline{u}T_p^+(u) - \int_{\Gamma_{-H}}\overline{u}T_p^-(u)\\
&= -2\int_\Omega (x_2+H)\Re\left(\overline{F}\frac{\partial u}{\partial x_2}\right) - \int_\Omega F\overline{u}.
\end{aligned} \tag{5.6}$$

Proof. Since $\varepsilon \in C^\infty(\mathbb{R}^2)$ and $F \in L^2(\Omega)$, $u \in H^2(\Omega)$. Using the identity

$$2\Re\left(\frac{\partial u}{\partial x_2}\overline{u}\right) = \frac{\partial}{\partial x_2}\left|u\right|^2, \tag{5.7}$$

and the Helmholtz equation $\nabla\cdot\left(\frac{1}{\varepsilon}\nabla\overline{u}\right) = \overline{F} - \kappa^2\overline{u}$, we obtain

$$2\Re\int_\Omega (x_2+H)\frac{\partial u}{\partial x_2}\nabla\cdot\left(\frac{1}{\varepsilon}\nabla\overline{u}\right) = \int_\Omega (x_2+H)2\Re\left(\frac{\partial u}{\partial x_2}\overline{F}\right) - \kappa^2\int_\Omega (x_2+H)\frac{\partial}{\partial x_2}\left|u\right|^2. \tag{5.8}$$

Integrating the last term in (5.8) by parts, we have

$$\begin{aligned}
2\Re\int_\Omega (x_2+H)\frac{\partial u}{\partial x_2}\nabla\cdot\left(\frac{1}{\varepsilon}\nabla\overline{u}\right) &= \int_\Omega (x_2+H)2\Re\left(\frac{\partial u}{\partial x_2}\overline{F}\right)\\
&- 2H\int_{\Gamma_H}\kappa^2\left|u\right|^2 + \kappa^2\int_\Omega\left|u\right|^2.
\end{aligned} \tag{5.9}$$

Using the divergence theorem, we obtain

$$\begin{aligned}
\int_\Omega (x_2+H)\frac{\partial u}{\partial x_2}\nabla\cdot\left(\frac{1}{\varepsilon}\nabla\overline{u}\right) &= -\int_\Omega\left[\frac{1}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 + (x_2+H)\frac{1}{\varepsilon}\nabla\left(\frac{\partial u}{\partial x_2}\right)\cdot\overline{\nabla u}\right]\\
+ 2H\int_{\Gamma_H}\frac{1}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 &+ \int_{\Gamma_R}(x_2+H)\frac{1}{\varepsilon}\frac{\partial u}{\partial x_2}\overline{\frac{\partial u}{\partial x_1}} - \int_{\Gamma_L}(x_2+H)\frac{1}{\varepsilon}\frac{\partial u}{\partial x_2}\overline{\frac{\partial u}{\partial x_1}}.
\end{aligned} \tag{5.10}$$

Using the quasi-periodicity of the solution, we see that

$$\int_{\Gamma_R}(x_2+H)\frac{1}{\varepsilon}\frac{\partial u}{\partial x_2}\overline{\frac{\partial u}{\partial x_1}} - \int_{\Gamma_L}(x_2+H)\frac{1}{\varepsilon}\frac{\partial u}{\partial x_2}\overline{\frac{\partial u}{\partial x_1}} = 0. \tag{5.11}$$

We take twice the real part of (5.10), use the identity $2\Re\left[\nabla\left(\frac{\partial u}{\partial x_2}\right)\cdot\overline{\nabla u}\right]=\frac{\partial}{\partial x_2}\left|\nabla u\right|^2$ therein, and integrate by parts. Then using quasi-periodicity again, we obtain

$$\int_{\Gamma_L}(x_2+H)\frac{1}{\varepsilon}\left|\nabla u\right|^2-\int_{\Gamma_R}(x_2+H)\frac{1}{\varepsilon}\left|\nabla u\right|^2=0. \tag{5.12}$$

Hence, (5.8) can be rewritten as follows:

$$\begin{aligned}2\Re\int_\Omega(x_2+H)\frac{\partial u}{\partial x_2}\nabla\cdot\left(\frac{1}{\varepsilon}\nabla\overline{u}\right)=&-\int_\Omega\left[\frac{2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2-(x_2+H)\frac{\partial}{\partial x_2}\left(\frac{1}{\varepsilon}\right)\left|\nabla u\right|^2\right]\\&+2H\int_{\Gamma_H}\left(\frac{2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2-\frac{1}{\varepsilon}\left|\nabla u\right|^2\right)+\int_\Omega\frac{1}{\varepsilon}\left|\nabla u\right|^2.\end{aligned} \tag{5.13}$$

To complete the proof, we equate the right hand sides of (5.9) and (5.13). The last term on the right side of (5.13) is replaced by the identity

$$\int_\Omega\frac{1}{\varepsilon}\left|\nabla u\right|^2=\kappa^2\int_\Omega\left|u\right|^2+\int_{\Gamma_H}\overline{u}T_p^+(u)+\int_{\Gamma_{-H}}\overline{u}T_p^-(u)-\int_\Omega F\overline{u}, \tag{5.14}$$

which can be obtained by setting $v=u$ in the variational problem (5.3). After rearranging some terms, the Rellich identity is shown. □

Now we use the Rellich identity to show that an *a-priori* estimate holds, and that the continuity constant has explicit dependences on κ and ε as long as certain non-trapping conditions are met. We first prove a lemma about controlling the L^2 norm of u, and then use it to determine the *a-priori* estimate.

Lemma 5.3.2. *If $u\in H^1_{qp}(\Omega)$ is a solution to the variational problem (5.3), then*

$$\begin{aligned}\left\|u\right\|^2_{L^2(\Omega)}\leq&\left[4H\varepsilon_+(a+1)+\frac{2H^2}{\min_\Omega\left(\frac{1}{\varepsilon}\right)}\right]\\&\times\left[\Im\int_{\Gamma_H}\overline{u}T_p^+(u)-\Re\int_{\Gamma_H}\overline{u}T_p^+(u)+\left\|\left(\frac{2}{\epsilon}\right)^{1/2}\frac{\partial u}{\partial x_2}\right\|_{L^2(\Omega)}\right],\end{aligned} \tag{5.15}$$

where

$$a=\max_{|\kappa^2\varepsilon_+-\alpha_n^2|<1}\left(\frac{1}{|\beta_n^+|}\right). \tag{5.16}$$

Proof. By virtue of Lemma 4.3 of Ref. [16], we know that

$$\left\|u\right\|^2_{L^2(\Omega)}\leq 4H\int_{\Gamma_H}\left|u\right|^2+4H^2\left\|\frac{\partial u}{\partial x_2}\right\|^2_{L^2(\Omega)}, \tag{5.17}$$

holds for all $u \in H^1(\Omega)$. We notice that by Parseval's theorem and by the definition of β_n^+,

$$\begin{aligned}\int_{\Gamma_H} |u|^2 &\leq \sum_{1+\alpha_n^2<\kappa^2\varepsilon_+} |\beta_n^+||u_n(H)|^2 + \sum_{\alpha_n^2>\kappa^2\varepsilon_++1} |\beta_n^+||u_n(H)|^2 + \sum_{|\kappa^2\varepsilon_+-\alpha_n^2|<1} |u_n(H)|^2 \\ &\leq \varepsilon_+\left[\Im\int_{\Gamma_H}\overline{u}T_p^+(u) - \Re\int_{\Gamma_H}\overline{u}T_p^+(u)\right] + a\sum_{|\kappa^2\varepsilon_+-\alpha_n^2|<1}|\beta_n^+||u_n(H)|^2 \\ &+ \sum_{|\kappa^2\varepsilon_+-\alpha_n^2|<1}(1-a|\beta_n^+|)|u_n(H)|^2. \end{aligned} \tag{5.18}$$

By virtue of our choice of a, $1-a|\beta_n^+| \leq 0$ and the last sum is non-positive. We add back all the missing terms (where $|\kappa^2\varepsilon_+ - \alpha_n^2| > 1$) into the second to last sum and thus obtain

$$\|u\|_{L^2(\Omega)}^2 \leq 4H\varepsilon_+(a+1)\left[\Im\int_{\Gamma_H}\overline{u}T_p^+(u) - \Re\int_{\Gamma_H}\overline{u}T_p^+(u)\right] + 4H^2\left\|\frac{\partial u}{\partial x_2}\right\|_{L^2(\Omega)}^2, \tag{5.19}$$

whereby Lemma 5.3.2 is proved.

□

Our next result is the desired continuity estimate when ε is smooth.

Theorem 5.3.3. *Assume that $\varepsilon \in C^\infty(\mathbb{R}^2)$ is real with $\Re(\epsilon) > 0$ and $\varepsilon = \varepsilon_+$ for $|x_2| > H$. Suppose also that the non-trapping conditions*

1. $\frac{\partial}{\partial x_2}\left(\frac{1}{\varepsilon}\right) \leq 0$ *in* Ω, *and*
2. $\varepsilon = \varepsilon_+$ *on* Γ_H

hold. Then, given an $F \in L^2(\Omega)$, there exists a unique solution $u \in H^1_{qp}(\Omega)$ to the variational problem (5.3) and a continuity constant $C(\kappa,\varepsilon) > 0$ such that

$$\|u\|_{H^1(\Omega)} \leq C(\kappa,\varepsilon)\|F\|_{L^2(\Omega)} \tag{5.20}$$

with

$$C(\kappa,\epsilon) = \min\left(\min_\Omega\left(\frac{1}{\varepsilon}\right),1\right)^{-1} \tag{5.21}$$

$$\times\left\{1+(\kappa^2+1)\left[4H\varepsilon_+(a+1)+\frac{2H^2}{\min_\Omega\left(\frac{1}{\varepsilon}\right)}\right]\left[4H\left(1+\varepsilon_+^{1/2}\kappa\right)+2\right]\right\}. \tag{5.22}$$

Proof. Using the first non-trapping condition and taking the real part of the Rellich identity, we get

$$\int_\Omega \frac{2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 - \Re\int_{\Gamma_H}\overline{u}T_p^+(u) \le -2\int_\Omega (x_2+H)\Re\left(\overline{F}\frac{\partial u}{\partial x_2}\right) - \Re\int_\Omega F\overline{u} + 2H\int_{\Gamma_H}\left(\frac{2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 - \frac{1}{\varepsilon}\left|\nabla u\right|^2 + \kappa^2\left|u\right|^2\right). \tag{5.23}$$

after using the inequality $-\Re\int_{\Gamma_{-H}}\overline{u}T_p^-(u)\ge 0$. Let us recall that u satisfies the same Rayleigh expansion as provided in (2.30). Then, using Parseval's theorem and the second non-trapping condition, we obtain

$$\begin{aligned}
2H\int_{\Gamma_H}\left(\frac{2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 - \frac{1}{\varepsilon}\left|\nabla u\right|^2 + \kappa^2\left|u\right|^2\right) &= 2H\sum_{n\in\mathbb{Z}}\left(\frac{1}{\varepsilon_+}\left|\kappa^2\varepsilon_+ - \alpha_n^2\right| - \frac{1}{\varepsilon_+}\alpha_n^2 + \kappa^2\right)\left|u_n(H)\right|^2 \\
&= 4H\sum_{\alpha_n^2<\kappa^2\varepsilon_+}\left(\kappa^2 - \frac{1}{\varepsilon_+}\alpha_n^2\right)\left|u_n(H)\right|^2 \\
&= 4H\varepsilon_+^{-1}\sum_{\alpha_n^2<\kappa^2\varepsilon_+}\left(\kappa^2\varepsilon_+ - \alpha_n^2\right)\left|u_n(H)\right|^2 \\
&\le 4H\varepsilon_+^{1/2}\kappa\Im\int_\Omega F\overline{u}.
\end{aligned} \tag{5.24}$$

This argument is similar to Lemma 2.2 of Ref. [15], but we have used different Dirichlet-to-Neumann operators. The inequality $\Im\int_{\Gamma_H}\overline{u}T_p^+(u)\le\Im\int_\Omega F\overline{u}$ follows on setting $u=v$ in the variational problem (5.3) and taking the imaginary part therof. We combine this result with (5.23) and add $\Im\int_{\Gamma_H}\overline{u}T_p^+(u)$ to both sides to obtain

$$\int_\Omega \frac{2}{\varepsilon}\left|\frac{\partial u}{\partial x_2}\right|^2 - \Re\int_{\Gamma_H}\overline{u}T_p^+(u) + \Im\int_{\Gamma_H}\overline{u}T_p^+(u) \le -2\int_\Omega (x_2+H)\Re\left(\overline{F}\frac{\partial u}{\partial x_2}\right) - \Re\int_\Omega F\overline{u} + \left(4H\varepsilon_+^{1/2}\kappa + 1\right)\Im\int_\Omega F\overline{u}. \tag{5.25}$$

Then we combine (5.25) and Lemma 5.3.2 to get

$$\begin{aligned}\|u\|^2_{L^2(\Omega)} &\leq 4H\varepsilon_+(a+1)\left[\Im\int_{\Gamma_H}\overline{u}T_p^+(u) - \Re\int_{\Gamma_H}\overline{u}T_p^+(u)\right] + 4H^2\left\|\frac{\partial u}{\partial x_2}\right\|^2_{L^2(\Omega)} \\ &\leq \left[4H\varepsilon_+(a+1) + \frac{2H^2}{\min_\Omega\left(\frac{1}{\varepsilon}\right)}\right] \\ &\quad\times\left[-2\int_\Omega(x_2+H)\Re\left(\overline{F}\frac{\partial u}{\partial x_2}\right) - \Re\int_\Omega F\overline{u} + \left(4H(\varepsilon_+)^{1/2}\kappa+1\right)\Im\int_\Omega F\overline{u}\right] \\ &\leq \left[4H\varepsilon_+(a+1) + \frac{2H^2}{\min_\Omega\left(\frac{1}{\varepsilon}\right)}\right]\left[4H\left(1+\varepsilon_+^{1/2}\kappa\right)+2\right]\|F\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)}.\end{aligned} \tag{5.26}$$

After setting $v = u$ in the variational problem (5.3) and taking the real part thereof, we have

$$\|u\|^2_{H^1(\Omega)} \leq \min\left(\min_\Omega\left(\frac{1}{\varepsilon}\right),1\right)^{-1}\left[\|F\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)} + (\kappa^2+1)\|u\|^2_{L^2(\Omega)}\right]. \tag{5.27}$$

After first combining (5.26) and (5.27) and then dividing the result by $\|u\|_{H^1(\Omega)}$, we obtain the *a-priori* estimate. Existence and uniqueness of u follow because the *a-priori* estimate implies an inf-sup condition (4.30) for $B_\varepsilon(\cdot,\cdot)$ [15].

□

So far we have not discussed problem (5.4) at all. In the trivial case where ε is constant, $\varepsilon = \varepsilon_h$ and there is nothing new to prove. Generally however, ε_h is only piecewise smooth even if $\varepsilon \in C^\infty(\mathbb{R}^2)$, because it has jumps over the inter-slice boundaries. Therefore, to show existence uniqueness and to find an *a-priori* estimate for the ε_h problem (and to cover applications to multilayered devices), we need to allow for coefficients with less smoothness.

5.4 *A-priori* Estimates for L^∞ Coefficients

We assumed in Section 5.3 that $\varepsilon \in C^\infty(\mathbb{R}^2)$ and showed that an *a-priori* estimate holds for non-trapping domains. In this section, we extend the *a-priori* estimates to $\varepsilon \in L^\infty(\mathbb{R}^2)$. Remarkably, the continuity constant defined in the forthcoming Lemma can be used even for general ε, and a general right hand side. Here we use the technique

of Graham et al. [35], but modify their argument slightly for our use. They showed these estimates for an exterior Dirichlet problem, so we check that the results hold for our quasi-periodic problem. First, we prove an *a-priori* estimate where the right side lies in the dual space $(H^1_{qp}(\Omega))'$, but the ε is smooth.

Lemma 5.4.1. *Assume that $\varepsilon \in C^\infty(\mathbb{R}^2)$ is real with $\Re(\varepsilon) > 0$ and that ε satisfies the non-trapping conditions given in Theorem 5.3.3. For general data, $F \in (H^1_{qp}(\Omega))'$, let $\tilde{u} \in H^1_{qp}(\Omega)$ satisfy*

$$B_\epsilon(\tilde{u}, v) = F(v) \tag{5.28}$$

for all $v \in H^1_{qp}(\Omega)$. Then $\tilde{u}$ exists and is unique; furthermore,

$$\|\tilde{u}\|_{H^1(\Omega)} \leq C_1(\kappa, \varepsilon) \|F\|_{\left(H^1_{qp}(\Omega)\right)'}, \tag{5.29}$$

where

$$C_1(\kappa, \varepsilon) = \min\left(\min_\Omega \left(\frac{1}{\varepsilon}\right), \kappa^2 \right)^{-1} \Big[1 + 2\kappa^2 C(\kappa, \varepsilon) \Big]. \tag{5.30}$$

Proof. Define $B^+_\varepsilon(w, v) = B_\varepsilon(w, v) + 2\kappa^2 \int_\Omega w\overline{v}$ for all $w \in H^1_{qp}(\Omega)$ and $v \in H^1_{qp}(\Omega)$. Since the signs of the real and imaginary parts of the Dirichlet-to-Neumann integrals on $\Gamma_{\pm H}$ are known, we have

$$-\Re \int_{\Gamma_H} \overline{v} T^+_p(v) - \Re \int_{\Gamma_{-H}} \overline{v} T^-_p(v) \geq 0, \tag{5.31}$$

and the sesquilinear form $B^+_\varepsilon(\cdot, \cdot)$ is coercive since

$$\left|B^+_\varepsilon(v, v)\right| \geq \Re\big(B^+_\varepsilon(v, v)\big) \geq \min\left(\min_\Omega \left(\frac{1}{\varepsilon}\right), \kappa^2 \right) \|v\|^2_{H^1(\Omega)} \tag{5.32}$$

for every $v \in H^1_{qp}(\Omega)$.

