## University of Strathclyde

# Analysis of electromagnetic waves in a periodic diffraction grating using a priori error estimates and a dual weighted residual method. 

Natacha Hajanirina Lord<br>Department of Mathematics and Statistics<br>University of Strathclyde<br>Glasgow, UK<br>January 17, 2012

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material in, or derived from, this thesis.

## Acknowledgements

I am deeply grateful to my supervisor Dr Anthony J. Mulholland for all his help, support, advice and valuable time throughout my thesis. I also want to thank my second supervisor Dr Alison Ramage for her contribution and the Faculty of Science for giving me the opportunity to finish my studies.
I also want to thank all my family and friends for their love, support and encouragement.
My special thanks to my husband whose patient love enabled me to complete this work, my son for all the joy he brings in my life and my dearest mother for never letting me down.
Finally, I would like to dedicate this thesis to the memory of my beloved father who guided me into Mathematics and who always believed in me.

Teny fohy, teny tsotra
noho ny fitiavana miavosa
noho ny tolona sy ny hasasarana
Niaretan'ny dada sy neny andro aman' alina.
Misaotra ry ray aman-dreny.
Tsy tambo isaina ny fitiavanareo sesehena.
Tsy adinoko ianao ry vady malalako.
Nanohina hatrany tao anaty hasasarako.
Ny faniriako raha mbola eto an-tany.
Dia ny hiara-dalana aminao ihany.
Ianareo mirahavavy sy izy telo mianaka koa,
Tadidiko ireo fiarahantsika taloha,
Fahatsiarovana mamy, tena tsaroko.
Rehefa tojo tolona aho ka nitady hitoloko.
Ary ho anao ry sombin' aiko.
Ho reharehan' ny dada sy neny tokoa.
Amin'ny fahendrena sy fomba fiaina.
Ho tafita soa ary ho lava andro iainana ianao sombin' aina.

## Abstract

The problem of using the $\alpha, 0$ and the $\alpha, \beta$-quasi periodic transformations within a finite element method in studying electromagnetic waves in a periodic space is addressed. We investigate an a priori error estimate for both transformations which allows us to solve our problem numerically on a uniform mesh. We also analyse the Dual Weighted Residual (DWR) method with the $\alpha, 0$-quasi periodic transformation to derive an a posteriori error estimate. This error estimate is later used to compute efficiently the numerical solution using an adaptive method. We then implement the above finite element methods. It is shown numerically that our numerical results are in good agreement with those in the literature, the $\alpha, \beta$-quasi periodic method converges at a far lower number of degrees of freedom than the $\alpha, 0$-quasi periodic method and the DWR method converges faster and requires fewer degrees of freedom than the global a posteriori error estimate or the uniform mesh. We also explore the geometrical freedom given by the finite element method and examine wave scattering by the Morpho butterfly wing.

## Contents

1 Introduction ..... 1
1.1 Background ..... 2
1.2 Outline of the thesis ..... 3
2 Physical and mathematical description of the problem ..... 6
2.1 Maxwell's equations and polarization ..... 6
2.2 Problem statement ..... 9
2.3 Matching the analytical solution on to the finite element solution at the transparent boundary ..... 17
2.3.1 Fundamental solution of Helmholtz problem in a homoge- neous domain ..... 17
2.3.2 The Dirichlet to Neumann maps $T, T^{\alpha, 0}$ and $T^{\alpha, \beta}$ ..... 19
2.3.3 Properties of the operators $T, T^{\alpha, 0}, T^{\alpha, \beta}$ ..... 22
2.4 Mathematical formulation of the truncated problem ..... 39
2.5 Summary ..... 44
3 A regularity estimate for the inhomogeneous Helmholtz problem for periodic gratings ..... 45
3.1 General case ..... 45
3.1.1 Case 1A: Perfectly conducting grating: TE case ..... 45
3.1.1.1 The inhomogeneous Helmholtz equation ..... 45
3.1.1.2 The $\alpha$-quasi-periodic Green functions of the Helmholtz equation ..... 46
3.1.1.3 Regularity of the solution of Helmholtz problem for periodic grating ..... 52
3.1.2 Case 2A: Transmitting dielectric grating: TE case ..... 57
3.1.2.1 The inhomogeneous Helmholtz equation ..... 57
3.1.2.2 $\quad$ The $\alpha$-quasi-periodic Green functions of the Helmholtz equation
3.1.2 C
3.1.2.3 Regularity of the solution of Helmholtz problem for periodic grating ..... 260
3.2 A special case ..... 267
3.2.1 Solution operator $N_{\beta^{n}}$
3.2.2 First case: $k^{2}>n_{\alpha}^{2}$ ..... 69
3.2.2.1 Properties of $U^{\left(n_{\alpha}\right)}$ ..... 70
3.2.3 Second case: $k^{2}<n_{\alpha}^{2}$ ..... 73
3.2.4 Regularity of the solution corresponding to the special case ..... 75
3.2.5 Summary ..... 76
4 A priori error estimates using the $\alpha, 0$-quasi periodic transforma- tion ..... 78
4.1 Restatement of the boundary value problems for the periodic solution ..... 79
4.1.1 Variational formulation ..... 81
4.1.2 Equivalence of the variational forms for the periodic and $\alpha$ - quasi periodic problems ..... 82
4.1.3 Well posedness of the variational problem ..... 84
4.1.3.1 Existence and uniqueness of the solution ..... 84
4.1.3.2 Regularity estimate of the exact solution ..... 87
4.1.4 Efficiency of the diffraction grating ..... 91
4.1.5 Conservation of the energy or energy balance ..... 93
4.2 The discrete problem ..... 94
4.2.1 Variational formulation ..... 94
4.2.2 Truncation of the DtN map ..... 95
4.2.3 Efficiency of the grating ..... 97
4.2.4 Checking the energy balance using the truncated DtN map ..... 97
4.3 A priori error estimates for the exact solution ..... 98
4.3.1 A priori error estimation of the discretised problem ..... 98
4.3.2 A priori error estimation of the continuous problem by trun- cating the DtN operator ..... 105
4.3.3 Total error made by solving numerically the problem ..... 109
4.4 Summary ..... 112
5 An a posteriori error estimate using the dual weighted residual method and the $\alpha, 0$-quasi periodic transformation ..... 114
5.1 Introduction ..... 115
5.1.1 Global error estimate ..... 116
5.1.2 Goal oriented error estimate ..... 117
5.2 Direct problem ..... 119
5.2.1 Continuous problem ..... 119
5.2.2 Discretised problem ..... 120
5.3 Dual problem ..... 120
5.3.1 Quantity of interest: Q ..... 120
5.3.2 Strong form of the dual problem ..... 122
5.3.3 Variational form ..... 124
5.4 A posteriori error estimation ..... 124
5.4.1 Continuous problem ..... 124
5.4.2 Discretised problem ..... 125
5.4.3 Error estimation ..... 125
5.5 Summary and conclusion ..... 127
5.5.1 Summary ..... 127
5.5.2 Conclusion ..... 127
6 New formulation of the grating problem and a priori error esti- mates using the $\alpha, \beta$-quasi periodic transformation ..... 129
6.1 Restatement of the boundary value problem for the periodic solution ..... 130
6.1.1 Variational formulation ..... 132
6.1.2 Well posedness of the variational formulation ..... 133
6.1.2.1 Existence and uniqueness ..... 133
6.1.2.2 Regularity of the solution ..... 136
6.1.3 Efficiency of the grating ..... 138
6.2 Discrete problem ..... 140
6.2.1 Variational formulation ..... 140
6.2.2 Truncation of the DtN map ..... 140
6.2.3 Efficiency of the grating ..... 142
6.3 A priori error estimates for the exact solution $U_{\alpha, \beta}$ ..... 142
6.3.1 Dual problem ..... 143
6.3.2 An a priori error estimation of the discretised problem ..... 144
6.3.3 Comparison between the continuous variational formulation with a truncated DtN map and that with a full DtN map ..... 149
6.3.4 An a priori error estimate from the truncation of the DtN operators ..... 150
6.3.5 Estimate of the total error ..... 153
6.4 Summary and conclusion ..... 155
6.4.1 Summary ..... 155
6.4.2 Conclusion ..... 155
7 Numerical methods and numerical results ..... 158
7.1 Numerical methods ..... 158
7.1.1 Uniform mesh ..... 158
7.1.2 Adaptive mesh ..... 162
7.2 Numerical results ..... 163
7.2.1 Code validation ..... 163
7.2.2 Variation with $k$ of the $\alpha, 0$ and the $\alpha, \beta$-quasi periodic meth- ods ..... 175
7.2.3 Comparison with the DWR method ..... 181
7.2.4 Application of the DWR to a scatterer with a complex ge- ometry: the butterfly wings ..... 185
7.3 Summary and Conclusion ..... 188
8 Conclusion ..... 190
A Sobolev spaces ..... 203
A. 1 Some useful definitions ..... 203
A. 2 Quasi-periodicity ..... 207
B The Finite element method ..... 210
B. 1 Variational formulations of elliptic boundary value problems ..... 210
B. 2 Analysis background ..... 211
B. 3 Finite element space ..... 213
C A regularity estimate for the inhomogeneous Helmholtz problem for periodic gratings for Case 1B, Case 2A and Case 2B ..... 214
C. 1 General case ..... 214
C.1.1 The inhomogeneous Helmholtz equation ..... 214
C.1.2 The $\alpha$-quasi-periodic Green functions of the Helmholtz equa- tion ..... 216
C.1.3 Regularity of the solution of Helmholtz problem for periodic grating ..... 217
D A priori error estimates using the $\alpha, 0$
in medical x-ray imaging [62, 132]. There are also acousto-optic devices, which are dynamic gratings used to make tunable optical filters. They are used in many applications for example in the pharmaceutical industry, to characterize the composition of drugs, and in spectroscopic sensing, for detecting trace gases [108].