By virtue of the Lax–Milgram Lemma [39, 40], given an $F \in (H^1_{qp}(\Omega))'$ we define $u^+ \in H^1_{qp}(\Omega)$ to be the solution to the problem

$$B^+_\varepsilon(u^+, v) = F(v), \tag{5.33}$$

for all $v \in H^1_{qp}(\Omega)$; furthermore,

$$\left\|u^+\right\|_{H^1(\Omega)} \leq \min\left(\min_\Omega \left(\frac{1}{\varepsilon}\right), \kappa^2 \right)^{-1} \|F\|_{\left(H^1_{qp}(\Omega)\right)'}. \tag{5.34}$$

First, we let $q \in H^1_{qp}(\Omega)$ be the solution to

$$B_\varepsilon(q, v) = 2\kappa^2 \int_\Omega u^+ \overline{v} \tag{5.35}$$

for all $v \in H^1_{qp}(\Omega)$, which exists and is unique because $2\kappa^2 u^+ \in L^2(\Omega)$; then, we apply Theorem 5.3.3. To complete the proof, we notice that $B_\varepsilon(u^+ + q, v) = B^+_\varepsilon(u^+, v) = F(v)$ for all $v \in H^1_{qp}(\Omega)$. Since $\tilde{u} = q + u^+$, a solution exists and the *a-priori* estimate shows it is unique. □

Now we extend our *a-priori* results to $L^\infty(\mathbb{R}^2)$ coefficients. For a periodic $\varepsilon \in L^\infty(\mathbb{R}^2)$, we seek a sequence of smooth and periodic $C^\infty(\mathbb{R}^2)$ functions that converge to ε in the sense of L^2. To use our previous results, this sequence must be uniformly bounded and each function in the sequence has to satisfy the non-trapping conditions given in Theorem 5.3.3. To do this, we first prove the following theorem, which is an analog of Theorem 2.7 of Ref. [35].

Theorem 5.4.2. *Let $\phi \in L^\infty(\mathbb{R}^2)$ be given such that ϕ is periodic with period L_x in x_1 almost everywhere and there are two constants ϕ_{min} and $\phi_{max} \geq \phi_{min}$ such that*

$$\phi_{min} \leq \phi \leq \phi_{max}, \tag{5.36}$$

almost everywhere in Ω. We can write

$$\phi(\boldsymbol{x}) = \phi_{min} + \Pi(\boldsymbol{x}), \tag{5.37}$$

where $\Pi \in L^\infty(\mathbb{R}^2)$ is almost everywhere L_x-periodic in x_1. Provided that for all $\tau \geq 0$, ϕ is monotonically increasing in the x_2-direction, i.e.,

$$\operatorname*{ess\,inf}_{\boldsymbol{x} \in \Omega} \Big[\Pi(\boldsymbol{x} + \tau \boldsymbol{e}_2) - \Pi(\boldsymbol{x}) \Big] \geq 0, \tag{5.38}$$

there is a sequence $\phi_\delta \in C^\infty(\mathbb{R}^2)$ of L_x-periodic functions in x_1 such that

1. *$\|\phi - \phi_\delta\|_{L^2(\Omega)} \to 0$ as $\delta \to 0$,*
2. *$\phi_{min} \leq \phi_\delta \leq \phi_{max}$, and*

3. $\frac{\partial}{\partial x_2}\phi_\delta \geq 0$.

Proof. Consider the extended domain $\mathcal{U} = \{\boldsymbol{x} \in \mathbb{R}^2,\ -L_x < x_1 < 2L_x\}$ and define $\psi \in C_0^\infty(\mathbb{R}^2)$ as

$$\psi(\boldsymbol{x}) = \begin{cases} C \exp\left(\frac{1}{|\boldsymbol{x}|-1}\right) & \text{if } |\boldsymbol{x}| < 1, \\ 0 & \text{if } |\boldsymbol{x}| > 1, \end{cases} \tag{5.39}$$

where we choose C so that $\int_{\mathbb{R}^2} \psi = 1$. Let $\psi_\delta(\boldsymbol{x}) = \delta^{-2}\psi(\boldsymbol{x}/\delta)$ for $\delta > 0$. Furthermore, let $\phi_\delta \in C^\infty(\mathcal{U}_\delta)$ be defined as

$$\phi_\delta(\boldsymbol{x}) = \phi_{\min} + (\Pi * \psi_\delta)(\boldsymbol{x}) = \phi_{\min} + \int_{\mathbb{R}^2} \Pi(\boldsymbol{x} - \boldsymbol{y})\psi_\delta(\boldsymbol{y})\ d\boldsymbol{y}, \tag{5.40}$$

where

$$\mathcal{U}_\delta = \{\boldsymbol{x} \in \mathcal{U},\ \operatorname{dist}(\boldsymbol{x}, \partial\mathcal{U}) > \delta\}. \tag{5.41}$$

Using standard properties of mollifiers (e.g., Theorem 7 in Ref. [40, Sec. C.5]), we have that $\|\phi - \phi_\delta\|_{L^2(\mathcal{V})} \to 0$ as $\delta \to 0$, for any compact subset $\mathcal{V}$ of $\mathcal{U}$. If we choose $\delta < L_x/2$ then $\Omega \subset \mathcal{U}_\delta \subset\subset \mathcal{U}$, so that $\|\phi - \phi_\delta\|_{L^2(\Omega)} \to 0$ as $\delta \to 0$. The condition $\phi_{\min} \leq \phi_\delta \leq \phi_{\max}$ follows from the definition of ϕ_δ. To finish the proof, we notice that

$$\begin{aligned} \phi_\delta(x_1 + L_x, x_2) &= \phi_{\min} + \int_{|\boldsymbol{y}|<\delta} \Pi(\boldsymbol{x} - \boldsymbol{y} + L_x\boldsymbol{e}_1)\psi_\delta(\boldsymbol{y})\ d\boldsymbol{y} \\ &= \phi_\delta(x_1, x_2), \end{aligned} \tag{5.42}$$

since $\Pi(\boldsymbol{x})$ is an L_x-periodic function in x_1. To show that each ϕ_δ satisfies the non-trapping condition, we see that

$$(\Pi * \psi_\delta)(\boldsymbol{x} + \tau\boldsymbol{e}_2) - (\Pi * \psi_\delta)(\boldsymbol{x}) \geq \operatorname*{ess\,inf}_{\boldsymbol{x}\in\Omega} \left[\Pi(\boldsymbol{x} + \tau\boldsymbol{e}_2) - \Pi(\boldsymbol{x})\right] \int_{|\boldsymbol{y}|<\delta} \psi_\delta(\boldsymbol{y})\ d\boldsymbol{y}, \tag{5.43}$$

for every $\tau \geq 0$. Since the ψ_δ are positive functions of compact support, this implies that $\frac{\partial}{\partial x_2}\phi_\delta \geq 0$.

□

The next result is the main result of this section, and it proves an *a-priori* estimate for our problem with a general source term and a general non-trapping condition.

Theorem 5.4.3. *Given $\varepsilon \in L^\infty(\mathbb{R}^2)$, assume that the generalized non-trapping condition*

$$\operatorname*{ess\,inf}_{\boldsymbol{x}\in\Omega}\left[\tilde{\Pi}(\boldsymbol{x}+\tau\boldsymbol{e}_2)-\tilde{\Pi}(\boldsymbol{x})\right]\geq 0 \tag{5.44}$$

holds for all $\tau \geq 0$, where $\varepsilon(\boldsymbol{x}) = \varepsilon_{min} + \tilde{\Pi}(\boldsymbol{x})$. Then for $F \in \left(H^1_{qp}(\Omega)\right)'$, the solution $u \in H^1_{qp}(\Omega)$ of

$$B_\epsilon(u,v) = F(v) \tag{5.45}$$

for all $v \in H^1_{qp}$ exists and is unique; furthermore,

$$\|u\|_{H^1(\Omega)} \leq C_1(\kappa,\varepsilon)\,\|F\|_{\left(H^1_{qp}(\Omega)\right)'}.$$

Remark 8. *The generalized non-trapping condition means that ε is monotonically increasing in the x_2-direction. We can also prove the same result for ε monotonically decreasing in the x_2-direction.*

Proof. Since $H^1_{qp}(\Omega) = \overline{C^\infty_{qp}(\Omega)}$ where the closure is taken in the sense of $H^1(\Omega)$, given a $\xi > 0$ we can choose a $u_\xi \in C^\infty_{qp}(\Omega)$ such that

$$\|u-u_\xi\|_{H^1(\Omega)} < \xi. \tag{5.46}$$

We see that $\phi = \varepsilon$ satisfies the conditions of Theorem 5.4.2, and so we have a sequence of smooth and periodic functions $\phi_\delta \in C^\infty(\mathbb{R}^2)$ such that $\|\phi_\delta - \varepsilon\|_{L^2(\Omega)} \to 0$ as $\delta \to 0$. The ϕ_δ also satisfy the non-trapping conditions of Theorem 5.3.3. For each $\delta > 0$, we consider the sesquilinear form $B_\delta(w,v)$ defined as in (5.3) but with ϕ_δ instead of ε. Then

$$B_\delta(w,v) = B_\varepsilon(w,v) - \int_\Omega \left(\frac{1}{\varepsilon}-\frac{1}{\phi_\delta}\right)\nabla w\cdot\nabla\overline{v} \tag{5.47}$$

for all $w \in H^1_{qp}(\Omega)$ and $v \in H^1_{qp}(\Omega)$. We also see that

$$B_\varepsilon(u_\xi,v) = F(v) - B_\varepsilon(u-u_\xi,v) \tag{5.48}$$

for all $v \in H^1_{qp}(\Omega)$. Combining the last two equalities with $w = u_\xi$, we have

$$B_\delta(u_\xi,v) = F(v) - B_\varepsilon(u-u_\xi,v) - \int_\Omega \left(\frac{1}{\varepsilon}-\frac{1}{\phi_\delta}\right)\nabla u_\xi\cdot\nabla\overline{v} \tag{5.49}$$

for all $v \in H^1_{qp}(\Omega)$.

Let u and u_ξ be given, $u' \in H^1_{qp}(\Omega)$ be the solution of the variational problem

$$B_\delta(u', v) = F(v) \tag{5.50}$$

for all $v \in H^1_{qp}(\Omega)$, and $u'' \in H^1_{qp}(\Omega)$ be the solution of the variational problem

$$B_\delta(u'', v) = -B_\varepsilon(u - u_\xi, v) - \int_\Omega \left(\frac{1}{\varepsilon} - \frac{1}{\phi_\delta}\right) \nabla u_\xi \cdot \nabla \overline{v} \tag{5.51}$$

for all $v \in H^1_{qp}(\Omega)$. It follows from Lemma 5.4.1 that the solutions u' and u'' exist since the right sides in the two variational problems are in the dual space $(H^1_{qp}(\Omega))'$. We can choose a $\delta > 0$ small enough so that

$$\begin{aligned}
\|u''\|_{H^1(\Omega)} &\le C_1(\kappa, \phi_\delta) \sup_{0 \neq v \in H^1_{qp}(\Omega)} \frac{\left| B_\epsilon(u - u_\xi, v) + \int_\Omega \left(\frac{1}{\varepsilon} - \frac{1}{\phi_\delta}\right) \nabla u_\xi \cdot \nabla \overline{v} \right|}{\|v\|_{H^1(\Omega)}} \\
&\le C_1(\kappa, \phi_\delta) \left[\gamma \|u - u_\xi\|_{H^1(\Omega)} + \left\| \left(\frac{1}{\varepsilon} - \frac{1}{\phi_\delta}\right) \nabla u_\xi \right\|_{L^2(\Omega)} \right] \\
&\le C_1(\kappa, \phi_\delta) \left[\gamma \xi + \|\nabla u_\xi\|_{L^\infty(\Omega)} \left(\frac{1}{\varepsilon_{\min}}\right)^2 \|\varepsilon - \phi_\delta\|_{L^2(\Omega)} \right] \\
&\le C_1(\kappa, \phi_\delta)(\gamma + 1)\xi,
\end{aligned} \tag{5.52}$$

where γ is the continuity constant of $B_\varepsilon(\cdot, \cdot)$. Finally we have

$$\begin{aligned}
\|u\|_{H^1(\Omega)} &\le \|u - u_\xi\|_{H^1(\Omega)} + \|u_\xi\|_{H^1(\Omega)} \\
&\le \xi + \|u'\|_{H^1(\Omega)} + \|u''\|_{H^1(\Omega)} \\
&\le \xi + C_1(\kappa, \phi_\delta) \left(\|F\|_{\left(H^1_{qp}(\Omega)\right)'} + (\gamma + 1)\xi \right)
\end{aligned} \tag{5.53}$$

for all $\xi > 0$. To complete the proof, we recall that the ϕ_δ are uniformly bounded and that $C_1(\kappa, \phi_\delta) \le C_1(\kappa, \varepsilon)$ follows from the definition of $C_1(\kappa, \varepsilon)$ for all $\delta > 0$.

□

Remark 9. *For ε piecewise C^2 in $\mathbb{R}^2$ satisfying the non-trapping condition of Theorem 5.4.3, it follows that ε_h is piecewise C^2 in $\mathbb{R}^2$ and also satisfies the non-trapping conditions. Thus, the associated problem (5.4) has a unique u^h as its solution. This*

follows because the same a-priori estimate holds for the problems (5.3) and (5.4), and the two continuity constants are $C_1(\kappa,\varepsilon)$ and $C_1(\kappa,\varepsilon_h)$. The non-trapping conditions allow us to write $C_1(\kappa,\varepsilon)$ explicitly in terms of κ and ε. This is important because $C_1(\kappa,\varepsilon_h) \leq C_1(\kappa,\varepsilon)$ for all $h>0$, which follows from this explicit dependence. Therefore, for the solution of the problem (5.4), we have an a-priori estimate where the continuity constant is independent of h.

5.5 *A-priori* Bounds on the Solution

To prove convergence we need two additional *a-priori* bounds on the solution.

Theorem 5.5.1. *Let ε be piecewise C^2 in $\mathbb{R}^2$, and satisfy the non-trapping condition of Theorem 5.4.3. Let $u_F \in H^1_{qp}(\Omega)$ denote the solution of problem (5.3) with $F \in L^2_{qp}(\Omega)$ on the right hand side. Then there is a constant $C_2(\kappa,\varepsilon)>0$ and an index s_1, such that*

$$\|u_F\|_{H^{1+s_1}(\Omega)} \leq C_2(\kappa,\varepsilon)\,\|F\|_{L^2(\Omega)}\,, \tag{5.54}$$

where $s_1 \in (0,1/2)$.

Proof. We extend the domain Ω by ℓ periods on the left and right, and then above and below by including the infinite half-spaces where $x_2 > H$ and $x_2 < -H$. This extended domain is then defined as

$$\Omega^E = \{\boldsymbol{x} \in \mathbb{R}^2,\ -\ell L_x < x_1 < (\ell+1)L_x\}. \tag{5.55}$$

As it is also useful to define a circular restricted domain, we choose an $R>0$ such that

$$\Omega_R = \{\boldsymbol{x} \in \mathbb{R}^2, \left|\boldsymbol{x} - (L_x/2,0)\right| < R\} \tag{5.56}$$

satisfies the set inclusion $\Omega \subset \Omega_R \subset \Omega^E$. The right hand side F is extended to Ω_R by quasi-periodicity in x_1 and by zero above and below. We can also extend the solution u_F to the domain Ω^E by quasi-periodicity to the left and right in x_1, and using the Rayleigh–Bloch expansion (2.30) above and below, to obtain $u_F^E \in H^1_{qp}(\Omega^E)$.

Let χ be a smooth cut-off function such that $\chi = 1$ in Ω, $\chi = 0$ on $\partial\Omega_R$ and $|\nabla\chi| < 1$ in Ω_R. We consider $w = \chi u_F^E$, and notice immediately that $w = u$ in Ω and $w = 0$ on $\partial\Omega_R$. Then $w \in H^1(\Omega_R)$ solves the elliptic problem

$$\left.\begin{aligned} \nabla\cdot\left(\tfrac{1}{\varepsilon}\nabla w\right) &= F^* \qquad && \text{in } \Omega_R \\ w &= 0 && \text{on } \partial\Omega_R \end{aligned}\right\}, \tag{5.57}$$

where the source function $F^* \in H^{s-1}(\Omega_R)$ for all $s \in [0, 1/2)$. To show that this is true, we let $\xi \in H^1(\Omega_R)$ and consider

$$\begin{aligned} \int_{\Omega_R} \nabla\cdot\left(\frac{1}{\varepsilon}\nabla w\right)\overline{\xi} = &-\int_{\Omega_R}\frac{1}{\varepsilon}u_F^E\nabla\chi\cdot\nabla\overline{\xi} - \int_{\Omega_R}\frac{1}{\varepsilon}\nabla u_F^E\cdot\nabla(\chi\overline{\xi}) \\ &+\int_{\Omega_R}\frac{1}{\varepsilon}\nabla u_F^E\cdot\nabla(\chi)\overline{\xi}, \end{aligned} \tag{5.58}$$

which can be obtained by the divergence theorem and adding and subtracting terms. The first term on the right side of (5.58) can be rewritten for all $\xi \in H^1(\Omega_R)$ as follows:

$$-\int_{\Omega_R}\frac{1}{\varepsilon}u_F^E\nabla\chi\cdot\nabla\overline{\xi} = \int_{\Omega_R}\nabla\cdot\left(\frac{1}{\varepsilon}u_F^E\nabla\chi\right)\overline{\xi}. \tag{5.59}$$

Since $\nabla\cdot\left(\frac{1}{\varepsilon}\nabla u_F^E\right) = F^E - \kappa^2 u_F^E$ in Ω_R, we find using the Divergence theorem that the second term on the right hand side of (5.58) may be rewritten as

$$-\int_{\Omega_R}\frac{1}{\varepsilon}\nabla u_F^E\cdot\nabla(\chi\overline{\xi}) = \int_{\Omega_R}F^E\overline{\xi}\chi - \int_{\Omega_R}\kappa^2 u_F^E\overline{\xi}\chi, \tag{5.60}$$

for all $\xi \in H^1(\Omega_R)$. The boundary integrals on $\partial\Omega_R$ cancel because $\chi\overline{\xi} = 0$ on $\partial\Omega_R$. Hence, (5.58) simplifies to

$$\int_{\Omega_R}\nabla\cdot\left(\frac{1}{\varepsilon}\nabla w\right)\overline{\xi} = \int_{\Omega_R}\left[\nabla\cdot\left(\frac{1}{\varepsilon}u_F^E\nabla\chi\right) + F^E\chi - \kappa^2 u_F^E\chi + \frac{1}{\varepsilon}\nabla u_F^E\cdot\nabla\chi\right]\overline{\xi} \tag{5.61}$$

for all $\xi \in H^1(\Omega_R)$. Then F^* is given by the terms enclosed in the square bracket on the right side of (5.61), and we can write $F^* = \nabla\cdot\left(\frac{1}{\varepsilon}u_F^E\nabla\chi\right) + \hat{F}$ where $\hat{F} \in L^2(\Omega_R)$. Since ε^{-1} is piecewise C^2, ε^{-1} satisfies the conditions of Proposition 2.1 of Ref. [48], i.e., $\varepsilon^{-1} \in L^\infty(\Omega_R)$ and $\nabla\varepsilon^{-1}$ is piecewise $L^\infty(\Omega_R)$. Since the product $u_F^E\nabla\chi \in H^s(\Omega_R)$, $\nabla\cdot\left(\frac{1}{\varepsilon}u_F^E\nabla\chi\right) \in H^{s-1}(\Omega_R)$ for every $s \in [0, 1/2)$. This shows that w solves the elliptic

problem given in (5.57), and so we apply Proposition 2.2 of Ref. [41] to this problem. It follows that there is a constant $c > 0$ and an index $s_1 \in (0, 1/2)$ such that

$$\begin{aligned}\|w\|_{H^{1+s_1}(\Omega_R)} &\leq c\,\|F^*\|_{H^{s_1-1}(\Omega_R)} \\ &\leq c\Big(\Big\| \frac{1}{\varepsilon} u_F^E \nabla\chi \Big\|_{L^2(\Omega_R)} + \Big\| \hat{F} \Big\|_{L^2(\Omega_R)} \Big) \\ &\leq c(2\ell+1)\Big[1 + C_1(\kappa,\varepsilon)(1+C(\kappa))(\kappa^2 + 2\left\|\varepsilon^{-1}\right\|_{L^\infty(\Omega)}) \Big] \|F\|_{L^2(\Omega)}\,.\end{aligned}$$

This follows by repeated use the *a-priori* estimate for u_F and Theorem 3 of Ref. [36]. To complete the proof, we note that $\|u_F\|_{H^{1+s_1}(\Omega)} \leq \|w\|_{H^{1+s_1}(\Omega_R)}$. □

Using the previous theorem we have the following regularity result for u.

Corollary 5.5.1.1. *Suppose ε is piecewise C^2 in $\mathbb{R}^2$ and satisfies the non-trapping condition of Theorem 5.4.3. Let $u \in H^1_{qp}(\Omega)$ be the solution to problem (5.3) where the right hand side is $f = \nabla\cdot\left[(\varepsilon_+^{-1} - \varepsilon^{-1})\nabla u^i\right]$. Then*

$$\|u\|_{H^{1+s_1}(\Omega)} \leq C_2(\kappa,\varepsilon)\left\|(\varepsilon_+^{-1}\Delta + \kappa^2)(\chi u^i)\right\|_{L^2(\Omega_\delta)} + \left\|u^i\right\|_{H^{1+s_1}(\Omega_\delta)}, \tag{5.62}$$

for some $s_1 \in (0,1/2)$, an appropriately chosen domain Ω_δ, and a smooth cut-off function χ.

Remark 10. *As $f = \nabla\cdot\left[(\varepsilon_+^{-1} - \varepsilon^{-1})\nabla u^i\right] \in (H^1_{qp}(\Omega))'$ is singular at first sight, we might only expect $u \in H^1_{qp}(\Omega)$. However, the special form of the solution allows for some extra regularity.*

Proof. By virtue of Theorem 5.4.3, we know that $u \in H^1(\Omega)$ exists and is unique. Given a $\delta > 0$, we define an extended domain

$$\Omega_\delta = \{\boldsymbol{x} \in \mathbb{R}^2,\ 0 < x_1 < L_x, -H-\delta < x_2 < H+\delta\}. \tag{5.63}$$

The top and bottom boundaries of Ω_δ are $\Gamma_{H+\delta}$ and $\Gamma_{-H-\delta}$, respectively. The smooth cut-off function χ is defined so that $\chi = 1$ in Ω and $\chi = 0$ on $\Gamma_{H+\delta}$ and $\Gamma_{-H-\delta}$. We

recall that the total field $u^t = u + u^{\text{inc}}$ and define $\tilde{w} = \chi u^t + (1-\chi)u$. Now by definition, $\tilde{w} = u^t$ in Ω, and

$$\nabla \cdot \left(\frac{1}{\varepsilon}\nabla \tilde{w}\right) + \kappa^2 \tilde{w} = \left(\varepsilon_+^{-1}\Delta + \kappa^2\right)(\chi u^{\text{inc}}) \tag{5.64}$$

in Ω_δ. As the right side of (5.64) is in $L^2(\Omega_\delta)$, we apply Theorem 5.5.1 to find a constant $C_2(\kappa, \varepsilon) > 0$ such that

$$\left\|u^t\right\|_{H^{1+s_1}(\Omega)} \leq \|\tilde{w}\|_{H^{1+s_1}(\Omega_\delta)} \leq C_2(\kappa,\varepsilon)\left\|\left(\varepsilon_+^{-1}\Delta + \kappa^2\right)(\chi u^{\text{inc}})\right\|_{L^2(\Omega_\delta)}. \tag{5.65}$$

The proof follows by the triangle inequality. □

5.6 Convergence of RCWA in h

We can now prove convergence of the solution u^h of the perturbed problem (see (5.4)):

Theorem 5.6.1. *Let ε be piecewise C^2 in $\mathbb{R}^2$ be real with $\Re(\varepsilon) > 0$ and satisfy the non-trapping condition of Theorem 5.4.3. Suppose that the interfaces are the graphs of piecewise C^2 functions. Let $u \in H^1_{qp}(\Omega)$ be the solution of problem (5.3) with $f = \nabla \cdot \left[(\varepsilon_+^{-1} - \varepsilon^{-1})\nabla u^i\right]$ on the right side. Also let $u^h \in H^1_{qp}(\Omega)$ be the solution of problem (5.4) with $f = \nabla \cdot \left[(\varepsilon_+^{-1} - \varepsilon_h^{-1})\nabla u^i\right]$ on the right side. Then there is a constant $C > 0$ independent of $h > 0$ such that*

$$\left\|u - u^h\right\|_{H^1(\Omega)} \leq Ch^{s_1/2}, \tag{5.66}$$

where $s_1 \in (0,1)$ is related to the regularity of u, i.e., $u \in H^{1+s_1}(\Omega)$.

Proof. Since ε is periodic and piecewise C^2 in $\mathbb{R}^2$, both ε and ε_h are periodic and are in $L^\infty(\mathbb{R}^2)$. By the definition of ε_h, we can write $\varepsilon_h(\boldsymbol{x}) = (\varepsilon_h)_{\min} + \Pi(\boldsymbol{x})$ where $\Pi(\boldsymbol{x}) = \varepsilon(x_1, h_{j-\frac{1}{2}}) - (\varepsilon_h)_{\min}$ in slice S_j. We also write $\varepsilon(\boldsymbol{x}) = \varepsilon_{\min} + \tilde{\Pi}(\boldsymbol{x})$. Then given a $\tau \geq 0$ there is an integer $n \geq 0$ such that

$$\begin{aligned}
\Pi(\boldsymbol{x} + \tau\boldsymbol{e}_2) - \Pi(\boldsymbol{x}) &= \varepsilon(x_1, h_{j-\frac{1}{2}} + nh) - \varepsilon(x_1, h_{j-\frac{1}{2}}) \\
&= \tilde{\Pi}(\boldsymbol{x}_h + nh\boldsymbol{e}_2) - \tilde{\Pi}(\boldsymbol{x}_h) \\
&\geq \inf_{\boldsymbol{x}\in\Omega}\left[\tilde{\Pi}(\boldsymbol{x} + nh\boldsymbol{e}_2) - \tilde{\Pi}(\boldsymbol{x})\right].
\end{aligned} \tag{5.67}$$

Since ε is monotonically increasing in the x_2-direction, it follows that

$$\inf_{\boldsymbol{x}\in\Omega}\Big[\Pi(\boldsymbol{x}+\tau\boldsymbol{e}_2)-\Pi(\boldsymbol{x})\Big]\geq 0. \tag{5.68}$$

Thus, ε_h satisfies the non-trapping conditions of Theorem 5.4.3. We notice that

$$B_{\varepsilon_h}(u^t-u^{h,t},v)=\int_\Omega\left(\frac{1}{\varepsilon_h}-\frac{1}{\varepsilon}\right)\nabla u^t\cdot\nabla\overline{v} \tag{5.69}$$

for all $v\in H^1_{qp}(\Omega)$, where $u^{h,t}=u^h+u^i$. As the right side is in the dual space $(H^1_{qp}(\Omega))'$, we apply Theorem 5.4.3 to (5.69) to see that

$$\begin{aligned}\left\|u^t-u^{h,t}\right\|_{H^1(\Omega)}&\leq C_1(\kappa,\varepsilon_h)\sup_{0\neq v\in H^1_{qp}(\Omega)}\frac{\left|\int_\Omega(\varepsilon_h^{-1}-\varepsilon^{-1})\nabla u^t\cdot\nabla\overline{v}\right|}{\|v\|_{H^1(\Omega)}}\\&\leq C_1(\kappa,\varepsilon_h)\left\|\left(\frac{1}{\varepsilon_h}-\frac{1}{\varepsilon}\right)\nabla u^t\right\|_{L^2(\Omega)}\\&\leq C_1(\kappa,\varepsilon_h)\left\|(\varepsilon_h\varepsilon)^{-1}\right\|_{L^\infty(\Omega)}\|\varepsilon-\varepsilon_h\|_{L^{2p}(\Omega)}\|\nabla u^t\|_{L^{2q}(\Omega)},\end{aligned} \tag{5.70}$$

where $(1/p)+(1/q)=1$, which follows from Hölder's inequality. Using the Sobolev embedding theorem (cf. Theorem 6 of Ref. [40, Sec. 5.6.3]), we choose

$$\frac{1}{2q}=\frac{1}{2}-\frac{s_1}{2}, \tag{5.71}$$

which implies that there is a constant $C>0$ independent of $h>0$ such that

$$\|\nabla u^t\|_{L^{2q}(\Omega)}\leq C\left\|u^t\right\|_{H^{1+s_1}(\Omega)}. \tag{5.72}$$

Then $2p=2/s_1$, $2q=2/(1-s_1)$, and

$$\left\|u^t-u^{h,t}\right\|_{H^1(\Omega)}\leq CC_1(\kappa,\varepsilon_h)\left(\frac{1}{\min_\Omega\varepsilon}\right)^2\|\varepsilon-\varepsilon_h\|_{L^{2/s_1}(\Omega)}\left\|u^t\right\|_{H^{1+s_1}(\Omega)}. \tag{5.73}$$

Using Lemma 6 of Ref. [36], we know that there is a constant $c>0$ independent of h such that

$$\|\varepsilon-\varepsilon_h\|_{L^{2/s_1}(\Omega)}\leq ch^{s_1/2}. \tag{5.74}$$

To complete the proof, we recall that $C_1(\kappa,\varepsilon_h)\leq C_1(\kappa,\varepsilon)$ for all $h>0$, and $u^t-u^{h,t}=u-u^h$.

□

5.7 Convergence of RCWA in M

To show convergence with increasing number $2M+1$ of retained Fourier modes, we first consider an associated adjoint problem. To this end, for $F \in L^2_{qp}(\Omega)$ we seek a $z_F^h \in H^1_{qp}(\Omega)$ such that

$$\overline{B_{\varepsilon_h}(\xi, z_F^h)} = -\int_\Omega F\bar{\xi} \tag{5.75}$$

for all $\xi \in H^1_{qp}(\Omega)$. For ε piecewise in C^2, this problem has a unique solution in $H^1_{qp}(\Omega)$ because of Theorem 5.4.3. The Galerkin orthogonality

$$B_{\varepsilon_h}(u^h - u^{h,M}, v_M) = 0 \tag{5.76}$$

holds for all $v_M \in V_M$ holds because u^h solves problem (5.4) and the RCWA is a Galerkin method. Taking $\xi = u^h - u^{h,M}$, we have

$$\begin{aligned}\left\|u^h - u^{h,M}\right\|_{L^2(\Omega)} &\leq \gamma \left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} \sup_{F\in L^2_{qp}(\Omega)} \left(\frac{1}{\|F\|_{L^2(\Omega)}} \inf_{v_M \in V_M} \left\|z_F^h - v_M\right\|_{H^1(\Omega)}\right) \\ &\leq C_2(\kappa,\varepsilon)\gamma \left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} M^{-s_2}. \end{aligned} \tag{5.77}$$

This follows because $\left\|z_F^h - \mathcal{F}_M z_F^h\right\|_{H^1(\Omega)} \leq M^{-s_2}\left\|z_F^h\right\|_{H^{1+s_2}(\Omega)}$, and by virtue of Theorem 5.5.1.

Theorem 5.7.1. *Suppose that ε is piecewise C^2 in $\mathbb{R}^2$, real with $\Re(\varepsilon) > 0$ and satisfies the non-trapping condition of Theorem 5.4.3. Let $u^h \in H^1_{qp}(\Omega)$ be the solution to problem (5.4) with $F = \nabla\cdot\left[(\varepsilon_+^{-1} - \varepsilon_h^{-1})\nabla u^i\right]$ on the right hand side, and $u^{h,M,t}$ be the RCWA solution. Then there is a constant $C > 0$ independent of h and M such that*

$$\left\|u^{h,t} - u^{h,M,t}\right\|_s \leq C M^{(s-2)s_2}, \tag{5.78}$$

where $s \in \{0,1\}$, $s_2 \in (0,1/2)$ is related to the regularity of u^h and M is large enough.

Proof. After using Galerkin orthogonality again, it follows that

$$B_{\varepsilon_h}(u^h - u^{h,M}, u^h - u^{h,M}) = B_{\varepsilon_h}(u^h - u^{h,M}, u^h - \mathcal{F}_M u^h). \tag{5.79}$$

We also have that the sesquilinear form $B_{\varepsilon_h}(\cdot,\cdot)$ satisfies the Gärding inequality

$$\left|B_{\varepsilon_h}(w,w)\right| \geq c\|w\|^2_{H^1(\Omega)} - (\kappa^2+1)\|w\|^2_{L^2(\Omega)} \tag{5.80}$$

for all $w \in H^1_{qp}(\Omega)$, where $c = \min_\Omega \left(\frac{\Re(\varepsilon)}{|\varepsilon|^2}, 1\right)$.

We use an argument of Schatz (see [18]) and standard properties of Fourier series to obtain

$$\left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} \leq c^{-1}\left(\gamma M^{-s_2} + (\kappa^2+1)\left\|u^h - u^{h,M}\right\|_{L^2(\Omega)}\right). \tag{5.81}$$

Now by taking $M \geq \left(2(\kappa^2+1)\gamma C_2(\kappa,\varepsilon)c^{-1}\right)^{1/s_2}$ in (5.77), and combining the result with (5.81), we have that

$$\left\|u^h - u^{h,M}\right\|_{H^1(\Omega)} \leq 2c^{-1}\gamma M^{-s_2}. \tag{5.82}$$

This completes the proof. □

5.8 Convergence of the RCWA Method in Dissipative Media

So far we have only discussed the case where $\varepsilon > 0$ is real in Ω. This case is the more mathematically interesting one, because we had to use a Rellich identity along with density arguments to demonstrate convergence of the RCWA.

Now we consider Case II of Section 5.2: ε is complex in Ω with $\Im(\varepsilon) > c_1 > 0$ and $\Re(\varepsilon) > c_2 > 0$. Then, a part of electromagnetic energy incident on the domain during a finite interval of time is absorbed inside the domain. This case is of interest for optical modeling of solar cells [31, 33, 34, 36] and absorbing gratings [42, 43, 44].