### 1.1 Background

The problem of electromagnetic wave diffraction is based on solving Maxwell's equations in the diffraction grating region, to find the resulting electromagnetic field when an incident wave shines upon the grating [79, 95]. In many applications we are more interested in the efficiency of the grating. The reflection (transmission) efficiency is the ratio of the reflected energy (transmitted energy) to the incident energy of the electromagnetic field. In our case, we will assume that there is neither charge nor current so that the Maxwell equations reduce to the Helmholtz equation [95]. There are two types of grating known as the perfectly conducting (sound soft) grating, when there is no energy absorbed by the grating (i.e the electromagnetic field vanishes inside the scatterer), and the dielectric transmitting (sound hard) grating, where energy can be absorbed by the grating. There are also two fundamental polarizations [82, 95, 50]. The first is the Transverse Electric mode (TE), where the direction of the electric field is perpendicular to the direction of propagation; which means the electric field is $E=\left(0,0, E_{z}(x, y, t)\right)$ and the magnetic field is $H=\left(H_{x}(x, y, t), H_{y}(x, y, t), 0\right)$. The second polarization is the Transverse Magnetic mode (TM), where the direction of the magnetic field is perpendicular to the direction of propagation; which means the magnetic field is $H=\left(0,0, H_{z}(x, y, t)\right)$ and the electric field is $E=\left(E_{x}(x, y, t), E_{y}(x, y, t), 0\right)$. We seek to solve the Helmholtz equation for both grating types and for both polarizations, in two dimensions, for a periodic grating with respect to one direction when a plane wave is incident. To this end, we use a finite element method and use the periodicity of the grating with respect to one direction to focus our study over one period. Since the domain is infinite in the other direction we need to truncate the domain and apply some transparent boundary conditions. We then use the Rayleigh expansion of the electromagnetic fields from the region outside the truncated domain of the diffraction grating system to match with the solution inside the diffraction grating region. We then subdivide the diffraction grating region into finite elements.

There are of course other numerical methods in the literature to solve the problem of diffraction of waves [95, 99]. There is the differential method where the dielectric permittivity and the electromagnetic field are expanded in a Fourier series over the grating period [95]. Then the Fourier expansions are inserted into the Helmholtz equation and the problem is solved using differential equation shooting methods. There is also the boundary element method [91], where the wave problem is formulated as an integral equation. This method is deployable when we can derive a periodic Green's function which describes the connection of the
electrical surface current radiating from one point belonging to the grating to an arbitrary point in the space. Then there is the finite difference method [65], which is based on solving the wave equation in the diffraction grating region by dividing the diffraction grating over one period into grids. The spatial derivatives are then approximated by finite differences. Like the finite element method, it uses the Rayleigh expansion of the electromagnetic fields from the region outside the diffraction grating system to match on the boundary the electromagnetic field from the region inside.

The advantage of the finite element method is its flexibility in dealing with complex geometries. It also naturally gives rise to a variational formulation which provides a platform to rigorously derive existence and uniqueness results and regularity bounds. This allows us to make statements about the well-posedness of the problem and to derive a priori error estimates. In addition, the finite element method can achieve a desired level of accuracy and at the same time minimize the computational cost by applying a posteriori error estimates.

### 1.2 Outline of the thesis

The first aim of this work is to undertake a rigorous analysis of an a priori error estimate with what we call the $\alpha, 0$-quasi periodic method when we use the finite element method and the Rayleigh expansion to solve the Helmholtz problem. This original work is contained in Chapters 3 and 4.

In applications, we are also interested in controlling efficiently the error that arises when we solve numerically the scattering problem. This can be done in the finite element case, by calculating a posteriori error estimates. Since we are interested in the efficiency of the grating we then choose the efficiency as our quantity of interest. The second aim of our work is therefore to investigate a goal oriented a posteriori error estimate called the Dual Weighted Residual (DWR) method. This method consists of using both the direct solution and the dual solution of the problem in order to achieve a goal oriented result. This original work is contained in Chapter 5.

We also want to investigate different formulations of the problem, still based on finite elements and the Rayleigh expansion, and examine in which cases these approaches will be more efficient in solving the diffraction problem. Since it has been reported for single scattering that there is an instability in numerical methods when we have high wavenumbers [59, 16], we write the solution of the scattering problem $U$ as a product of the analytical solution of the scattering problem when the domain is scatterer free and another unknown function we call $U_{\alpha, \beta}$. By doing so, we get a new wave equation that we will solve for this new unknown function rather than solving directly the Helmholtz problem for $U$. We call this the $\alpha, \beta$ quasi periodic method. Hence our third aim is to investigate the a priori error associated with this approach. The details of this original work are contained in Chapter 6.

Our final aim is to implement these methods in a discretised finite element code and investigate their stability, accuracy, dependency on system parameters, and their applicability to a set of problems. The details of this work are contained in Chapter 7.