Again, we let ε be piecewise C^2 in $\mathbb{R}^2$. The case of ε complex is easier than the purely real case because $\Im(\varepsilon) > c_1 > 0$ and $\Re(\varepsilon) > c_2 > 0$ ensures that the sesquilinear form $B_\varepsilon(\cdot,\cdot)$ defined in (5.3) is coercive. Therefore, we can use the Strang lemmas [37] to prove convergence. To show that $B_\varepsilon(\cdot,\cdot)$ is coercive in this case, we first prove a lemma.

Lemma 5.8.1. *Let $w \in H^1_{qp}(\Omega)$. Then there is a constant $C > 0$ such that*

$$\|w\|^2_{L^2(\Omega)} \leq C\left(|w|^2_{H^1(\Omega)} + \Im\int_{\Gamma_H} \overline{w}T^+_p(w)\right) \tag{5.83}$$

where $|\cdot|_{H^1(\Omega)}$ is the H^1 seminorm.

Proof. By contradiction, suppose there is a sequence $\{w_n\}_{n=1}^{\infty}$ in $H^1_{qp}(\Omega)$ such that

1. $\|w_n\|_{L^2(\Omega)} = 1$, and
2. $\|w_n\|_{L^2(\Omega)} > n\Big(|w_n|^2_{H^1(\Omega)} + \Im \int_{\Gamma_H} \overline{w_n} T_p^+(w)\Big)$.

As the imaginary part of the Dirichlet-to-Neumann boundary integral is nonnegative according to $(2.83)_2$, we see by using the two aforementioned properties of the sequence $\{w_n\}_{n=1}^{\infty}$ that

$$\frac{1}{n} > |w_n|^2_{H^1(\Omega)} \quad \forall\, n \geq 1. \tag{5.84}$$

Then the sequence is bounded in the $H^1(\Omega)$ norm. Furthermore, there is a subsequence $\{w_{n(j)}\}_{n=1}^{\infty}$ that converges weakly in H^1 to some $q \in H^1(\Omega)$ but strongly to q in $L^2(\Omega)$. It follows that

$$\left\|w_{n(j)}\right\|^2_{L^2(\Omega)} + \left\|\nabla w_{n(j)}\right\|^2_{L^2(\Omega)} \to \|q\|^2_{L^2(\Omega)}\,, \tag{5.85}$$

so then $\|q\|_{L^2(\Omega)} = 1$. Since strong convergence implies weak convergence (in particular in L^2), it follows that $(\nabla w_{n(j)}, \nabla p)_{L^2(\Omega)} \to (\nabla q, \nabla p)_{L^2(\Omega)}$ for all $p \in H^1(\Omega)$. But $\lim_{j\to\infty} \nabla w_{n(j)} = 0$, and so $(\nabla q, \nabla p)_{L^2(\Omega)} = 0$ for all $p \in H^1(\Omega)$, and $\|\nabla q\|_{L^2(\Omega)} = 0$. Thus, q is constant in Ω and also $q \in H^1_{qp}(\Omega)$ since it is a closed subspace of $H^1(\Omega)$. Then q is a quasi-periodic constant function and $q = 0$, if the incidence angle $\theta \neq 0$. Since this is a contradiction, we assume that $\theta = 0$. Then the only non-zero Fourier coefficient for q is the zeroth one.

By construction,

$$\frac{1}{n} > \Im \int_{\Gamma_H} \varepsilon_+ \overline{w_{n(j)}} T_p^+(w_{n(j)}) = \sum_{\alpha_k^2 < \kappa^2\varepsilon_+} \sqrt{\kappa^2\varepsilon_+ - \alpha_k^2}\left|w^{(k)}_{n(j)}(H)\right|^2 \tag{5.86}$$

and then it follows that $\lim_{j\to\infty} w^{(k)}_{n(j)}(H) = 0$ for all k such that $\alpha_k^2 < \kappa^2\varepsilon_+$. In particular, $\alpha_0^2 = 0 < \kappa^2\varepsilon_+$. By the trace theorem [22], we know that there is a constant $c > 0$ such that

$$\left\|w_{n(j)} - q\right\|^2_{L^2(\Gamma_H)} \leq c^2\big(\left\|w_{n(j)} - q\right\|^2_{L^2(\Omega)} + |w_{n(j)}|^2_{H^1(\Omega)}\big). \tag{5.87}$$

It follows from Bessel's inequality that

$$\sum_{k\in\mathbb{Z}} \left|w_{n(j)}^{(k)}(H) - q_k(H)\right|^2 \leq \left\|w_{n(j)} - q\right\|_{L^2(\Gamma_H)}^2 \overset{j\to\infty}{\to} 0, \tag{5.88}$$

but then $q = 0$ on Γ_H, since the coefficient $q_0(H) = 0$. Thus $q = 0$ in Ω, contradicting that $\|q\|_{L^2(\Omega)} = 1$. □

Now we can prove the coercivity of $B_\varepsilon(\cdot,\cdot)$.

Corollary 5.8.1.1. *If $\Im(\varepsilon) > c_1 > 0$ and $\Re(\varepsilon) > c_2 > 0$ for some positive constants c_1 and c_2 in Ω, then the sesquilinear form $B_\varepsilon(\cdot,\cdot)$ is coercive in $H^1(\Omega)$.*

Proof. It suffices to show that $\Im B_\varepsilon(\cdot,\cdot)$ is coercive in $L^2(\Omega)$ and $\Re B_\varepsilon(\cdot,\cdot)$ is coercive in $H^1(\Omega)$ in the Gärding sense. Recalling that $\Im(1/\varepsilon) = -\Im(\varepsilon)/|\varepsilon|^2$ and the signs of the imaginary parts of the Dirichlet-to-Neumann boundary integral are known per (2.83), we have

$$\begin{aligned}\left|\Im B_\varepsilon(w,w)\right| &\geq \min_\Omega\left(\frac{\Im(\varepsilon)}{|\varepsilon|^2},1\right)\left(|w|_{H^1(\Omega)}^2 + \Im\int_{\Gamma_H}\overline{w}T_p^+(w)\right)\\ &\geq C\min_\Omega\left(\frac{\Im(\varepsilon)}{|\varepsilon|^2},1\right)\|w\|_{L^2(\Omega)}^2\end{aligned} \tag{5.89}$$

for all $w \in H^1_{qp}(\Omega)$ per Lemma 5.8.1. To see that $|\Re B_\varepsilon(\cdot,\cdot)|$ is coercive in $H^1(\Omega)$ in the Gärding sense [63], we recall that $\Re(1/\varepsilon) = \Re(\epsilon)/|\varepsilon|^2$, and

$$\left|\Re B_\varepsilon(w,w)\right| \geq \min_\Omega\left(\frac{\Re(\varepsilon)}{|\varepsilon|^2},1\right)\|w\|_{H^1(\Omega)}^2 - (\kappa^2+1)\|w\|_{L^2(\Omega)}^2 \tag{5.90}$$

for all $w \in H^1_{qp}(\Omega)$. To show this implies coercivity, there are two cases to consider. We define $\lambda(\varepsilon) = C\min_\Omega\left(\frac{\Im(\varepsilon)}{|\varepsilon|^2},1\right)$, $\sigma(\varepsilon) = \min_\Omega\left(\frac{\Re(\varepsilon)}{|\varepsilon|^2},1\right)$ and $\gamma = \kappa^2+1$. If $\sigma\|w\|_{H^1(\Omega)}^2 - \gamma\|w\|_{L^2(\Omega)}^2 \leq 0$, then

$$\begin{aligned}\lambda\sigma\|w\|_{H^1(\Omega)}^2 &\leq \gamma\lambda\|w\|_{L^2(\Omega)}^2\\ &\leq \gamma\left|\Im B_\varepsilon(w,w)\right|,\end{aligned} \tag{5.91}$$

follows from (5.89). If $\sigma\|w\|^2_{H^1(\Omega)} - \gamma\,\|w\|^2_{L^2(\Omega)} > 0$, we let $c = \lambda^2/(2\gamma^2)$ so that $\lambda^2 - c\gamma^2 > 0$ and set $\tilde{c} = 2(c+1)$. By defining $a = \tilde{c}\gamma^2\,\|w\|^4_{L^2(\Omega)}$ and $b = 4(\tilde{c})^{-1}\sigma^2\,\|w\|^4_{H^1(\Omega)}$, it follows from the inequality of arithmetic and geometric means that

$$\begin{aligned}
\big|B_\varepsilon(w,w)\big|^2 &= \big|\Im B_\varepsilon(w,w)\big|^2 + \big|\Re B_\varepsilon(w,w)\big|^2 \\
&\geq \lambda^2\,\|w\|^4_{L^2(\Omega)} + \sigma^2\,\|w\|^4_{H^1(\Omega)} + \gamma^2\,\|w\|^4_{L^2(\Omega)} - 2\sigma\gamma\,\|w\|^2_{H^1(\Omega)}\,\|w\|^2_{L^2(\Omega)} \\
&\geq \big[\lambda^2 + \gamma^2(1-\tilde{c}/2)\big]\,\|w\|^4_{L^2(\Omega)} + \sigma^2\big[1 - 2(\tilde{c})^{-1}\big]\,\|w\|^4_{H^1(\Omega)}\,. \\
&= \frac{\lambda^2}{2}\,\|w\|^4_{L^2(\Omega)} + \sigma^2\big(1 - \frac{1}{c+1}\big)\,\|w\|^4_{H^1(\Omega)} \\
&\geq \frac{\sigma^2\lambda^2}{\lambda^2 + 2\gamma^2}\,\|w\|^4_{H^1(\Omega)}\,. \qquad (5.92)
\end{aligned}$$

It follows in either case that

$$\big|B_\varepsilon(w,w)\big| \geq \frac{\sigma(\varepsilon)}{\sqrt{1 + 2(\frac{\gamma}{\lambda(\varepsilon)})^2}}\,\|w\|^2_{H^1(\Omega)} \qquad (5.93)$$

for all $w \in H^1_{qp}(\Omega)$.

□

Existence and uniqueness of the variational problem (5.3) now follow by virtue of the Lax–Milgram Lemma [40]. Problem (5.4) has the same characteristics but ε_h is constructed by sampling the true relative permittivity at the inter-slice boundaries. Since $\Im(\varepsilon) > c_1 > 0$ and $\Re(\varepsilon) > c_2 > 0$, it follows that $\Im(\varepsilon_h) > c_1 > 0$ and $\Re(\varepsilon_h) > c_2 > 0$ for all $h > 0$, and we have existence and uniqueness for the ε_h problem as well. To obtain an *a-priori* estimate where the continuity constant is explicit, we note that

$$\big\|u^h\big\|_{H^1(\Omega)} \leq \frac{\sqrt{1 + 2(\frac{\gamma}{\lambda(\varepsilon)})^2}}{\sigma(\varepsilon)}\,\|f\|_{\left(H^1_{qp}(\Omega)\right)'}\,. \qquad (5.94)$$

follows from the Lax–Milgram Lemma, and the fact that $\lambda(\varepsilon_h)^{-1} \leq \lambda(\varepsilon)^{-1}$ and $\sigma(\varepsilon_h)^{-1} \leq \sigma(\varepsilon)^{-1}$ for all $h > 0$.

The foregoing discussion leads us to the following theorem.

Theorem 5.8.2. *Suppose that ε is piecewise C^2 in $\mathbb{R}^2$, $\Im(\varepsilon) > c_1 > 0$, $\Re(\varepsilon) > c_2 > 0$ in Ω, and the interfaces are graphs of piecewise C^2 functions. Then the two variational problems (5.3) and (5.4) have unique solutions u and $u^h \in H^1_{qp}(\Omega)$, respectively. Furthermore, there is a constant $C > 0$ independent of $h > 0$ such that*

$$\left\|u^t - u^{h,t}\right\|_{H^1(\Omega)} \leq Ch^{s_1/2}, \tag{5.95}$$

where $s_1 \in (0, 1/2)$ is related to the regularity of u, i.e., $u \in H^{1+s_1}(\Omega)$.

Proof. By virtue of (5.93), the family of sesquilinear forms $\left(B_{\varepsilon_h}(\cdot,\cdot)\right)_{h>0}$ is uniformly $H^1_{qp}(\Omega)$-elliptic. According to Strang's first lemma [22], there is a constant $c > 0$ independent of $h > 0$ such that

$$\begin{aligned}\left\|u^t - u^{h,t}\right\|_{H^1(\Omega)} &\leq c \inf_{w \in H^1_{qp}(\Omega)} \left(\left\|u^t - w\right\|_{H^1(\Omega)} + \sup_{v \in H^1_{qp}(\Omega)} \frac{\left|B_\varepsilon(w,v) - B_{\varepsilon_h}(w,v)\right|}{\|v\|_{H^1(\Omega)}} \right) \\ &\leq c \left(\sup_{v \in H^1_{qp}(\Omega)} \frac{\left|B_\varepsilon(u^t,v) - B_{\varepsilon_h}(u^t,v)\right|}{\|v\|_{H^1(\Omega)}} \right),\end{aligned} \tag{5.96}$$

where we put $u^t = w$. We then bound the consistency error

$$\left|B_\varepsilon(u^t,v) - B_{\varepsilon_h}(u^t,v)\right| \leq \left(\frac{1}{\min_\Omega |\varepsilon|}\right)^2 \|\varepsilon - \varepsilon_h\|_{L^{2p}(\Omega)} \|\nabla u^t\|_{L^{2q}(\Omega)} \|v\|_{H^1(\Omega)}, \tag{5.97}$$

where $(1/q) + (1/p) = 1$. The proof follows just like for Theorem 5.6.1. □

Theorem 5.8.3. *Suppose that ε is piecewise C^2 in $\mathbb{R}^2$, $\Im(\varepsilon) > c_1 > 0$, and $\Re(\varepsilon) > c_2 > 0$ in Ω. Let $u^h \in H^1_{qp}$ be the solution to problem (5.4) and $u^{h,M,t}$ be the RCWA solution. Then there exists a constant $C > 0$ independent of h and M such that*

$$\left\|u^{h,t} - u^{h,M,t}\right\|_s \leq CM^{(s-2)s_2}, \tag{5.98}$$

where $s \in \{0, 1\}$ and $s_2 \in (0, 1/2)$ is chosen so that $u^h \in H^{1+s_2}(\Omega)$.

Proof. By virtue of (5.93), the family of sesquilinear forms $\left(B_{\varepsilon_h}(\cdot,\cdot)\right)_{h>0}$ is uniformly V_M-elliptic. The sesquilinear form is the same for both problems, and if we consider

the total fields, the variational problems have the same right hand side. Strang's first lemma [22] yields a constant $c > 0$ independent of h and M such that

$$\begin{aligned}
\left\|u^{h,t} - u^{h,M,t}\right\|_{H^1(\Omega)} &\leq c\left(\inf_{v_M \in V_M} \left\|u^{h,t} - v_M\right\|_{H^1(\Omega)}\right) \\
&\leq c\left(\left\|u^{h,t} - \mathcal{F}_M u^{h,t}\right\|_{H^1(\Omega)}\right) \\
&\leq c\left(M^{-s_2}\left\|u^{h,t}\right\|_{H^{1+s_2}(\Omega)}\right) \\
&\leq cC_3(\kappa,\varepsilon)\left\|\left(\varepsilon_+^{-1}\Delta + \kappa^2\right)(\chi u^{\text{inc}})\right\|_{L^2(\Omega_\delta)} M^{-s_2}. \qquad (5.99)
\end{aligned}$$

Here, $C_3(\kappa,\varepsilon)$ is the same as $C_2(\kappa,\varepsilon)$ but with the continuity constant in (5.94) instead of $C_1(\kappa,\varepsilon)$. We have used standard properties of Fourier series [45] in the third line, and the last line follows from Corollary 5.5.1.1. The extra order of convergence in the L^2 norm follows from the Aubin–Nitsche trick [46].

□

We now summarize the convergence results of this chapter:

Theorem 5.8.4. *Suppose the conditions of Theorem 5.6.1 or Theorem 5.8.2 are met, $u \in H^1_{qp}(\Omega)$ is the solution to problem (5.3), and $u^{h,M,t}$ is the RCWA solution. For M large enough, there is a constant $C > 0$ independent of h and M such that*

$$\left\|u^t - u^{h,M,t}\right\|_s \leq C\left(h^{s_1/2} + M^{(s-2)s_2}\right), \qquad (5.100)$$

with $s \in \{0,1\}$, where s_1 is related to the regularity of u and s_2 is related to the regularity of u^h.

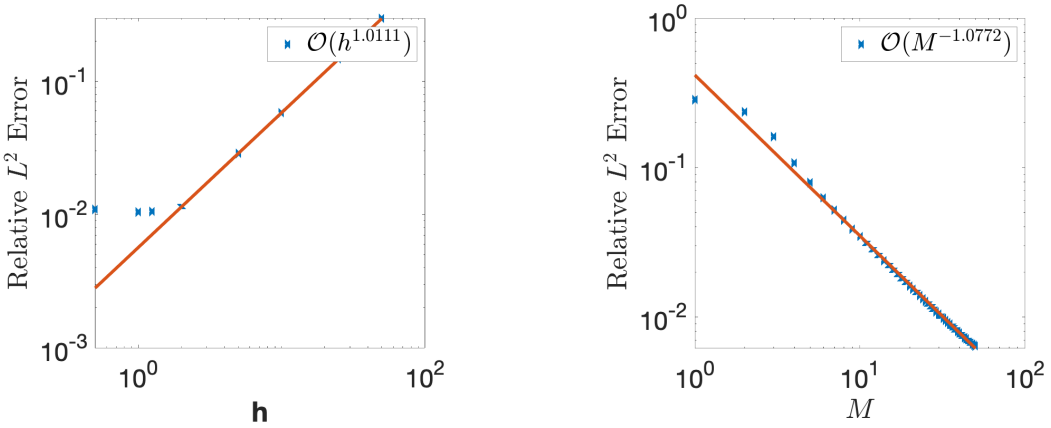
Proof. This follows from either Theorems 5.6.1 and 5.7.1 or Theorems 5.8.2 and 5.8.3, and the triangle inequality.

□

5.9 Numerical Examples

In this section, for two different gratings, we compute the convergence rate of the RCWA solution by comparing $u^{h,M,t}$ with the solution obtained using the FEM

with a highly refined grid, since an analytical solution is not possible to obtain. Both gratings have a period $L = 500$ nm and a triangular profile of base $L/2$. The first grating, shown in Fig. 4.2(a), has a symmetric triangular profile of height 100 nm and sits atop a 100-nm-thick strip. The second grating, shown in Fig. 4.2(b), has an asymmetric triangular profile of height 50 nm with its vertex shifted off-center by $L/8$ and sits atop a 50-nm-thick strip. The grating and the underlying strip are made of either a metal with relative permittivity $\varepsilon_m = -15 + 4i$ or a dissipative material with relative permittivity $\varepsilon_d = 15 + 4i$. The domain height $2H = 1700$ nm in Fig. 4.2(a) but $2H = 1600$ nm in Fig. 4.2(b). Whatever portion of Ω is not occupied by metal or dissipative material is filled with air with relative permittivity $\varepsilon_a = 1 + 10^{-6}i$. The small imaginary part $\Im(\varepsilon_a) = 10^{-6}$ was used for the sake of numerical stability of the RCWA algorithm [1]. Calculations were made for normal incidence (i.e., $\theta = 0$) at free-space wavenumber $\lambda_0 = 2\pi/\kappa = 600$ nm.



(a) (b)

Figure 5.1: Convergence plots comparing the relative L^2 error between the RCWA and FEM solutions for the symmetric grating of Fig. 4.2(a). Whereas $h \in \{1/2, 1, 1.25, 2, 5, 10, 25, 50\}$ nm but $M = 30$ in (a), $h = 1$ nm but $M = [1, 50]$ in (b). The grating and underlying strip are made of a dissipative material with relative permittivity $\varepsilon_d = 15 + 4i$. In all plots the error saturates below 10^{-2}. For the least-squares-fit lines, data in the convergent regime only where used.

The FEM solution u_{FE} was computed using an adaptive method in NGSolve

version 6.2.1908 [47]. The domain Ω was taken to be sandwiched between two perfectly matched layers (PMLs) of thickness equal to λ_0 and PML parameter equal to $1.5+2.5i$. The FEM solutions were computed using 4th-degree continuous finite elements, with the mesh adaptivity terminating when the algorithm would reach 100,000 degrees of freedom. The relative L^2 error between the RCWA and FEM solutions was calculated as

$$\frac{\left\|u^{h,M,t} - u_{FE}\right\|_{L^2(\Omega)}}{\|u_{FE}\|_{L^2(\Omega)}}. \tag{5.101}$$

The convergence results in Figs. 5.1(a) and 5.1(b) are for the symmetric case of Fig. 4.2(a) with $\varepsilon_d = 15 + 4i$. Since $\Re(\varepsilon) > 0$ and $\Im(\varepsilon) > 0$, this case is covered by Theorem 5.8.4. Again, as our analysis predicted the convergence rate with respect to M to be $\mathcal{O}(M^{-2s_2})$ with $s_2 \in (0, 1/2)$. In fact, we know that $u^h \in H^{1+s_2}$ for every $s_2 \in (0, 1/2)$ which follows from Corollary 3.2 of Ref. [49]. The numerical results yielding $\mathcal{O}(M^{-1.0772})$ match the prediction.

Theorem 5.8.4 predicts the convergence rate with respect to h to be $\mathcal{O}(h^{s_1/2})$. In practice, however, we see faster convergence than predicted by our analysis. This was also observed for s-polarized light [36]. It would be desirable to improve the predicted convergence rates with respect to h to closely match the higher rates seen in our numerical results. The RCWA converges in a stable way in this case, as the error data closely falls on the least-squares-fit line.

However, there is some numerical instability for the metallic grating and we were not able to analyze that case for p-polarized incident light.

Figures 5.2(a) and 5.2(c) show the convergence of the RCWA solution for the symmetric grating with respect to M and h, respectively, when the grating and the underlying strip are made of a metal with relative permittivty of $\varepsilon_m = -15 + 4i$. We observe order $\mathcal{O}(M^{-1.028})$ and $\mathcal{O}(h^{0.73068})$. Figures 5.2(b) and 5.2(d) show the convergences of the RCWA solution of the asymmetric grating as being order $\mathcal{O}(M^{-0.91695})$ and $\mathcal{O}(h^{0.92497})$, respectively. Although our analysis in the previous sections did not cover the case where $\Im(\varepsilon) < 0$, we included these numerical results because metal

gratings are a common application for RCWA.

5.10 Conclusion

The convergence properties of the two-dimensional RCWA for p- polarized light was studied for domains with piecewise smooth relative permittivity that is either

I. purely real and positive, or

II. complex with both real and imaginary parts positive.

Since the RCWA approximates the solution using slices of thickness h and a Fourier truncation parameter M, we provided theorems about the convergence rates with respect to both h and M. We showed that the RCWA is a Galerkin scheme for p-polarized light, which allowed investigation using FEM techniques. If the relative permittivity is purely real, we proved a Rellich identity for non-trapping domains. Our theory predicts convergence even for trapping domains, as long as the continuity constant (5.22) remains bounded independently of h. In the case where $\Im(\varepsilon) > c_1 > 0$ and $\Re(\varepsilon) > c_2 > 0$, we proved convergence using the Strang lemmas.

Our study has two limitations which need to be overcome in future work. We need to extend our analysis to the case where $\Re(\epsilon) < 0$ in some parts of the domain. This would allow us to assert convergence for metallic scatterers as seen in Figure 4.2. Secondly we need to allow for absorption in only some parts of the domain when $\Re(\epsilon) > 0$.

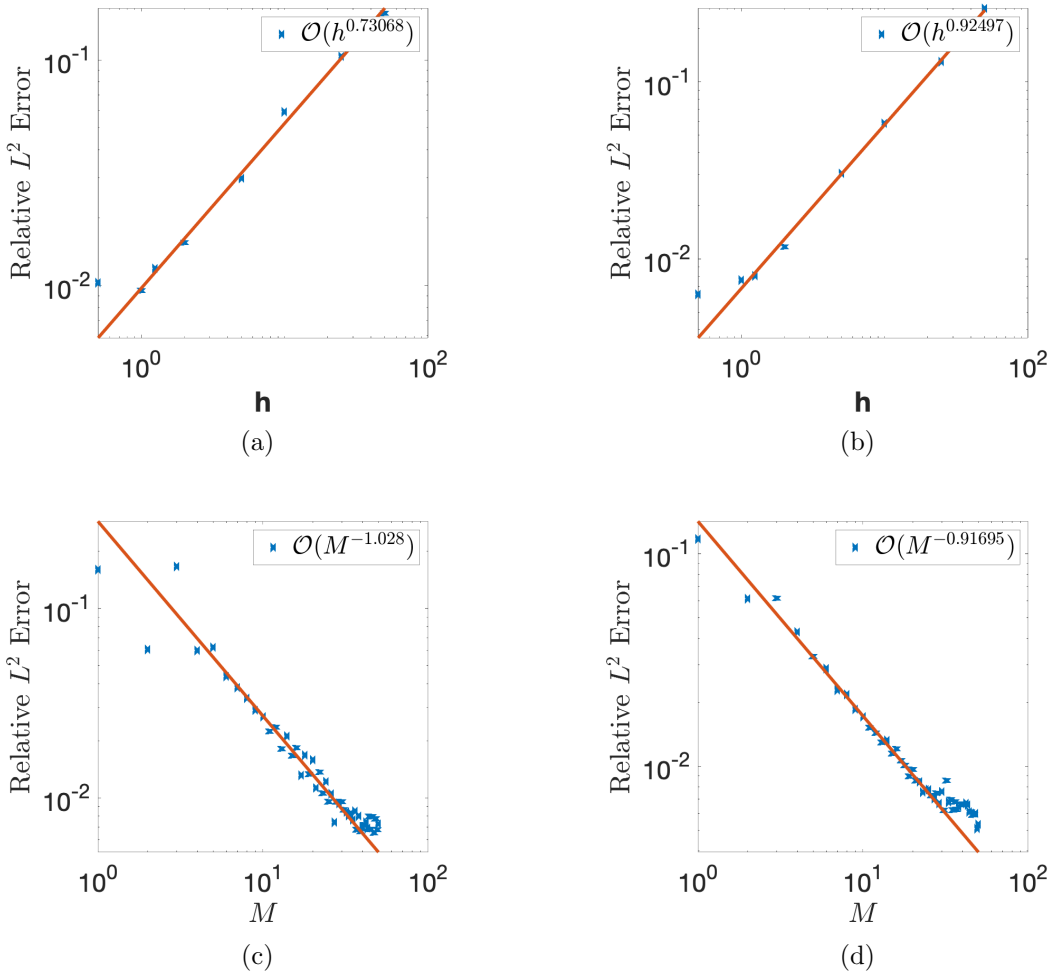


Figure 5.2: Convergence plots comparing the relative L^2 error between the RCWA and FEM solutions for (a,c) the symmetric grating of Fig. 4.2(a) and (b,d) the asymmetric grating of Fig. 4.2(b). Whereas $h \in \{1/2, 1, 1.25, 2, 5, 10, 25, 50\}$ nm but $M = 30$ in (a) and (b), $h = 1$ nm but $M = [1, 50]$ in (c) and (d). The grating and underlying strip are made of a metal with relative permittivity $\varepsilon_m = -15 + 4i$. In all plots the error saturates below 10^{-2}. For the least-squares-fit lines, data in the convergent regime only where used.

Chapter 6

THE C METHOD

6.1 Introduction

The C Method, first introduced by Chandezon [60] in 1980, is another numerical method to solve the scattering problem (2.72)–(2.74). Instead of approximating the grating interfaces, the C Method uses a coordinate transformation such that all the interfaces in the new coordinate system are flat. This is at the cost of spatially dependent anisotropic coefficients in the transformed system. In this chapter, we propose a hybrid method that combines the C Method with the RCWA method. When applying the RCWA algorithm to the transformed problem, all the interfaces will line up with the grid for the slices. The benefit of this method is that the interfaces are not approximated, whereas the standard RCWA replaces the interfaces with a stairstep.

After transformation, in the new coordinates $(\hat{x}_1, \hat{x}_2, \hat{x}_3)$ the geometry of the domain $\hat{\Omega}$ is simple compared to that of the original domain Ω. Under a coordinate transformation, Maxwell's equations are invariant up to the coefficients $\boldsymbol{\varepsilon}$ and $\boldsymbol{\mu}$. The scattering problem we have considered has scalar $\mu \equiv 1$ and $\varepsilon = \varepsilon(x_1, x_2)$, but under transformation the coefficients will both be anisotropic. Therefore, this hybrid C-RCWA method relies upon the RCWA for bianisotropic media where $\boldsymbol{\varepsilon}$ and $\boldsymbol{\mu}$ are chosen so that the solution solves the appropriate transformed Helmholtz equation. For simplicity, we derive the method for the case where there is only one interface Γ, assumed to be the graph of a C^1 function $g(x_1)$. However, our method can be applied to problems with any number of interfaces.

A known issue with the RCWA for p-polarized incident light is numerical instability of the method when $\Re(\varepsilon) < 0$ [29, 58, 5, 59]. This instability manifests as

spurious low-amplitude oscillations near the grating. Many methods, such as an improved formulation [5] and the normal vector method [59], exist to try and eliminate these oscillations in the near-field. The improved formulation by Li is not applicable to the full Maxwell system, and the normal vector method is difficult to apply to this problem. In this chapter, we offer an alternative method C-RCWA to solve this issue. The benefit of our method is that it should be relatively straightforward to apply to the full 3D scattering problem for crossed gratings. We offer some preliminary numerical results that show that our method does eliminate these oscillations.

This chapter is organized as follows. We introduce the coordinate transform method in Section 6.2, and state the transformed scattering problem for the p-polarization case. In Section 6.3, we derive our hybrid C-RCWA and in Section 6.4 we give some preliminary numerical results comparing the C-RCWA solution to the RCWA solution.

6.2 The Transformed Scattering Problem

In the case considered here, the domain $\Omega = [0, L] \times [-H, H]$ contains a single interface given by $x_2 = g(x_1)$. The mapped domain $\hat{\Omega} = [0, L] \times [-H, H]$ is bisected by a single flat interface. We define the family of coordinate transforms $G(\hat{\boldsymbol{x}})$ by

$$x_1 = \hat{x}_1, \tag{6.1}$$

$$x_2 = S(\hat{x}_2)g(\hat{x}_1) + \hat{x}_2, \tag{6.2}$$

that map the domain $\hat{\Omega}$ bijectively into Ω. An illustration of the coordinate transformation is given in Figure 6.1. In Chandezon's original method, $S \equiv 1$, but we consider a general $S(\hat{x}_2)$ for now. We want the inverse transformation $G^{-1}(\boldsymbol{x})$ to map the interface defined by $x_2 = g(x_1)$ onto $\hat{x}_2 = 0$, and therefore an essential property is that $S(0) = 1$. The rectangular domain Ω is bounded by the lines $x_2 = H$ and $x_2 = -H$. Thus, $S(H) = 0$ and $S(-H) = 0$ together ensure that the upper and lower boundaries are the same in both coordinate systems.

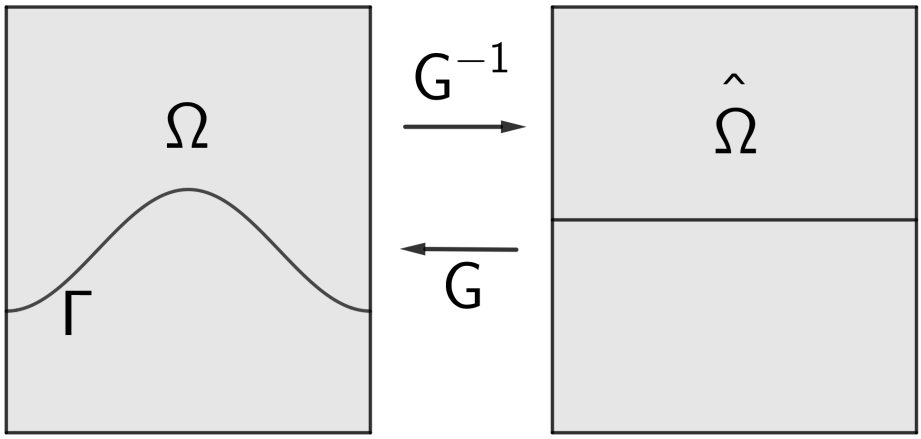


Figure 6.1: The coordinate transformation G maps the domain $\hat{\Omega}$ bijectively into Ω. The interface Γ is the graph of the C^1 function $g(x_1)$ and gets mapped to $\hat{x}_2 = 0$ by G^{-1}.

For the p-polarization case, the scattered field $\hat{u} = u \circ G$ solves the Helmholtz problem

$$\hat{\nabla} \cdot \left(\hat{\varepsilon}^{-1} A \hat{\nabla} \hat{u} \right) + \kappa^2 a \hat{u} = \hat{F} \qquad \text{in } \hat{\Omega}, \tag{6.3}$$

$$\exp(-i\alpha_0 L)\hat{u}(0, \hat{x}_2) = \hat{u}(L, \hat{x}_2) \qquad \forall\, \hat{x}_2, \tag{6.4}$$

$$\exp(-i\alpha_0 L)(A\hat{\nabla}\hat{u}) \cdot \hat{\nu}_L \Big|_{\hat{x}_1 = 0} = (A\hat{\nabla}\hat{u}) \cdot \hat{\nu}_R \Big|_{\hat{x}_1 = L} \qquad \forall\, \hat{x}_2. \tag{6.5}$$

The 2×2 symmetric matrix $A = |\det DG| DG^{-1} DG^{-T}$ and scalar $a = |\det DG|$, where DG is the Jacobian of the transformation G. In general,

$$A = \begin{pmatrix} |\det DG| & -\operatorname{sgn}(\det DG) S g' \\ -\operatorname{sgn}(\det DG) S g' & \frac{\left(Sg'\right)^2 + 1}{|\det DG|} \end{pmatrix}. \tag{6.6}$$

We have $A = I$ on Γ_H and Γ_{-H} because of our choice of $S(\hat{x}_2)$ and the additional requirement that $S'(H) = 0$ and $S'(-H) = 0$. Also, $\det DG = S'(\hat{x}_2) g(\hat{x}_1) + 1$ is unity on the boundaries Γ_H and Γ_{-H}.