As there are two types of wave to consider (transverse magnetic (TM) and transverse electric (TE)) and there are two types of gratings (perfectly conducting and transmitting dielectric) there are in fact four cases to investigate. We denote these cases by Case 1A/B: perfectly conducting and Case 2A/B: transmitting dielectric where $\mathrm{A}(\mathrm{B})$ denotes the $\mathrm{TE}(\mathrm{TM})$ wave. Whilst we have derived results for all four cases, in order to emphasise the salient points, we have relegated the analysis of Cases 1B, 2A and 2B to the appendices (in Chapter 3 we also give details for Case 2A).

We start by describing the geometry of the problem, in Chapter 2, when an incident wave is radiating on the grating profile $P=f(x)$ and we give the mathematical formulation of the problem. Due to the periodicity of the grating, we can focus our study over one period $d$. When we solve the problem numerically, we truncate the domain at $|y|=B<\infty$ by using the Dirichlet to Neumann (DtN) operators. The DtN operator is used to match the Fourier coefficients of the field obtained by finite elements inside this domain with the Rayleigh expansion of the field outside. Once we have formulated the problem mathematically, we study the properties of the DtN map. There have been a small number of theoretical investigations into the use of the Finite Element Method as a tool for studying the electromagnetic waves interacting with a diffraction grating [9, 12, 13]. In these studies the continuity of the DtN map was simply assumed and hence the dependency of the regularity constant on the system parameters such as the wavenumber was not derived. These are essential components of our analysis and so in Section 2.3.3 we derive these results for the first time. We then study the regularity estimate for the Helmholtz problem for a periodic grating in two dimensions in Chapter 3. This is needed in Chapters 4 and 6 to show the continuous dependence of the solution of the Helmholtz problem on the data. We then describe the $\alpha, 0$-quasi periodic method in Chapter 4 . We start with an examination of the continuous problem where we give its variational formulation in an appropriate Sobolev space. We then investigate the well-posedness of the continuous problem; that is, the solution exists, is unique and depends continuously on the data. Hence, a new regularity result is required in order to show this continuous dependence. The study of the well-posedness of the problem allows us to know in advance that we can solve the problem. We then discretise the problem to approximate its solution and derive a new a priori error estimate. This result guarantees the uniqueness of the approximate solution.

There are essentially two approaches to studying a posteriori error estimates. We can either study the difference between the exact solution and its approximate solution or we just choose to focus on a particular linear functional which depends on these solutions. In our case, we will adopt the second approach in Chapter 5. A
linear functional $Q$ is defined in such a way that it captures the error made between the efficiency of the grating in the continuous problem and that obtained from the discretised problem. By finding an upper bound for $Q$ we derive a formula for an error indicator function. This indicator function is subsequently used in the finite element implementation to adaptively refine the size of the individual elements. We start with the direct problem, then introduce the dual problem before proving an estimate of the linear functional of the error.

We also want to investigate different transformations of the problem, still based on finite elements and the Rayleigh expansion. To this end we introduce the $\alpha, \beta$ quasi periodic method in Chapter 6. Similar to Chapter 4, we start with an examination of the continuous problem where we give its variational formulation. We investigate the well-posedness of the continuous problem, discretise the problem to approximate its solution and derive a new a priori error estimate which guarantees the uniqueness of this approximate solution. We finish by comparing quantitatively the a priori error estimate from the $\alpha, 0$-quasi periodic method and the $\alpha, \beta$-quasi periodic method.

In order to validate the theoretical results that are derived in this thesis a series of numerical experiments were devised and implemented. There are two types of experiment to be performed. The first set of experiments involve solving the problem numerically using a standard implementation and confirming that a bounded solution does indeed exist. We will also use this implementation to examine the relative convergence of the two problem formulations given in Chapters 4 and 6. The second set of experiments utilise the new approach detailed in Chapter 5 to construct an adaptive grid implementation. The aim of these experiments is to verify that the proposed method does produce correct solutions and converges faster than alternative strategies. Details of these implementations and the corresponding results are contained in Chapter 7. We begin by validating our code with numerical methods and experiments from the literature. We then compare the $\alpha, \beta$-quasi periodic method with the $\alpha, 0$-quasi periodic method across a range of wavenumbers. We produce numerical results that show the advantage of using the Dual Weighted Residual (DWR) method as compared to the uniform mesh or the global a posteriori error method proposed in [13]. We have chosen the finite element method because of its flexibility to adapt to complex scattering geometry. In our final illustration, we will consider a scattering geometry from a real world application concerning the Morpho butterfly wing [98, 129, 124, 102].

## Chapter 2

## Physical and mathematical description of the problem

In this chapter, we start by introducing Maxwell's equations and show how these reduce to the Helmholtz equation in two simplified cases. Then, we give the geometrical description of the diffraction grating and the mathematical formulation of the problem concerning wave interaction with this. After we study the Dirichlet to Neumann map, we use this to construct a transparent boundary between an outer analytic solution and an inner numerical (finite element) solution so that we can truncate the domain of our problem. We then give the mathematical formulation of our problem in the truncated domain $\Omega \subset \mathbb{R}^{2}$.

### 2.1 Maxwell's equations and polarization

There are two types of grating [95, 82], the perfectly conducting (sound soft) grating when there is no energy absorbed by the grating (i.e the electromagnetic field vanishes inside the scatterer) and the transmitting dielectric (sound hard) grating where energy can be absorbed by the grating. In this latter case, the wavenumber inside the scatterer has a positive imaginary part which captures the energy loss. There are two fundamental polarizations [82, 95, 50]. The first is the Transverse Electric mode (TE), where the direction of the electric field is perpendicular to the direction of propagation; which means the electric field $E=\left(0,0, E_{z}\right)$ and the magnetic field $H=\left(H_{x}, H_{y}, 0\right)$. The second polarization is the Transverse Magnetic mode (TM), when the direction of the magnetic field is perpendicular to the direction of propagation; which means the magnetic field $H=$ $\left(0,0, H_{z}\right)$ and the electric field $E=\left(E_{x}, E_{y}, 0\right)$. Let us start by studying Maxwell's equations and derive the TE and TM problems for the perfectly conducting grating. After that we will deal with the transmitting dielectric grating.

The time harmonic form of Maxwell's equations, with $e^{-i w t}$ dependence, is
given by $[95,60]$

$$
\begin{align*}
\nabla \times \underline{E} & =-i w \underline{B}  \tag{2.1}\\
\nabla \times \underline{H} & =i w \underline{D}+\underline{J}  \tag{2.2}\\
\nabla \cdot \underline{J} & =-i w \rho  \tag{2.3}\\
\nabla \cdot \underline{D} & =\rho  \tag{2.4}\\
\nabla \cdot \underline{B} & =0 \tag{2.5}
\end{align*}
$$

where $\underline{E}$ is the electric field, $\underline{H}$ is the magnetic field, $\underline{B}$ is the magnetic flux density, $\underline{D}$ is the electric displacement, $\underline{J}$ is the electric current density, $w$ is the angular frequency, and $\rho$ is the electric charge density. The constitutive relations may be written

$$
\begin{align*}
& \underline{D}=\varepsilon \underline{E}  \tag{2.6}\\
& \underline{B}=\mu \underline{H} \tag{2.7}
\end{align*}
$$

where $\varepsilon$ and $\mu$ are, respectively, the electric permittivity and magnetic permeability of the medium. At this stage let us assume that $\mu=1$ and that there is no charge and no current, so that $\rho=0$ and $\underline{J}=0$.