Using the Divergence theorem in the usual way, we see that $\hat{u} \in H^1_{qp}(\hat{\Omega})$ solves the variational problem

$$\int_{\hat{\Omega}} \left(\hat{\varepsilon}^{-1} A \hat{\nabla} \hat{u} \cdot \hat{\nabla} \overline{\hat{v}} - \kappa^2 a \hat{u} \overline{\hat{v}} \right) - \int_{\Gamma_H} \overline{\hat{v}} T_p^+(\hat{u}) - \int_{\Gamma_{-H}} \overline{\hat{v}} T_p^-(\hat{u}) = - \int_{\hat{\Omega}} \hat{F} \overline{\hat{v}} \tag{6.7}$$

for all $\overline{\hat{v}} \in H^1_{qp}(\hat{\Omega})$.

6.3 The C-RCWA Method

In this section, we derive the hybrid C-RCWA method. For ease of reading, we drop the $\hat{}$ over the mapped variables. First, we consider the problem of having general bianisotropic coefficients of the form

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \tilde{\boldsymbol{\varepsilon}} & 0 \\ 0 & \varepsilon_{33} \end{pmatrix} \qquad\qquad \boldsymbol{\mu} = \begin{pmatrix} \tilde{\boldsymbol{\mu}} & 0 \\ 0 & \mu_{33} \end{pmatrix},$$

where $\tilde{\boldsymbol{\varepsilon}}$ and $\tilde{\boldsymbol{\mu}}$ are 2×2 matrices such that $\det \tilde{\boldsymbol{\varepsilon}}$ and $\det \tilde{\boldsymbol{\mu}}$ are constant. We also define

$$\tilde{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{pmatrix} \qquad\qquad \tilde{\boldsymbol{\mu}} = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix}.$$

With this choice of bianisotropic coefficients, we obtain the following system of PDEs:

$$\frac{\partial E_3}{\partial x_2} = i\omega\mu_0\big(\mu_{11}H_1 + \mu_{12}H_2\big), \tag{6.8}$$

$$-\frac{\partial E_3}{\partial x_1} = i\omega\mu_0\big(\mu_{21}H_1 + \mu_{22}H_2\big), \tag{6.9}$$

$$\frac{\partial E_2}{\partial x_1} - \frac{\partial E_1}{\partial x_2} = i\omega\mu_0\mu_{33}H_3, \tag{6.10}$$

$$\frac{\partial H_3}{\partial x_2} = -i\omega\varepsilon_0\big(\varepsilon_{11}E_1 + \varepsilon_{12}E_2\big), \tag{6.11}$$

$$\frac{\partial H_3}{\partial x_1} = i\omega\varepsilon_0\big(\varepsilon_{21}E_1 + \varepsilon_{22}E_2\big), \tag{6.12}$$

$$\frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2} = -i\omega\varepsilon_0\varepsilon_{33}\varepsilon E_3. \tag{6.13}$$

The s-polarization case is given by (6.8), (6.9) and (6.13), and the p-polarization case is given by (6.10)–(6.12). Each system is equivalent to solving a Helmholtz equation for E_3 and H_3, respectively. The general procedure to do this is as follows. For the first case, multiply (6.8) by μ_{21} and (6.9) by μ_{11} and add the resulting equations. We obtain

$$\mu_{21}\frac{\partial E_3}{\partial x_2} + \mu_{11}\frac{\partial E_3}{\partial x_1} = -i\omega\mu_0 \det(\tilde{\boldsymbol{\mu}})H_2. \tag{6.14}$$

On the other hand, we can eliminate H_2 in a similar way. We multiply (6.8) by μ_{22} and (6.9) by μ_{12} and add the result to obtain

$$\mu_{22}\frac{\partial E_3}{\partial x_2} + \mu_{12}\frac{\partial E_3}{\partial x_1} = i\omega\mu_0 \det(\tilde{\boldsymbol{\mu}})H_1. \tag{6.15}$$

The use of (6.14) and (6.15) in (6.13) shows that E_3 solves the Helmholtz equation

$$\nabla \cdot \left(\tilde{\boldsymbol{\mu}}^T \nabla E_3\right) + \kappa^2 \varepsilon_{33} \det(\tilde{\boldsymbol{\mu}}) E_3 = 0. \tag{6.16}$$

In a similar way, we can show that H_3 solves the Helmholtz equation

$$\nabla \cdot \left(\det(\tilde{\boldsymbol{\varepsilon}})^{-1}\tilde{\boldsymbol{\varepsilon}}^T \nabla H_3\right) + \kappa^2 \mu_{33} H_3 = 0. \tag{6.17}$$

In order to solve the correct transformed Helmholtz problems, we must choose

$$\boldsymbol{\varepsilon} = \hat{\varepsilon}\begin{pmatrix} A & 0 \\ 0 & |\det DG| \end{pmatrix}, \tag{6.18}$$

$$\boldsymbol{\mu} = (\hat{\varepsilon})^{-1}\boldsymbol{\varepsilon}. \tag{6.19}$$

For example, we set $\tilde{\boldsymbol{\varepsilon}} = \hat{\varepsilon}A$. Since $\det A = 1$ and A is symmetric, it is easy to see that $\det(\tilde{\boldsymbol{\varepsilon}})^{-1}\tilde{\boldsymbol{\varepsilon}}^T = \hat{\varepsilon}^{-1}A$. This choice of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\mu}$ is equivalent to solving the transformed Helmholtz equation (6.3).

We now express the system of PDEs (6.8)–(6.13) in Fourier space, where $\boldsymbol{X}_\sigma$ is the vector of Fourier coefficients of X_σ where $\boldsymbol{X} \in \{\boldsymbol{E}, \boldsymbol{H}\}$ and $\sigma \in \{1, 2, 3\}$. We obtain the system

$$\frac{\partial}{\partial x_2}\boldsymbol{E}_3 = i\omega\mu_0\big[\boldsymbol{\mathcal{T}}\big(\mu_{11}\big)\boldsymbol{H}_1 + \boldsymbol{\mathcal{T}}\big(\mu_{12}\big)\boldsymbol{H}_2\big], \tag{6.20}$$

$$-\frac{\partial}{\partial x_1}\boldsymbol{E}_3 = i\omega\mu_0\big[\boldsymbol{\mathcal{T}}\big(\mu_{21}\big)\boldsymbol{H}_1 + \boldsymbol{\mathcal{T}}\big(\mu_{22}\big)\boldsymbol{H}_2\big], \tag{6.21}$$

$$\frac{\partial}{\partial x_1}\boldsymbol{E}_2 - \frac{\partial}{\partial x_2}\boldsymbol{E}_1 = i\omega\mu_0\boldsymbol{\mathcal{T}}\big(\mu_{33}\big)\boldsymbol{H}_3, \tag{6.22}$$

$$\frac{\partial}{\partial x_2}\boldsymbol{H}_3 = -i\omega\varepsilon_0\big[\boldsymbol{\mathcal{T}}\big(\varepsilon_{11}\big)\boldsymbol{E}_1 + \boldsymbol{\mathcal{T}}\big(\varepsilon_{12}\big)\boldsymbol{E}_2\big], \tag{6.23}$$

$$\frac{\partial}{\partial x_1}\boldsymbol{H}_3 = i\omega\varepsilon_0\big[\boldsymbol{\mathcal{T}}\big(\varepsilon_{21}\big)\boldsymbol{E}_1 + \boldsymbol{\mathcal{T}}\big(\varepsilon_{22}\big)\boldsymbol{E}_2\big], \tag{6.24}$$

$$\frac{\partial}{\partial x_1}\boldsymbol{H}_2 - \frac{\partial}{\partial x_2}\boldsymbol{H}_1 = -i\omega\varepsilon_0\boldsymbol{\mathcal{T}}\big(\varepsilon_{33}\big)\boldsymbol{E}_3. \tag{6.25}$$

As usual for RCWA we set $\boldsymbol{f}(x_2) = \left[\boldsymbol{E}_1^T, \boldsymbol{E}_3^T, \eta_0\boldsymbol{H}_1^T, \eta_0\boldsymbol{H}_3^T\right]^T$. The dependence on $\boldsymbol{E}_2$ and $\boldsymbol{H}_2$ can be removed from the system (6.20)–(6.25). Using (6.21) we have

$$\eta_0\boldsymbol{H}_2 = -\boldsymbol{\mathcal{T}}\big(\mu_{22}\big)^{-1}\left[\frac{1}{\kappa}\boldsymbol{\alpha}\boldsymbol{E}_3 + \boldsymbol{\mathcal{T}}\big(\mu_{21}\big)\eta_0\boldsymbol{H}_1\right], \tag{6.26}$$

and from (6.24) we find

$$\boldsymbol{E}_2 = \boldsymbol{\mathcal{T}}\big(\varepsilon_{22}\big)^{-1}\left[\frac{1}{\kappa}\boldsymbol{\alpha}\eta_0\boldsymbol{H}_3 - \boldsymbol{\mathcal{T}}\big(\varepsilon_{21}\big)\boldsymbol{E}_1\right]. \tag{6.27}$$

Using (6.26) and (6.27) in the remaining four equations, we find that

$$\frac{\partial}{\partial x_2}\boldsymbol{f}(x_2) = i\boldsymbol{P}(x_2)\boldsymbol{f}(x_2), \tag{6.28}$$

where

$$\boldsymbol{P}(x_2) = \begin{pmatrix} \boldsymbol{P}_{11}(x_2) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{14}(x_2) \\ \boldsymbol{0} & \boldsymbol{P}_{22}(x_2) & \boldsymbol{P}_{23}(x_2) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{32}(x_2) & \boldsymbol{P}_{33}(x_2) & \boldsymbol{0} \\ \boldsymbol{P}_{41}(x_2) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{44}(x_2). \end{pmatrix}, \tag{6.29}$$

and the submatrices are defined as

$$\begin{aligned}
\boldsymbol{P}_{11}(x_2) &= -\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\big(\varepsilon_{22}\big)^{-1}\boldsymbol{\mathcal{T}}\big(\varepsilon_{21}\big), \\
\boldsymbol{P}_{14}(x_2) &= \frac{1}{\kappa}\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\big(\varepsilon_{22}\big)^{-1}\boldsymbol{\alpha} - \kappa\boldsymbol{\mathcal{T}}\big(\mu_{33}\big), \\
\boldsymbol{P}_{22}(x_2) &= -\boldsymbol{\mathcal{T}}\big(\mu_{12}\big)\boldsymbol{\mathcal{T}}\big(\mu_{22}\big)^{-1}\boldsymbol{\alpha}, \\
\boldsymbol{P}_{23}(x_2) &= \kappa\left[\boldsymbol{\mathcal{T}}\big(\mu_{11}\big) - \boldsymbol{\mathcal{T}}\big(\mu_{12}\big)\boldsymbol{\mathcal{T}}\big(\mu_{22}\big)^{-1}\boldsymbol{\mathcal{T}}\big(\mu_{21}\big)\right], \\
\boldsymbol{P}_{32}(x_2) &= \kappa\boldsymbol{\mathcal{T}}\big(\varepsilon_{33}\big) - \frac{1}{\kappa}\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\big(\mu_{22}\big)^{-1}\boldsymbol{\alpha}, \\
\boldsymbol{P}_{33}(x_2) &= -\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\big(\mu_{22}\big)^{-1}\boldsymbol{\mathcal{T}}\big(\mu_{21}\big), \\
\boldsymbol{P}_{41}(x_2) &= \kappa\left[\boldsymbol{\mathcal{T}}\big(\varepsilon_{12}\big)\boldsymbol{\mathcal{T}}\big(\varepsilon_{22}\big)^{-1}\boldsymbol{\mathcal{T}}\big(\varepsilon_{21}\big) - \boldsymbol{\mathcal{T}}\big(\varepsilon_{11}\big)\right], \\
\boldsymbol{P}_{44}(x_2) &= -\boldsymbol{\mathcal{T}}\big(\varepsilon_{12}\big)\boldsymbol{\mathcal{T}}\big(\varepsilon_{22}\big)^{-1}\boldsymbol{\alpha}.
\end{aligned}$$

The only x_2 dependence in (6.29) is from $S(x_2)$, and so we sample S in the center line of each slice S_j. There are many different choices for S, and we can choose it to

be as smooth as we want (for example, see (6.30)). After setting $S = S(h_{j+\frac{1}{2}})$ in slice S_j, we apply the RCWA algorithm we described in Section 3.4 to (6.29). The boundary conditions (3.41)–(3.44) or (3.47)–(3.43) are enforced, depending on the desired polarization of the incident field. These boundary conditions are the same as for the standard RCWA method since we have chosen S such that $A = I$ and $\det DG = 1$ on Γ_H and Γ_{-H}. Chandezon's original method sets $S \equiv 1$, and therefore requires different boundary conditions. Our method avoids this issue altogether.

Once the C-RCWA solution $\hat{u}_C^t$ is found, the final step is to map this solution back into the original domain Ω. Taking $u_C^t = \hat{u}_C^t \circ G^{-1}$, we can compute u_C^t by using the coordinate transform (6.1).

6.4 Preliminary Numerical Results

In this section, we use the C-RCWA method to compute u_C^t for a grating $g(x_1) = 50\cos(\pi x_1)$ illuminated by a p-polarized incident field. Above the 50-nm-thick grating region there is an air layer 950-nm-thick, where $\varepsilon_a = 1 + 10^{-6}i$. Below the grating, there is a 1000-nm-thick layer of a fictitious metal with $\varepsilon_m = -15 + 4i$.

In this numerical test, where $-H < \hat{x}_2 < H$ we take

$$S(\hat{x}_2) = \begin{cases} 1 - \frac{3}{H^2}\hat{x}_2^2 + \frac{2}{H^3}\hat{x}_2^3 & \text{if } 0 \le \hat{x}_2 \le H, \\ 1 - \frac{3}{H^2}\hat{x}_2^2 - \frac{2}{H^3}\hat{x}_2^3 & \text{if } -H \le \hat{x}_2 \le 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.30}$$

Figure 6.2 shows the field maps of the C-RCWA solution $\hat{u}_C^t$ and the mapped solution u_C^t. In Figure 6.3, we compare the RCWA solution $u^{h,M,t}$ to the mapped C-RCWA solution u_C^t. Here, we are interested in the near-field, so these plots are zoomed into the grating region. The top row is the RCWA solution, and the bottom row is the C-RCWA solution. In all plots, the sinusoidal grating is outlined in white for clarity. In the top row, spurious low-amplitude oscillations in the near-field are seen. However, the C-RCWA removes these oscillations entirely.

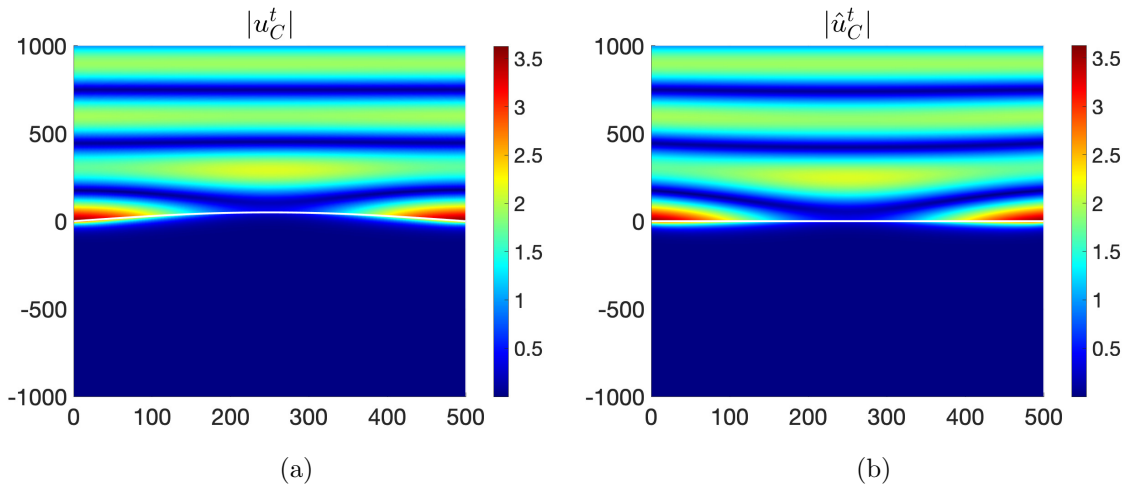


Figure 6.2: Field maps of the solution u_C^t in Ω (a) and $\hat{u}_C^t$ in the mapped domain $\hat{\Omega}$ (b). For clarity, the grating profiles are outlined in white.

6.5 Conclusion

We have derived a family of coordinate transform methods for our scattering problem. Our study of these methods shows that it is equivalent to the RCWA applied to bianisotropic media with flat interfaces. We studied the simple case where there is only one interface, although this method can be applied to a medium with any number of gratings. This can be achieved by stacking grating regions on top of one another, and applying the RCWA algorithm to the entire stack. Since $S(\hat{x}_2) = 0$ on the top and bottom boundaries, these regions fit together and the resulting coordinate transform would have the desired properties on the entire domain.

The issue of convergence of this coordinate transform method is still open, although could be resolved based on the analysis in this thesis. The analysis of the Helmholtz problem with anisotropic coefficients is simplified somewhat, since it comes from a coordinate transform. Another area of interest is applying this method to crossed gratings in 3D, which would open up the possibility of rapidly solving the full 3D Maxwell system in photonic crystals.