## - Case 1: Perfectly conducting grating

For the perfectly conducting case, in the domain exterior to the grating, equations (2.6) and (2.7) hold, where the electric permittivity of the medium $\varepsilon$ is constant. Inserting equation (2.7) into equation (2.1) gives

$$
\begin{equation*}
\nabla \times \underline{E}=-i w \mu \underline{H} . \tag{2.8}
\end{equation*}
$$

Similarly if we use equation (2.6) in equation (2.2) we get

$$
\begin{equation*}
\nabla \times \underline{H}=i w \varepsilon \underline{E} . \tag{2.9}
\end{equation*}
$$

Introducing the components of each field via $\underline{E}=\left(E_{x}, E_{y}, E_{z}\right)$ and $\underline{H}=$ $\left(H_{x}, H_{y}, H_{z}\right)$ then, if we assume that $\underline{E}(x, y)$ and $\underline{H}(x, y)$, we have from equation (2.8)

$$
\begin{align*}
\partial_{y} E_{z} & =-i w \mu H_{x}  \tag{2.10}\\
\partial_{x} E_{z} & =i w \mu H_{y}  \tag{2.11}\\
\partial_{x} E_{y}-\partial_{y} E_{x} & =-i w \mu H_{z} \tag{2.12}
\end{align*}
$$

and from equation (2.9)

$$
\begin{align*}
\partial_{y} H_{z} & =i w \varepsilon E_{x}  \tag{2.13}\\
\partial_{x} H_{z} & =-i w \varepsilon E_{y}  \tag{2.14}\\
\partial_{x} H_{y}-\partial_{y} H_{x} & =i w \varepsilon E_{z} . \tag{2.15}
\end{align*}
$$

The perfectly conducting boundary conditions for the fields on the grating surface are given by

$$
\begin{array}{r}
\underline{n} \times \underline{E}=\underline{0} \\
\underline{n} \cdot \underline{H}=0 \tag{2.17}
\end{array}
$$

where $\underline{n}=\left(n_{x}, n_{y}, 0\right)$ is the outward unit normal vector on the grating surface [82]. Examining the components of equation (2.16) we get on the boundary

$$
\begin{equation*}
E_{z}=0, \tag{2.18}
\end{equation*}
$$

and

$$
\begin{equation*}
n_{x} E_{y}-n_{y} E_{x}=0 \tag{2.19}
\end{equation*}
$$

Combining equation (2.19) with equations (2.13) and (2.14) gives

$$
\begin{equation*}
n_{x} \partial_{x} H_{z}+n_{y} \partial_{y} H_{z}=\frac{\partial H_{z}}{\partial n}=0 \tag{2.20}
\end{equation*}
$$

where $\partial_{x}$ is the first order partial derivative with respect to $x$ ( $\partial_{y}$ is the first order partial derivative with respect to $y$ ). We note that equations (2.10) to (2.20) can be separated into two independent sets [82]. The first set consists of (2.10), (2.11), (2.15) and (2.18). These equations only contain the transverse component $E_{z}$ of the electric field and the components $H_{x}$ and $H_{y}$ of the magnetic field. Differentiating equation (2.10) with respect to $y$, differentiating equation (2.11) with respect to $x$ and adding them gives

$$
\Delta E_{z}=i w \mu\left(\partial_{x} H_{y}-\partial_{y} H_{x}\right)
$$

Using equation (2.15) then gives the Helmholtz equation

$$
\begin{equation*}
\Delta E_{z}+k^{2} E_{z}=0 \tag{2.21}
\end{equation*}
$$

where

$$
k^{2}=w^{2} \varepsilon \mu,
$$

subject to the boundary condition given by equation (2.18). This is called the Transverse Electric (TE) problem. The second set consists of (2.12), (2.13), (2.14) and (2.20). These equations only contain the transverse component $H_{z}$ of the magnetic field and the components $E_{x}$ and $E_{y}$ of the electric field. In a similar way if we combine equations (2.13) and (2.14) and use equation (2.12) we get a second Helmholtz problem

$$
\begin{equation*}
\Delta H_{z}+k^{2} H_{z}=0 \tag{2.22}
\end{equation*}
$$

subject to the boundary condition given by equation (2.20). This is called the Transverse Magnetic (TM) problem.

## - Case 2. Transmitting dielectric grating

We follow the same procedure as in Case 1 but this time we also have to satisfy Maxwell's equations inside the scatterer, and so we need to take into account that the permittivity $\varepsilon$ is no longer constant. For the TE case the combination of equations (2.10), (2.11) and (2.15) (with permittivity $\varepsilon(x, y)$ ) lead to the Helmholtz problem

$$
\begin{equation*}
\Delta E_{z}+k^{2}(x, y) E_{z}=0 \tag{2.23}
\end{equation*}
$$

with

$$
k^{2}(x, y)=w^{2} \varepsilon(x, y) \mu
$$

For the TM case a complication arises when we take the derivatives of equations (2.13) and (2.14) since the permittivity is now spatially dependent. Hence equations (2.13) and (2.14) can be rearranged and differentiated to give

$$
\partial_{y}\left(\frac{1}{i w \varepsilon(x, y)} \partial_{y} H_{z}\right)=\partial_{y} E_{x}
$$

and

$$
\partial_{x}\left(\frac{1}{i w \varepsilon(x, y)} \partial_{x} H_{z}\right)=-\partial_{x} E_{y} .
$$

Adding these and using equation (2.12) gives

$$
\begin{equation*}
\partial_{x}\left(\frac{1}{k^{2}(x, y)} \partial_{x} H_{z}\right)+\partial_{y}\left(\frac{1}{k^{2}(x, y)} \partial_{y} H_{z}\right)+H_{z}=0 \tag{2.24}
\end{equation*}
$$

since the permeability is constant.
For both Case 1 and Case 2, in order to formulate the scattering problem as a boundary value problem, we need to include an appropriate radiation condition (outgoing wave condition). In this thesis we will be considering electromagnetic waves interacting with a periodic diffraction grating. The usual Sommerfeld radiation condition is therefore not appropriate [28] as the radiating energy does not diminish in the direction of periodicity. Hence the so-called upward propagating radiation condition (UPRC) was introduced in [28]. It has been used to establish the uniqueness of the solution of scattering from a periodic grating. It was shown that if the grating is periodic then the UPRC is equivalent to the Rayleigh condition which means when $y$ goes to infinity, $E_{z}$ and $H_{z}$ must remain bounded and can be described as a superposition of outgoing plane waves $[97,95,4,5]$.

### 2.2 Problem statement

We wish to solve the Helmholtz equation in $\mathbb{R}^{2}$ for a periodic grating of period $d$ (with respect to $x$ ), as shown in Figure 2.1. Since we will consider two types of


Figure 2.1: Diagram representing an incident wave with an angle of incidence $\theta$ with respect to the $y$ axis on a periodic grating in $\mathbb{R}^{2}$. The parameter $d$ represents the period of the grating in the $x$ direction. The scatterers are represented by the shaded regions.
grating (the perfectly conducting grating and the transmitting dielectric grating) we denote, for the perfectly conducting case, the region of $\mathbb{R}^{2}$ outside the scatterers by $\mathbb{R}_{+}^{2}$ (the unshaded region in Figure 2.1) and by $\mathbb{R}_{-}^{2}$, the scatterers in $\mathbb{R}^{2}$ (shown as the shaded regions in Figure 2.1).

## - Case 1A: Perfectly conducting grating: TE case

Here we are solving for $U=E_{z}$ and, using equations (2.21) and (2.16), we solve the Helmholtz problem to find $U(x, y) \in C^{2}\left(\mathbb{R}_{+}^{2}\right)$ such that

$$
\begin{align*}
\Delta U(x, y)+k^{2}(x, y) U(x, y) & =0, & & (x, y) \in \mathbb{R}_{+}^{2} \\
U(x, y) & =0, & (x, y) & \in \partial \mathbb{R}_{-}^{2}, \tag{2.25}
\end{align*}
$$

subject to the upward propagating radiation condition (UPRC)

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{2.26}
\end{equation*}
$$

## - Case 1B: Perfectly conducting grating: TM case

Here we are solving for $U=H_{z}$ and, using equations (2.22) and (2.17), we solve the Helmholtz problem to find $U(x, y) \in C^{2}\left(\mathbb{R}_{+}^{2}\right)$ such that

$$
\begin{align*}
\Delta U(x, y)+k^{2}(x, y) U(x, y) & =0, & & (x, y) \in \mathbb{R}_{+}^{2} \\
\frac{\partial U(x, y)}{\partial n} & =0, & & (x, y) \in \partial \mathbb{R}_{-}^{2} \tag{2.27}
\end{align*}
$$

subject to the UPRC

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{2.28}
\end{equation*}
$$

where $\frac{\partial}{\partial n}$ denotes the normal derivative operator on the boundary of each scatterer.

## - Case 2A: Transmitting dielectric grating: TE case

Here we are solving for $U=E_{z}$ and, using equations (2.23), we solve the Helmholtz problem to find $U(x, y) \in C^{2}\left(\mathbb{R}^{2}\right)$ such that

$$
\begin{equation*}
\Delta U(x, y)+k^{2}(x, y) U(x, y)=0, \quad(x, y) \in \mathbb{R}^{2} \tag{2.29}
\end{equation*}
$$

subject to the UPRC

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{2.30}
\end{equation*}
$$

## - Case 2B: Transmitting dielectric grating: TM case

Here we are solving for $U=H_{z}$ and, using equation (2.24), we solve the Helmholtz problem to find $U(x, y) \in C^{2}\left(\mathbb{R}^{2}\right)$ such that

$$
\begin{equation*}
\nabla \cdot\left(\frac{1}{k^{2}} \nabla U(x, y)\right)+U(x, y)=0, \quad(x, y) \in \mathbb{R}^{2} \tag{2.31}
\end{equation*}
$$

subject to the UPRC

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{2.32}
\end{equation*}
$$

We will utilize the periodicity of the grating and restrict our problem to a single vertical strip, $S=[0, d] \times \mathbb{R}$, as shown in Figure 2.2. We restrict the effects of the scatterers (the grating) to a horizontal strip $\Omega_{0}=[0, d] \times[-b, b]$ and we define the wavenumber $k$ to be

$$
k(x, y)= \begin{cases}k_{1} \in \mathbb{R}, & \text { for }(x, y) \in \Omega_{1}  \tag{2.33}\\ k_{0}(x, y) \in \mathbb{C}, & \text { for }(x, y) \in \Omega_{0} \\ k_{2} \in \mathbb{C}, & \text { for }(x, y) \in \Omega_{2}\end{cases}
$$

where $\Omega_{1}=[0, d] \times[b,+\infty)$ and $\Omega_{2}=[0, d] \times(-\infty,-b]$. The first reason for having $b$ is for computational efficiency, so that we can have coarse mesh inside the region away from the scatterers $(|y|>b)$. The second reason is that we need to use the Rayleigh expansion inside the region $b<|y|<B$ when we derive an a priori error estimates when we truncate the DtN map. The third reason is to make our geometry sufficiently general to cope with the situation where there is a different material in regions $\Omega_{i}$.

## Incident wave

The incident wave is

$$
\begin{equation*}
U_{I}=e^{i \alpha x-i \beta_{1}^{0} y} \tag{2.34}
\end{equation*}
$$

where

$$
\begin{align*}
\alpha & =k_{1} \sin \theta,  \tag{2.35}\\
\beta_{1}^{0} & =k_{1} \cos \theta
\end{align*}
$$

and $\theta$ is the angle of incidence of the wave as shown in Figure 2.2. We demand that

$$
\left.\begin{array}{l}
\mathfrak{R}\left(k_{j}\right)>0,  \tag{2.36}\\
\mathfrak{J}\left(k_{j}\right) \geq 0,
\end{array}\right\}
$$

where $\mathfrak{R}\left(k_{j}\right)\left(\mathfrak{J}\left(k_{j}\right)\right)$ denotes the real (imaginary) part of $k_{j}$ so that the scattered and diffracted waves are composed of bounded outgoing waves. We use the notation $\beta_{1}^{0}$ to be consistent with the notation $\beta_{j}^{n}$ that we will use later on for the


Figure 2.2: Diagram depicting a periodic grating in $\mathbb{R}^{2}$ over one period $d$ in the $x$ direction. The scattering region is denoted by $\Omega_{0}$, the region occupied by the scatterer is denoted by $\Omega_{3}, \Omega_{1}$ is the upper region with constant wavenumber $k_{1} \in \mathbb{R}$ and $\Omega_{2}$ is the lower region with constant wavenumber $k_{2} \in \mathbb{C}$. $U_{I}$ is the incident wave.
different wavenumbers of the diffracted waves in the upper region $\Omega_{1}(j=1)$ and lower region $\Omega_{2}(j=2)$. The reason that there is no subscript or superscript on $\alpha$ is just a consequence of the Snell's law for stratified media which makes $\alpha$ the same in $\Omega_{1}$ and $\Omega_{2}$.

## Transparent boundary conditions (Dirichlet to Neumann Maps)

To solve the grating problem numerically for a wide range of grating geometries we will use a finite element method. We therefore need to truncate our domain to render it finite. To provide suitable boundary conditions for the finite element solver we use an analytical solution in the adjacent domains and apply transparent boundary conditions that match this analytical solution continuously and smoothly with the finite element solution inside the truncated region. To achieve this we study the analytic solution of the Helmholtz equation when the wavenumber is a complex constant; known as the Rayleigh expansion. These transparent boundary conditions are captured by the Dirichlet to Neumann (DtN) operators $T_{ \pm}$(see Section 2.3.2) which match the Rayleigh expansion of the electromagnetic field on the boundary of the truncated region with the finite element solution inside the truncated domain. To assist us in this we redefine $\Omega_{1}, \Omega_{2}$ to be

$$
\begin{array}{ll}
\Omega_{1}=\{(x, y): & 0 \leq x \leq d, b \leq y \leq B\} \\
\Omega_{2}=\{(x, y): & 0 \leq x \leq d,-B<y \leq-b\},
\end{array}
$$

where $B$ is a positive real number and $B>b$ (see Figure 2.3). We also denote

$$
\begin{array}{lll}
\Omega_{+}=\{(x, y): & 0 \leq x \leq d, & y \geq B\} \\
\Omega_{-}=\{(x, y): & 0 \leq x \leq d, & y \leq-B\}, \\
\Gamma_{+}=\{(x, y): & 0 \leq x \leq d, & y=B\} \\
\Gamma_{-}=\{(x, y): & 0 \leq x \leq d, & y=-B\}, \tag{2.38}
\end{array}
$$

and we denote by $\Omega_{3} \subset \Omega_{0}$ the scatterer (see Figure 2.3). In our study, the structure of the diffraction grating in one period can be shown either by Figure 2.3 or Figure 2.4. Then instead of finding $U$ which satisfies one of the equations (2.25), (2.27), (2.29) and (2.31) in $[0, d] \times \mathbb{R}$ we will find the equivalent solution of the problem in

$$
\begin{equation*}
\Omega=\{(x, y): 0 \leq x \leq d,-B \leq y \leq B\} . \tag{2.39}
\end{equation*}
$$

Hence we need an analytical solution to the Helmholtz problem in the homogeneous regions that lie above and below $\Omega$ so that we can solve our problem in $\Omega$, this is done in the following section.