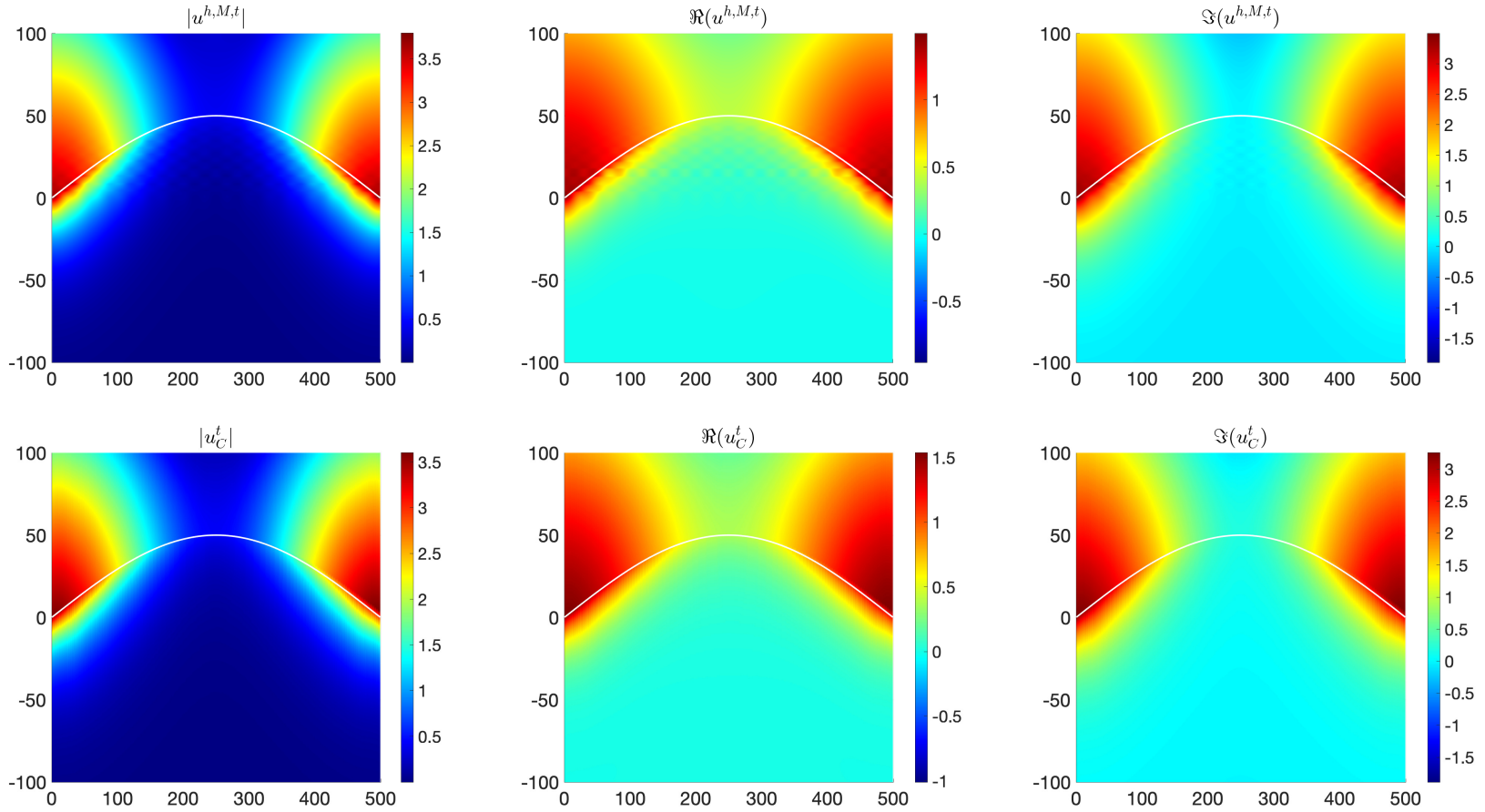


Figure 6.3: A comparison of the RCWA solution (top row when viewed horizontally) with the C-RCWA solution (bottom row when viewed horizontally). The RCWA solutions exhibit low-amplitude spurious oscillations near the grating.

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this concluding chapter, we discuss the scope of our analysis and some extensions and cases that remain open. We also discuss some shortcomings of our analysis.

1. In Chapter 4 we proved the convergence of the RCWA for s-polarized incident light in three different cases, depending on the sign of the real and imaginary part of ε. A case we did not consider is where $\Re(\varepsilon) < 0$. In this case the variational problem is coercive and the analysis would follow by using the Strang lemmas. We restricted our analysis to the mathematically interesting cases, and also to the cases applicable to solar-cell modeling.

2. In Chapter 5 we did not analyze the cases when $\Im(\varepsilon) > c_1 > 0$ and $\Re(\varepsilon) < 0$. In solar-cell applications this is often the case, for example, with metallic gratings for certain wavelengths of incident light. This case can probably be analyzed by using our *a-priori* estimates when the right hand side $F \in (H^1_{qp}(\Omega))'$. The argument would be similar to the one used in Chapter 4, where the *a-priori* estimate for the purely real ε case is extended to the other cases by first rewriting the Helmholtz equation. A similar argument can be found in [16], but there $\Re(\varepsilon) > 0$ and $\Im(\varepsilon) > 0$. However, we have already considered this case by first proving that these conditions imply that the variational problem is coercive.

3. In Chapters 4 and 5, the convergence rate with respect to slice thickness $h > 0$ does not match the observed higher convergence rate. It is not clear from our analysis how the theoretical convergence rate can be improved, so perhaps another technique is needed.

4. In order to mitigate some shortcomings of our analysis, in Chapter 6 we introduced a hybrid method where no stairstep approximation of the interface between materials is needed. Instead, a coordinate transform is applied so that the interfaces in the transformed domain are flat. Preliminary numerical results indicate that this method is more stable than the standard RCWA in the case where $\Re(\varepsilon) < 0$. Therefore, this hybrid C-RCWA is promising for solar-cell applications like scattering by a metallic grating.

We now make some comments about future work that could extend the results of this thesis to crossed gratings, i.e., a grating that is periodic in both x_1 and x_2. In this thesis, we focused on problems where the domain is invariant in the x_3–direction, and therefore focused on the analysis of Helmholtz equations in periodic media. It is only natural to extend our analysis to the case of general electromagnetic scattering, and therefore we need to analyze the full Maxwell system.

In the future, we hope to derive the C-RCWA method for the full Maxwell system, which is equivalent to the RCWA for flat gratings with bianisotropic coefficients. As we showed in Chapter 6, the coefficients must be chosen so that the solution solves the appropriate transformed variational problem. As in the 2D case, this method should be able to handle the case where $\Re(\varepsilon) < 0$.

To analyze this method, the general technique of this thesis may be used. Rellich identities for the continuous problems can be used to prove *a-priori* estimates and then existence and uniqueness of the solutions would follow. The goal would be to show convergence of this method in the slice thickness parameter $h > 0$ and retained Fourier modes M.

REFERENCES

[1] M. Faryad and A. Lakhtakia, Grating-coupled excitation of multiple surface plasmon-polariton waves, *Phys. Rev. A*, 84(3):033852, 2011.

[2] J.A. Polo Jr., T.G. Mackay, and A. Lakhtakia, *Electromagnetic Surface Waves: A Modern Perspective*, Elsevier, Waltham, MA, USA, 2013.

[3] H. Kogelnik, Coupled wave theory for thick hologram gratings, *Bell Syst. Tech. J.*, 48(9):2909–2947, 1969.

[4] M.G. Moharam and T.K. Gaylord, Rigorous coupled-wave analysis of planar grating diffraction, *J. Opt. Soc. Am.*, 71(7):811–818, 1981.

[5] L. Li, Use of Fourier series in the analysis of discontinuous periodic structures, *J. Opt. Soc. Am. A*, 13(9):1870–1876, 1996.

[6] J. Homola (Ed.), *Surface Plasmon Resonance Based Sensors*, Springer, Heidelberg, Germany, 2006.

[7] L.M. Anderson, Harnessing surface plasmons for solar energy conversion, *Proc. SPIE*, 408(1):172–178, 1983.

[8] D. Alonso-Álvarez, T. Wilson, P. Pearce, M. Führer, D. Farrell, and N. Ekins-Daukes, Solcore: a multi-scale, Python-based library for modelling solar cells and semiconductor materials, *J. Comput. Electron.*, 17(3):1099–1123, 2018.

[9] J. J. Hench and Z. Strakoš, The RCWA method–A case study with open questions and perspectives of algebraic computations, *Electron. Trans. Numer. Anal.*, 31:331–357, 2008.

[10] B. D. Guenther, *Modern Optics*, Wiley, Hoboken, NJ, USA, 1990.

[11] M.V. Shuba, M. Faryad, M.E. Solano, P.B. Monk, and A. Lakhtakia, Adequacy of the rigorous coupled-wave approach for thin-film silicon solar cells with periodically corrugated metallic backreflectors: spectral analysis, *J. Opt. Soc. Am. A*, 32(7):1222–1230, 2015.

[12] F. Ahmad, T.H. Anderson, P.B. Monk, and A. Lakhtakia, Optimization of light trapping in ultrathin nonhomogeneous Cu1–ξGa–$\xi$$Se_2$ solar cell backed by 1D periodically corrugated backreflector, *Proc. SPIE*, 10731(1):107310L, 2018.

[13] T.H. Anderson, A. Lakhtakia, and P.B. Monk, Optimization of nonhomogeneous indium-gallium nitride Schottky-barrier thin-film solar cells, *J. Photon. Energy*, 8(3):034501, 2018.

[14] J.A. DeSanto, Scattering by rough surfaces, in: R. Pike and P. Sabatier (Eds.), *Scattering: Scattering and Inverse Scattering in Pure and Applied Science*: 15–36, Academic Press, San Diego, CA, US, 2002.

[15] S.N. Chandler-Wilde, P. Monk, and M. Thomas, The mathematics of scattering by unbounded, rough, inhomogeneous layers, *J. Comput. Appl. Math.*, 204(2):549–559, 2007.

[16] A. Lechleiter and S. Ritterbusch, A variational method for wave scattering from penetrable rough layers, *IMA J. Appl. Math.*, 75:366–391, 2010.

[17] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, Germany, 1998.

[18] A.H. Schatz, An observation concerning Ritz–Galerkin methods with indefinite bilinear forms, *Math. Comput.*, 28(128):959–962, 1974.

[19] Z. Chen and H. Wu, An adaptive finite element method with perfectly matched absorbing layers for the wave scattering by periodic structures, *SIAM J. Numer. Anal.*, 41(3):799–826, 2003.

[20] M. Ainsworth, J.Z. Zhu, A.W. Craig, and O.C. Zienkiewicz, Analysis of the Zienkiewicz–Zhu *a-posteriori* error estimator in the finite element method, *Int. J. Numer. Math. Eng.*, 28(9):2161–2174, 1989.

[21] H. Ammari and G. Bao, Maxwell's equations in periodic chiral structures, *Math. Nachr.*, 251(1):3–18, 2003.

[22] L.R. Scott and S. Brenner, *The Mathematical Theory of Finite Element Methods*, Springer, New York, USA, 2008.

[23] W. Dörfler, A. Lechleiter, M. Plum, G. Schneider and C. Wieners, *Photonic Crystals: Mathematical Analysis and Numerical Approximation*, Birkhäuser, Basel, Switzerland, 2011.

[24] D. Maystre (Ed.), *Selected Papers on Diffraction Gratings*, SPIE, Bellingham, WA, USA, 1993.

[25] E.G. Loewen and E. Popov,*Diffraction Gratings and Applications*, Marcel Dekker, New York, NY, USA,1997.

[26] M.G. Moharam, E.B. Grann, D.A. Pommet, and T.K. Gaylord, Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings, *J. Opt. Soc. Am. A*, 12(5):1068–1076, 1995.

[27] V.A. Yakubovich and V.M. Starzhinskii, *Linear Differential Equations with Periodic Coefficients*, Wiley, New York, NY, USA, 1975.

[28] M.G. Moharam, D.A. Pommet, E.B. Grann, and T.K. Gaylord, Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach, *J. Opt. Soc. Amer. A*, 12(5):1077–1086, 1995.

[29] P. Lalanne and G.M. Morris, Highly improved convergence of the coupled-wave method for TM polarization, *J. Opt. Soc. Am. A*, 13(4):779–784, 1996.

[30] V. Liu and S. Fan, S^4: A free electromagnetic solver for layered periodic structures, *Comput. Phys. Commun.*, 183(10):2233–2244, 2012.

[31] F. Ahmad, T.H. Anderson, P.B. Monk, and A. Lakhtakia, Efficiency enhancement of ultrathin CIGS solar cells by optimal bandgap grading, *Appl. Opt.*, 58(22):6067–6078, 2019.

[32] F. Ahmad, T. H. Anderson, P. B. Monk, and A. Lakhtakia, Efficiency enhancement of ultrathin CIGS solar cells by optimal bandgap grading: erratum, *Appl. Opt.*, 59:2615-2615, 2020.

[33] F. Ahmad, A. Lakhtakia, and P.B. Monk, Optoelectronic optimization of graded-bandgap thin-film AlGaAs solar cells, *Appl. Opt.*,58(22):6067-6078, 2019.

[34] T.H. Anderson, B.J. Civiletti, P.B. Monk, and A. Lakhtakia, Combined optoelectronic simulation and optimization of solar cells, *J. Comput. Phys.*, 407:109242, 2020.

[35] I.G. Graham, O.R. Pembery, and E.A. Spence, The Helmholtz equation in heterogeneous media: a priori bounds, well-posedness, and resonances, *J. Diff. Eqns.*, 266(6):2869–2923, 2019.

[36] B.J. Civiletti, A. Lakhtakia, and P.B. Monk, Analysis of the rigorous coupled wave approach for s-polarized light in gratings, *J. Comput. Appl. Math.*, 368(1):12478, 2020.

[37] E. Previato (Ed.), *Dictionary of Applied Math for Engineers and Scientists*, CRC Press, Boca Raton, FL, USA, 2003.

[38] C.H. Palmer Jr., Parallel diffraction grating anomalies, *J. Opt. Soc. Am.*, 42(4):269–276, 1952.

[39] W.V. Petryshyn, Constructional proof of Lax–Milgram Lemma and its application to non-K-P.D. abstract and differential operator equations, *J. SIAM Numer. Anal. B*, 2(3):404–420, 1965.

[40] L.C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, USA, 1998.

[41] C. Bernardi and R. Verfürth, Adaptive finite element methods for elliptic equations with non-smooth coefficients, *Numer. Math*, 85(4):579–608, 2000.

[42] F. Kahmann, Separate and simultaneous investigation of absorption gratings and refractive-index gratings by beam-coupling analysis, *J. Opt. Soc. Am. A*, 10(7):1562–1569, 1993.

[43] M. Fally, Separate and simultaneous investigation of absorption gratings and refractive-index gratings by beam-coupling analysis: comment, *J. Opt. Soc. Am. A*, 23(10):2662–2663, 2006.

[44] A.W. Brown and M. Xiao, All-optical switching and routing based on an electromagnetically induced absorption grating, *Opt. Lett.*, 30(7) 699–701, 2005.

[45] J.S. Walker, *Fourier Analysis*, Oxford University Press, New York, NY, USA, 1988.

[46] G.C. Hsiao, W.L. Wendland, The Aubin–Nitsche lemma for integral equations, *J. Integral Eqns.*, 3(4):299–315, 1981.

[47] J. Schöberl, Netgen/NGsolve, 2019. https://ngsolve.org.

[48] A. Bonito, J.-L. Guermond, and F. Luddens, Regularity of the Maxwell equations in heterogeneous media and Lipschitz domains, *J. Math. Anal. Appl.*, 408(2):498–512, 2013.

[49] J. Elschner and G. Schmidt, Diffraction in periodic structures and optimal design of binary gratings. Part I: Direct problems and gradient formulas, *Math. Methods Appl. Sci.*, 21(4):1297–1342, 1998.

[50] J. Hadamard, *Lectures on the Cauchy Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, CT, USA,1923.

[51] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, New York, NY, USA, 2000.

[52] R. Kress, *Linear Integral Equations*. Springer, Heidelberg, Germany, 1999.

[53] S.T. Peng, H.L. Bertoni, and T. Tamir, Analysis of periodic thin-film structures with rectangular profile, *Opt. Commun.*, 10: 91–94, 1974.

[54] B.J. Civiletti, A. Lakhtakia, and P.B. Monk, Analysis of the rigorous coupled wave approach for p-polarized light in gratings (submitted).

[55] Renewable Resource Data Center. Reference solar spectral irradiance: ASTM G-173, 2003. URL rredc.nrel.gov/solar/spectra/am1.5/.

[56] A. Barnett and L. Greengard, A new integral representation for quasi-periodic fields and its application to two dimensional band structure calculations, *J. Comput. Phys.*, 229:6898–6914, 2010.

[57] A. Lechleiter, The Floquet-Bloch transform and scattering from locally perturbed periodic surfaces, *J. Math. Anal. Appl.*, 446:605–627, 2017.

[58] G. Granet and B. Guizal, Efficient implemenation of the coupled-wave method for metallic lamellar grating in TM polarization, *J. Opt. Soc. Am. A*, 13:1019–1023, 1996.

[59] T. Schuster, J. Ruoff, N. Kerwien, S. Rafler, and W. Osten, Normal vector method for convergence improvement using the RCWA for crossed gratings, *J. Opt. Soc. Am. A*, 24:2880–2890, 2007.

[60] J. Chandezon, G. Raoult, and D. Maystre, A new theoretical method for diffraction gratings and its numerical application, *J. Opt. (France)*, 11: 235–241, 1980.

[61] P.B. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, 2003.

[62] L. Li., Fourier Modal Method. in: E. Popov, ed., *Gratings: Theory and Numeric Applications, Second Revisited Edition*: 13.1–13.40, Aix Marseille Université, CNRS, Centrale Marseille, Institut Fresnel, 2014.