Figure 2.3: Diagram showing the truncated periodic grating domain. We redefine $\Omega_{1}$, the region above the scattering region to be $\{(x, y): 0 \leq x<d, b \leq y \leq B\}$, and the substrate $\Omega_{2}$ to be $\{(x, y): 0 \leq x<d,-B \leq y \leq-b\}$.


Figure 2.4: Another type of the grating profile when the interface which separates the grating with the region above is an open curve.

### 2.3 Matching the analytical solution on to the finite element solution at the transparent boundary

As explained earlier we will need an analytical solution to the Helmholtz problem in the homogeneous regions that lie above and below $\Omega$ defined in (2.39). In the following section we derive this solution.

### 2.3.1 Fundamental solution of Helmholtz problem in a homogeneous domain

In Section 2.2 the methodology for solving these Helmholtz problems for a general scatterer geometry required an analytical solution in the outer domains; where the wavenumber is constant. Let $u \in H_{\alpha \#}^{1}\left(\Omega_{ \pm}\right) \bigcap C^{2}\left(\Omega_{ \pm}\right), k_{j} \in \mathbb{C}$ and let us find the analytical solution of

$$
\Delta u+k_{j}^{2} u=0, \quad x \in[0, d],|y| \geq B, \text { i.e. } x \in \Omega_{ \pm}, \quad j \in\{1,2\} .
$$

For that, we need the following lemma. For completeness sake Appendix A contains the definitions that we shall need in this thesis regarding Sobolev spaces.

Lemma 1. If u is $\alpha$-quasi-periodic and if $k_{j} \in \mathbb{C}$, then the solution of the Helmholtz problem

$$
\begin{equation*}
\Delta u(x, y)+k_{j}^{2} u(x, y)=0, \quad x \in[0, d],|y| \geq B, \quad j \in\{1,2\} \tag{2.40}
\end{equation*}
$$

is composed only of outgoing waves (apart from the incident wave). In order to satisfy the UPRC conditions at $y=\infty$ so that it is unique, it can be written as [97, 95, 4, 5]

$$
\begin{equation*}
u(x, y)=\sum_{n \in \mathbb{Z}} F_{n}(x) G_{n}(y) \tag{2.41}
\end{equation*}
$$

where

$$
\begin{aligned}
F_{n}(x) & =c_{1} e^{i n_{\alpha} x}, \\
G_{n}(y) & =c_{n} e^{-i \beta_{j}^{n} y}+d_{n} e^{i \beta_{j}^{n} y} .
\end{aligned}
$$

For $y \geq B, c_{n}=0$ for all $n \in \mathbb{Z}$ except for $n=0$ which corresponds to the incident wave and for $y \leq-B d_{n}=0$ for all $n \in \mathbb{Z}$ in order to satisfy the radiation condition (UPRC). We also have

$$
\begin{align*}
n_{\alpha} & =\alpha+\frac{2 \pi n}{d}  \tag{2.42}\\
\left|\beta_{j}^{n}\right|^{2} & =\left|k_{j}^{2}-n_{\alpha}^{2}\right|, \quad \beta_{j}^{n}=e^{i z_{n} / 2}\left(\left|k_{j}^{2}-n_{\alpha}^{2}\right|\right)^{1 / 2} \tag{2.43}
\end{align*}
$$

such that

$$
\begin{equation*}
z_{n}=\arg \left(k_{j}^{2}-n_{\alpha}^{2}\right), \tag{2.44}
\end{equation*}
$$

for $j \in\{1,2\}$.
Proof. Let us write $u=\sum_{n \in \mathbb{Z}} F_{n}(x) G_{n}(y)$. If $u$ satisfies equation (2.40) then so does each $u_{n}=F_{n}(x) G_{n}(y)$. Therefore,

$$
\begin{equation*}
\Delta u_{n}+k_{j}^{2} u_{n}=0, \quad x \in[0, d],|y| \geq B, \quad j \in\{1,2\} . \tag{2.45}
\end{equation*}
$$

Then, equation (2.45) becomes

$$
F_{n}^{\prime \prime}(x) G_{n}(y)+F_{n}(x) G_{n}^{\prime \prime}(y)+k_{j}^{2} F_{n}(x) G_{n}(y)=0
$$

which can be rewritten as follows

$$
\begin{equation*}
\frac{F_{n}^{\prime \prime}(x)}{F_{n}(x)}=-\frac{G_{n}^{\prime \prime}(y)}{G_{n}(y)}-k_{j}^{2} . \tag{2.46}
\end{equation*}
$$

We see that the left and right hand sides of equation (2.46), are dependent only on $x$, and on $y$, respectively. Therefore, they both must be constant and so there exists a constant $\lambda_{n}$ such that

$$
\begin{align*}
& F_{n}^{\prime \prime}(x)=\lambda_{n} F_{n}(x)  \tag{2.47}\\
& G_{n}^{\prime \prime}(y)=-\left(\lambda_{n}+k_{j}^{2}\right) G_{n}(y) . \tag{2.48}
\end{align*}
$$

Let us begin by solving equation (2.47). Because of the quasi-periodicity of $F_{n}$, we cannot have $\lambda_{n}>0$ so we write $\lambda_{n}=-n_{\alpha}^{2}$. Therefore,

$$
F_{n}(x)=C_{1} e^{-i n_{\alpha} x}+C_{2} e^{i n_{\alpha} x} .
$$

Applying the quasi-periodic condition

$$
\begin{equation*}
F_{n}(d)=e^{i \alpha d} F_{n}(0) \tag{2.49}
\end{equation*}
$$

gives

$$
\left(C_{1}+C_{2}\right) e^{i \alpha d}=C_{1} e^{-i n_{\alpha} d}+C_{2} e^{i n_{\alpha} d}
$$

and so equation (2.49) is satisfied if and only if

$$
\begin{equation*}
n_{\alpha}=\alpha+\frac{2 \pi n}{d} \tag{2.50}
\end{equation*}
$$

and

$$
C_{1}=0 .
$$

We choose the direction of propagation along the $x$ direction to be from left to right and so

$$
\begin{equation*}
F_{n}(x)=C_{2} e^{i n_{\alpha} x} . \tag{2.51}
\end{equation*}
$$

Equation (2.48) can be solved using equation (2.50) to give

$$
-G_{n}^{\prime \prime}(y)-\left(k_{j}^{2}-n_{\alpha}^{2}\right) G_{n}(y)=0
$$

Then, with $\left(\beta_{j}^{n}\right)^{2}=k_{j}^{2}-n_{\alpha}^{2}$, as given by equation (2.43),

$$
\begin{equation*}
G_{n}(y)=c_{n} e^{-i \beta_{j}^{n} y}+d_{n} e^{i \beta_{j}^{n} y} . \tag{2.52}
\end{equation*}
$$

Note that we avoid the case $k_{j}^{2}=n_{\alpha}^{2}$ which corresponds to the resonance phenomenon [95, 82].

### 2.3.2 The Dirichlet to Neumann maps $T, T^{\alpha, 0}$ and $T^{\alpha, \beta}$

We need the following property of $\alpha$-quasi-periodic functions before we study the DtN maps.

Lemma 2. For fixed $y=y_{0}, f\left(x, y_{0}\right)$ can be treated as a function of one variable, $x$. If $f\left(x, y_{0}\right) \in L_{\alpha \#}^{1}([0, d])$, then $f\left(x, y_{0}\right)$ can be expanded as the Fourier series

$$
\begin{equation*}
f\left(x, y_{0}\right)=\sum_{n \in \mathbb{Z}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right) e^{i n_{\alpha} x} \tag{2.53}
\end{equation*}
$$

where

$$
f^{\left(n_{\alpha}\right)}\left(y_{0}\right)=\frac{1}{d} \int_{0}^{d} f\left(x, y_{0}\right) e^{-i n_{\alpha} x} d x
$$

with $n_{\alpha}=\frac{2 \pi n}{d}+\alpha$.
Proof. Because $f\left(x, y_{0}\right)$ is $\alpha$-quasi-periodic with respect to $x$, we can use Lemma A16 and we can see that $g\left(x, y_{0}\right)=e^{-i \alpha x} f\left(x, y_{0}\right)$ is periodic with respect to $x$. Therefore we can express $g\left(x, y_{0}\right)$ as a Fourier series

$$
e^{-i \alpha x} f\left(x, y_{0}\right)=\sum_{n \in \mathbb{Z}} g^{(n)}\left(y_{0}\right) e^{i \frac{i \pi n}{d} x}
$$

such that

$$
\begin{aligned}
g^{(n)}\left(y_{0}\right) & =\frac{1}{d} \int_{0}^{d} e^{-i \alpha x} f\left(x, y_{0}\right) e^{-i \frac{2 \pi n}{d} x} d x, \\
& =\frac{1}{d} \int_{0}^{d} f\left(x, y_{0}\right) e^{-i n_{\alpha} x} d x, \\
& =f^{\left(n_{\alpha}\right)}\left(y_{0}\right) .
\end{aligned}
$$

Hence

$$
e^{-i \alpha x} f\left(x, y_{0}\right)=\sum_{n \in \mathbb{Z}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right) e^{i \frac{2 \pi n}{d} x}
$$

which finishes the proof.

We can now expand the solution to the Helmholtz problem detailed in Lemma 1, $u(x, y)$, as a Fourier series in $x$ (for each $y$ value), by using Lemma 2. This leads us to

$$
\begin{equation*}
u(x, y)=\sum_{n \in Z} U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} \tag{2.54}
\end{equation*}
$$

By using Lemma 1 for $(x, y) \in \Omega_{ \pm}$, this solution can be written as

$$
\begin{equation*}
u(x, y)=\sum_{n \in \mathbb{Z}} r_{j}^{n} e^{i \beta_{j}^{n} y+i n_{\alpha} x}+t_{j}^{n} e^{-i \beta_{j}^{n} y+i n_{\alpha} x}, j=1,2 \tag{2.55}
\end{equation*}
$$

where the unknowns $r_{j}^{n}$ and $t_{j}^{n}$, called Rayleigh coefficients, are complex scalars (we shall see later in Section 4.1.4 that $r_{j}^{n}\left(t_{j}^{n}\right)$ is related to the grating reflection (transmission) efficiency of order $n$ ). By requiring that $u$ is composed of bounded outgoing planes waves in $\Omega_{ \pm}$, except the incident wave in $\Omega_{+}$, we have

$$
\begin{aligned}
t_{1}^{n} & =\delta_{n 0} \\
r_{2}^{n} & =0
\end{aligned}
$$

for all $n \in \mathbb{Z}$ such that $\beta_{j}^{n}$ is given by equation (2.43) and $\delta$ is the Kronecker delta defined by

$$
\delta_{n m}= \begin{cases}1, & \text { if } n=m  \tag{2.56}\\ 0, & \text { otherwise }\end{cases}
$$

We again stipulate that $k_{j}^{2} G n_{\alpha}^{2}$ to avoid the phenomenon of resonance [95]. This leads us to write

$$
\begin{array}{lc}
U^{\left(n_{\alpha}\right)}(y)=r_{1}^{n} e^{i \beta_{1}^{n} y}+\delta_{n 0} e^{-i \beta_{1}^{0} y}, & \text { in } \Omega_{+} \\
U^{\left(n_{\alpha}\right)}(y)=t_{2}^{n} e^{-i \beta_{2}^{n} y}, & \text { in } \Omega_{-} \tag{2.57}
\end{array}
$$

In $\Omega_{+}$, since $\Gamma_{+} \subset \Omega_{+}$, when $y=B$, we have

$$
U^{\left(n_{\alpha}\right)}(B)=r_{1}^{n} e^{i \beta_{1}^{n} B}+\delta_{n 0} e^{-i \beta_{1}^{0} B}
$$

so

$$
\begin{equation*}
r_{1}^{n}=U^{\left(n_{\alpha}\right)}(B) e^{-i \beta_{1}^{n} B}-\delta_{n 0} e^{-2 i \beta_{1}^{0} B} . \tag{2.58}
\end{equation*}
$$

In $\Omega_{-}$, since $\Gamma_{-} \subset \Omega_{-}$, when $y=-B$,

$$
U^{\left(n_{\alpha}\right)}(-B)=t_{2}^{n} e^{i \beta_{2}^{n} B}
$$

so

$$
\begin{equation*}
t_{2}^{n}=U^{\left(n_{\alpha}\right)}(-B) e^{-i \beta_{2}^{n} B} \tag{2.59}
\end{equation*}
$$

Inserting the expressions from equations (2.58) and (2.59) in equation (2.57), we find that

$$
U^{\left(n_{\alpha}\right)}(y)=\left\{\begin{array}{l}
U^{\left(n_{\alpha}\right)}(B) e^{i \beta_{1}^{n}(y-B)}-\delta_{n 0}\left(e^{i \beta_{1}^{0}(y-2 B)}-e^{-i \beta_{1}^{0} y}\right), \text { in } \Omega_{+}  \tag{2.60}\\
U^{\left(n_{\alpha}\right)}(-B) e^{-i \beta_{2}^{n}(y+B)} \text { in } \Omega_{-} .
\end{array}\right.
$$

From now on, we will just denote by $\underline{n}$ the unit normal on $\Gamma_{+}$. When $y \geq B$, we derive from equation (2.60) that

$$
\frac{\partial U^{\left(n_{\alpha}\right)}(y)}{\partial n}=i \beta_{1}^{n} U^{\left(n_{\alpha}\right)}(y) e^{i \beta_{1}^{n}(y-B)}-\delta_{n 0}\left(i \beta_{1}^{0} e^{i \beta_{1}^{0}(y-2 B)}+i \beta_{1}^{0} e^{-i \beta_{1}^{0} y}\right) .
$$