[63] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Springer-Verlag, New York, NY, 1998.

Appendix A

2D RCWA CODE

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 3D RCWA code
%
% (1) You must load experiment.mat before running
% (2) To edit the experiment, edit experiment.mat
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% lambda         : incident light wavelength (nm)
% theta          : incidence angle of light in the x1,x2 plane (radians)
% psi            : incidence angle of light in the x2,x3 plane (radians)
% L1             : period of grating (nm) in x1--direction
% L2             : period of grating (nm) in x3--direction
% M1             : Fourier truncation parameter in the x1--direction
% M2             : Forier truncation parameter in the x2--direction
% numberOfLayers : total number of different material layers in simulation
% numberOfSlices : (1 x numberOfLayers) vector containing the number of
%                  slices in each layer, from top to bottom
% epsilonValues  : the constant value of relative epsilon in each layer,
%                  from top (x2=0) to bottom. 0 implies a grating layer.
% sliceBoundary  : (1 x numberOfSlices+1) vector containing slice boundary
%                  locations (nm)
% midSliceValues : (1 x numberOfSlices) vector containing locations of
%                  slice midpoints (nm)
% layerThickness : (1 x numberOfLayers) vector containing the thickness(nm)
%                  of each layer, from top to bottom
```

```
% layerBoundary  : (1 x numberOfLayers+1) vector containing the locations
%                : of the layer boundaries, from top to bottom (nm)
% x1Values       : (1 x L1) vector of x1 values, nanometer scaling in x1
%                : from -L1/2 to L1/2
% x3Values       : (1 x L2) vector of x3 values, nanometer scaling in x3
%                : from -L2/2 to L2/2
% Sample         : the number of samples to use for FFT in x1 and x3
% layerIndex     : (1 x numberOfSlices) vector of values for the layer
%                : index, from top to bottom.
% gratingProfile : chebfun2 object describing the grating surface, a
%                  function of x,y
% mu0            : permeability of free-space (H m^-1)
% epsilon0       : permittivity of free-space (F m^-1)
% kappa          : wave number
%
% belowGrating   : the value of the grating relative permittivity
% aboveGrating   : the value of the relative permittivity above grating
%                : (but still in the grating region)
% polarization   : polarization of incident field (0 is s, 1 is p)
%
% plotField      : Setting if you want to plot the field (1 for plot)
%
% Last Edited: 5/10/20

epsilon0=8.854*10^-12;
mu0=4*pi*10^-7;
eta0=sqrt(mu0/epsilon0);
kappa=2*pi/lambda;
omega=kappa/eta0;
x = chebfun2(@(x,y) x, [-L1/2 L1/2 -L2/2 L2/2]);
y = chebfun2(@(x,y) y, [-L1/2 L1/2 -L2/2 L2/2]);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% SET THE EXPERIMENT PARAMETERS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
polarization=experiment.polarization;
plotField=experiment.plotField;
M1=experiment.M1;
M2=experiment.M2;
L1=experiment.L1;
L2=experiment.L2;
%gratingProfile=100*cos(pi/L1*x).*cos(pi/L2*y).^2;
%gratingProfile=50*cos(pi/L1*x);
gratingProfile=experiment.gratingProfile;
invariantGrating=experiment.invariantGrating;
lambda=experiment.lambda;
epsilonValues=experiment.epsilonValues;
aboveGrating=experiment.aboveGrating;
belowGrating=experiment.belowGrating;
theta=experiment.theta;
psi=experiment.psi;
numberOfSlices=experiment.numberOfSlices;
layerThickness=experiment.layerThickness;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% PREALLOCATE MATRICES + BOUNDARY CONDITIONS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

numberOfLayers=length(numberOfSlices);
totalNumberOfSlices=sum(numberOfSlices);
sliceThicknessPerLayer=layerThickness./numberOfSlices;
layerBoundary=[0 cumsum(layerThickness)];

midSliceValues=[];
layerIndex=[];
for i=1:numberOfLayers
midSliceValues=[midSliceValues linspace(layerBoundary(i)+...
    sliceThicknessPerLayer(i)/2,layerBoundary(i+1)-...
    sliceThicknessPerLayer(i)/2,numberOfSlices(i))];
```

```
layerIndex=[layerIndex i*ones(1,numberOfSlices(i))];
end

x1Values=linspace(-L1/2+1/2,L1/2-1/2,L1);
x3Values=linspace(-L2/2+1/2,L2/2-1/2,L2);

Sample=200;
x1SampleValues=linspace(-L1/2+1/2,L1/2-1/2,Sample);
x3SampleValues=linspace(-L2/2+1/2,L2/2-1/2,Sample);

alpha1=kappa*sin(theta)*cos(psi)+(2*pi/L1)*(-M1:M1);
alpha2=kappa*sin(theta)*sin(psi)+(2*pi/L2)*(-M2:M2);
alpha=zeros(2*M1+1,2*M2+1);
beta=zeros(2*M1+1,2*M2+1);
s1=zeros(2*M1+1,2*M1+1);
s2=zeros(2*M2+1,2*M2+1);
pinc1=zeros(2*M2+1,2*M2+1);
pinc2=zeros(2*M2+1,2*M2+1);
pinc3=zeros(2*M2+1,2*M2+1);
pref1=zeros(2*M2+1,2*M2+1);
pref2=zeros(2*M2+1,2*M2+1);
pref3=zeros(2*M2+1,2*M2+1);
ptr1=zeros(2*M2+1,2*M2+1);
ptr2=zeros(2*M2+1,2*M2+1);
ptr3=zeros(2*M2+1,2*M2+1);
for i=1:2*M1+1
    for j=1:2*M2+1
        alpha(i,j)=sqrt(alpha1(i).^2+alpha2(j).^2);
        beta(i,j)=sqrt(kappa^2-alpha1(i).^2-alpha2(j).^2);
        s1(i,j)=-alpha2(j)/alpha(i,j);
        s2(i,j)=alpha1(i)/alpha(i,j);
        pinc1(i,j)=-(1/kappa)*alpha1(i)*beta(i,j)/alpha(i,j);
        pinc2(i,j)=-(1/kappa)*alpha2(j)*beta(i,j)/alpha(i,j);
        pinc3(i,j)=(1/kappa)*alpha(i,j);
        pref1(i,j)=(1/kappa)*alpha1(i)*beta(i,j)/alpha(i,j);
```

```
        pref2(i,j)=(1/kappa)*alpha2(j)*beta(i,j)/alpha(i,j);
        pref3(i,j)=(1/kappa)*alpha(i,j);
        ptr1(i,j)=-(1/kappa)*alpha1(i)*beta(i,j)/alpha(i,j);
        ptr2(i,j)=-(1/kappa)*alpha2(j)*beta(i,j)/alpha(i,j);
        ptr3(i,j)=(1/kappa)*alpha(i,j);
    end
end

tauIndex=2*M1*M2+M1+M2;
A=zeros(2*(2*tauIndex+1),1);
sx=zeros(2*tauIndex+1,2*tauIndex+1);
sy=zeros(2*tauIndex+1,2*tauIndex+1);
pincx=zeros(2*tauIndex+1,2*tauIndex+1);
pincy=zeros(2*tauIndex+1,2*tauIndex+1);
prefx=zeros(2*tauIndex+1,2*tauIndex+1);
prefy=zeros(2*tauIndex+1,2*tauIndex+1);
ptrx=zeros(2*tauIndex+1,2*tauIndex+1);
ptry=zeros(2*tauIndex+1,2*tauIndex+1);
for i=1:2*M1+1
    for j=1:2*M2+1
        tau=(i-1-M1)*(2*M2+1)+(j-1-M2);
        tau=tau+tauIndex+1;
        sx(tau,tau)=s1(i,j);
        sy(tau,tau)=s2(i,j);
        pincx(tau,tau)=pinc1(i,j);
        pincy(tau,tau)=pinc2(i,j);
        prefx(tau,tau)=pref1(i,j);
        prefy(tau,tau)=pref2(i,j);
        ptrx(tau,tau)=ptr1(i,j);
        ptry(tau,tau)=ptr2(i,j);
    end
end
Yeinc=[sx pincx; sy pincy];
Yhinc=[pincx -sx;pincy -sy];
Yetr=[sx ptrx;sy ptry];
```

```
Yhtr=[ptrx -sx;ptry -sy];
Yeref=[sx prefx; sy prefy];
Yhref=[prefx -sx; prefy -sy];
alphaMat1=diag(alpha1);
alphaMat2=diag(alpha2);

IdentityMatrix=eye(2*tauIndex+1,2*tauIndex+1);
ZeroMatrix=zeros(2*tauIndex+1,2*tauIndex+1);
bottomOfBoundary=layerBoundary(2:end);
locationMat1=zeros(Sample,Sample);
sampleEpsilonMat=zeros(Sample,Sample);
zeroVecPad1=zeros(1,M2*(2*M1+1));
zeroVecPad2=zeros(1,M1*(2*M2+1));
kx=kappa*sin(theta)*cos(psi)+(2*pi/L1)*(-M1:M1);
ky=(kappa*sin(theta)*sin(psi)+(2*pi/L2)*(-M2:M2));
KX=zeros(tauIndex,tauIndex);
KY=zeros(tauIndex,tauIndex);
for i=-tauIndex:tauIndex
    [m,n,¬] = findIndex(i,M1,M2);
    in=i+tauIndex+1;
    m=m+M1+1;
    n=n+M2+1;
    KX(in,in)=kx(m);
    KY(in,in)=ky(n);
end

Z=cell(1,totalNumberOfSlices+1);
T=cell(1,totalNumberOfSlices+1);
G=cell(1,totalNumberOfSlices);
D=cell(1,totalNumberOfSlices);
Wus=cell(1,totalNumberOfSlices);
Dus=cell(1,totalNumberOfSlices);
f=cell(1,totalNumberOfSlices+1);
ex=cell(1,totalNumberOfSlices+1);
ey=cell(1,totalNumberOfSlices+1);
```

```
ez=cell(1,totalNumberOfSlices+1);
hx=cell(1,totalNumberOfSlices+1);
hy=cell(1,totalNumberOfSlices+1);

if polarization==0
    A(tauIndex+1,1)=1;
elseif polarization==1
    A(3*tauIndex+2,1)=1;
end
dz=1;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% RCWA ALGORITHM BEGINS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Z{totalNumberOfSlices+1}=[Yeinc;Yhinc];

for j=totalNumberOfSlices:-1:1
    epsilonValueInSlice=epsilonValues(layerIndex(j));
    midSliceValue=midSliceValues(j);
    if epsilonValueInSlice==0
        bottomBoundary=bottomOfBoundary(layerIndex(j));
        gRoot=chebfun2(gratingProfile-(bottomBoundary-midSliceValue));
        intersectionCurve=roots(gRoot);
        x1ValsOfIntersection=real(intersectionCurve(2*x1SampleValues/L1));
        x2ValsOfIntersection=imag(intersectionCurve(2*x3SampleValues/L2));
        if invariantGrating==1
            intPointOnex1=x1ValsOfIntersection(1);
            intPointTwox1=x1ValsOfIntersection(end);
            ind=find(x1SampleValues>intPointOnex1 & ...
                x1SampleValues<intPointTwox1);
            sampleEpsilonMat=aboveGrating*ones(Sample,Sample);
            sampleEpsilonMat(:,ind(1):ind(end))=belowGrating;
        else
            for k=1:Sample
```

```
            locationMat1(k,:)=inpolygon(x1SampleValues(k)*...
                ones(1,Sample),x3SampleValues,x1ValsOfIntersection,...
                x2ValsOfIntersection);
            locationMat2=¬locationMat1;
            %grating region, so we sample epsilon with FFT
            sampleEpsilonMat=belowGrating*locationMat1+...
                aboveGrating*locationMat2;
        end
    end
    epfft=fftshift(fft2(sampleEpsilonMat))/(Sample^2);
    epMat=epfft(Sample/2+1-2*tauIndex:Sample/2+1+2*tauIndex,...
        Sample/2+1-2*tauIndex:Sample/2+1+2*tauIndex);
    epMat=epMat.';
    %construct the Toeplitz matrix of epsilon
    ToeplitzOfEpsilon=zeros(2*tauIndex+1,2*tauIndex+1);
    for kk=-tauIndex:tauIndex
        for ii=-tauIndex:tauIndex
            [m,n,¬] = findIndex(kk,M1,M2);
            [mPrime,nPrime,¬] = findIndex(ii,M1,M2);
            mMinusmPrime=m-mPrime+2*tauIndex+1;
            nMinusnPrime=n-nPrime+2*tauIndex+1;
            kn=kk+tauIndex+1;
            in=ii+tauIndex+1;
            ToeplitzOfEpsilon(kn,in)=epMat(mMinusmPrime,nMinusnPrime);
        end
    end
else
    ToeplitzOfEpsilon=epsilonValueInSlice*IdentityMatrix;
    ToeplitzOfOneUponEpsilon=(1/epsilonValueInSlice)*IdentityMatrix;
end
%construct the P matrix for d/dz f = i P f
P14=kappa*IdentityMatrix-(1/kappa)*KX/ToeplitzOfEpsilon*KX;
P13=(1/kappa)*KX/ToeplitzOfEpsilon*KY;
P23=(1/kappa)*KY/ToeplitzOfEpsilon*KY-kappa*IdentityMatrix;
P24=-(1/kappa)*KY/ToeplitzOfEpsilon*KY;
```

```
    P32=(1/kappa)*KX^2-kappa*ToeplitzOfEpsilon;
    P31=-(1/kappa)*KX*KY;
    P41=kappa*ToeplitzOfEpsilon-(1/kappa)*KY^2;
    P42=(1/kappa)*KY*KX;

    P=[ZeroMatrix ZeroMatrix P13         P14
       ZeroMatrix ZeroMatrix P23         P24
       P31        P32        ZeroMatrix  ZeroMatrix
       P41        P42        ZeroMatrix  ZeroMatrix
                                                ];
    [G,D] = eig(P);
    %sort the eigenvalues
    sortD=imag(diag(D));
    [¬,ind]=sort(sortD,'descend');
    Dtemp=diag(D);
    D=Dtemp(ind);
    D=diag(D);
    G=G(:,ind);
    Du=D(1:2*(2*tauIndex+1),1:2*(2*tauIndex+1));
    Dl=D(2*(2*tauIndex+1)+1:end,2*(2*tauIndex+1)+1:end);
    W=G\Z{j+1};
    Wu=W(1:2*(2*tauIndex+1),:);
    Wl=W(2*(2*tauIndex+1)+1:end,:);
    Wus{j}=Wu;
    Dus{j}=Du;
    aux1=expm(1i*dz*Du);
    aux2=expm(-1i*dz*Dl);
    Lower=aux2*(Wl)*(Wu\aux1);
    Z{j}=G*[eye(2*(2*tauIndex+1)); Lower];
end

Z0=Z{1};
Z0u=Z0(1:2*(2*tauIndex+1),:);
Z0l=Z0(2*(2*tauIndex+1)+1:end,:);
TRmat=[Z0u -Yeref; Z0l -Yhref]\([Yeinc;Yhinc]*A);
```

```
T0=TRmat(1:2*(2*tauIndex+1),1);
R=TRmat(2*(2*tauIndex+1)+1:end,1);
T{1}=T0;
for j=2:totalNumberOfSlices+1
    T{j}=Wus{j-1}\(expm(1i*dz*Dus{j-1})*T{j-1});
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% RCWA ALGORITHM ENDS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

for j=1:totalNumberOfSlices+1
    f{j}=Z{j}*T{j};
    ex{j}=f{j}(1:2*tauIndex+1);
    ey{j}=f{j}(2*tauIndex+2:4*tauIndex+2);
    hx{j}=f{j}(4*tauIndex+3:6*tauIndex+3);
    hy{j}=f{j}(6*tauIndex+4:8*tauIndex+4);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% RECONSTRUCT THE FIELDS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Ey=zeros(totalNumberOfSlices,L1,1);
Ex=zeros(totalNumberOfSlices,L1,1);
Hy=zeros(totalNumberOfSlices,L1,1);
Hx=zeros(totalNumberOfSlices,L1,1);
mat=zeros(2*tauIndex+1,1);
for j=1:totalNumberOfSlices
    for i=1:L1
            r=[x1Values(i)+L1/2,x3Values(1)+L2/2];
        for tt=-tauIndex:tauIndex
            [m,n,¬] = findIndex(tt,M1,M2);
            m=m+M1+1;
            n=n+M2+1;
```

```
            if M2==0
                mat=exp(1i*diag(KX)*(x1Values(i)+L1/2));
            else
                kvec=[alpha1(m),alpha2(n)];
                mat(tt+tauIndex+1,1)=exp(1i*kvec*r');
            end

        end
        Ex(j,k,1)=(ex{j+1}.')*mat;
        Ey(j,i,1)=(ey{j+1}.')*mat;
        Hx(j,k,1)=(hx{j+1}.')*mat;
        Hy(j,i,1)=(hy{j+1}.')*mat;
    end
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% PLOT THE FIELDS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

if plotField==1
    subplot(1,2,1)
    surf(x1Values,midSliceValues,abs(Ey));
    shading interp; colormap jet; view(2)
    title('|Ey|')
    colorbar
    set(gca,'ydir','reverse')
    subplot(1,2,2)
    surf(x1Values,midSliceValues,abs(Hy));
    shading interp; colormap jet; view(2)
    title('|Hy|')
    colorbar
    set(gca,'ydir','reverse')
else
end
```

Appendix B

PERMISSIONS

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright # Author-rights.