Hence

$$
\begin{equation*}
\left.\frac{\partial U^{\left(n_{\alpha}\right)}(y)}{\partial n}\right|_{\Gamma_{+}}=i \beta_{1}^{n} U^{\left(n_{\alpha}\right)}(B)-\delta_{n 0} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \tag{2.61}
\end{equation*}
$$

Similarly, let us denote by $\underline{n}$ the unit normal on $\Gamma_{-}$. When $y \leq-B$, we derive from equation (2.60) that

$$
\frac{\partial U^{\left(n_{\alpha}\right)}(y)}{\partial n}=i \beta_{2}^{n} U^{\left(n_{\alpha}\right)}(y) e^{-i \beta_{2}^{n}(y+B)}
$$

Hence

$$
\begin{equation*}
\left.\frac{\partial U^{\left(n_{\alpha}\right)}(y)}{\partial n}\right|_{\Gamma_{-}}=i \beta_{2}^{n} U^{\left(n_{\alpha}\right)}(-B) \tag{2.62}
\end{equation*}
$$

This leads us to,

$$
\begin{align*}
\left.\frac{\partial u}{\partial n}\right|_{\Gamma_{+}} & =\sum_{n \in Z} i \beta_{1}^{n} U^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x}-2 i \beta_{1}^{0} e^{i \alpha x-i \beta_{1}^{0} B}  \tag{2.63}\\
\left.\frac{\partial u}{\partial n}\right|_{\Gamma_{-}} & =\sum_{n \in Z} i \beta_{2}^{n} U^{\left(n_{\alpha}\right)}(-B) e^{i n_{\alpha} x} .
\end{align*}
$$

We now introduce Dirichlet to Neumann maps, $T_{ \pm}$, which are used to match continuously and smoothly the outer analytical solution on $\Gamma_{ \pm}$, with the inner solution given by the finite element methods in $\Omega$. They are called $\operatorname{DtN}$ maps, because they give the relationship between the value of the function $f$ on the boundary (i.e. Dirichlet data) to its normal derivative (i.e. Neumann data).
Definition 3. Let $f \in H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$. We define the DtN maps [9], $T$, where $T f \in$ $H_{\alpha \#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)$and

$$
T_{ \pm} f(x)=\sum_{n \in Z} i \beta_{j}^{n} f^{\left(n_{\alpha}\right)}( \pm B) e^{i n_{\alpha} x}
$$

where $\left.f(x, y)\right|_{\Gamma_{ \pm}}=f(x, \pm B)$, with $z_{n}$ given by equation (2.44), $\beta_{j}^{n}$ given by equation (2.43) such that $j=1$ corresponds to $\Gamma_{+}$and $j=2$ corresponds to $\Gamma_{-}, n_{\alpha}$ given by equation (2.42) and $f^{\left(n_{\alpha}\right)}( \pm B)$ given by Lemma 2 .

In Chapter 4 and Chapter 6, two different methods are used to solve the scattering problem. We will need different $\operatorname{DtN}$ maps for each method, which we will denote by $T_{ \pm}^{\alpha, 0}$ and $T_{ \pm}^{\alpha, \beta}$.
Definition 4. Let $f \in H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$. We define the $\operatorname{DtN}$ maps [11], $T_{ \pm}^{\alpha, 0}$, where $T_{ \pm}^{\alpha, 0} f \in H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)$, and

$$
\begin{aligned}
T_{ \pm}^{\alpha, 0} f(x) & =\sum_{n \in \mathbb{Z}} i \beta_{j}^{n} f^{(n)}( \pm B) e^{i \frac{2 \pi n}{d} x} \\
f^{(n)}( \pm B) & =\frac{1}{d} \int_{0}^{d} f(x, \pm B) e^{-i \frac{2 \pi n}{d} x} d x
\end{aligned}
$$

In Chapter 6, we will introduce a function $w(y)$ such that the solution to the scattering problem $U=e^{i \alpha x} w(y) U_{\alpha, \beta}$. In doing so, we will need to introduce $T_{ \pm}^{\alpha, \beta}$ which is defined as follows
Definition 5. Let $f \in H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$. We define the DtN map, $T_{ \pm}^{\alpha, \beta}$, where $T_{ \pm}^{\alpha, \beta} f \in$ $H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)$and

$$
\begin{equation*}
T_{ \pm}^{\alpha, \beta} f(x)=\sum_{n \in \mathbb{Z}} i\left(\beta_{j}^{n} \pm\left. i \frac{w^{\prime}(y)}{w(y)}\right|_{\Gamma_{ \pm}}\right) f^{(n)}( \pm B) e^{i \frac{2 \pi n}{d} x} \tag{2.64}
\end{equation*}
$$

where

$$
w(y)=\left\{\begin{array}{lr}
t_{1} e^{-i \beta_{1}^{0} y}+r_{1} e^{i \beta_{1}^{0} y}, & b \leq y \leq B  \tag{2.65}\\
t_{2} e^{-i \beta_{2}^{0} y}+r_{2} e^{i \beta_{2}^{0} y}, & -B \leq y \leq-b,
\end{array}\right.
$$

with $b$ and $B$ as shown in Figure 2.3. The parameters $t_{1}=1, r_{2}=0$ and $r_{1}, t_{2} \in \mathbb{C}$, which satisfy $\left|r_{1}\right| \leq 1,\left|t_{2}\right| \leq 1$, and will be explained in detail in Chapter 6 along with the function $w(y)$.
Note that we can combine Definition 4 and Definition 5 to obtain

$$
\begin{equation*}
T_{ \pm}^{\alpha, \beta} f(x)=\left.T_{ \pm}^{\alpha, 0} f(x) \mp \frac{w^{\prime}(y)}{w(y)}\right|_{\Gamma_{ \pm}} \quad I_{d} f(x), \tag{2.66}
\end{equation*}
$$

where $I_{d}$ is the identity operator.

### 2.3.3 Properties of the operators $T, T^{\alpha, 0}, T^{\alpha, \beta}$

In this section we will derive some results regarding the DtN maps defined above.
To ease the notation we will drop the subscript $\pm$ from the maps and write

$$
\begin{aligned}
T & =T_{ \pm} \\
T^{\alpha, 0} & =T_{ \pm}^{\alpha, 0} \\
T^{\alpha, \beta} & =T_{ \pm}^{\alpha, \beta}
\end{aligned}
$$

In order to show the uniqueness of the solution to the scattering problem, one of the prerequisites is that $T$ is a continuous operator. In many papers [14, 13, 9, 118], the authors state that $T$ is a continuous linear form but they do not calculate the constant of continuity. Here we calculate these constants explicitly.

Lemma 6. The operator $T: H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right) \rightarrow H_{\alpha \#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)$is a continuous linear form (the operator is linear and its norm in the image space is bounded by the norm of the function in the domain). There exists a positive constant $2 \leq C^{2} \leq \overline{5}$ such that

$$
\mathrm{k} T f \mathrm{k}_{H_{\alpha \#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \leq C^{2} \sup \left(\left|k_{j}^{2}\right|, 1\right) \mathrm{k} f \mathrm{k}_{H_{\alpha \#}^{2}}^{2}\left(\Gamma_{ \pm}\right)
$$

for $j=1,2$. In addition, for all $n \in \mathbb{N}$ and $j=1,2$

$$
\left|\beta_{j}^{n}\right|^{2} \leq \begin{cases}C^{2}\left|k_{j}^{2}\right| & \text { if }\left|k_{j}^{2}\right|>n_{\alpha}^{2}  \tag{2.67}\\ C^{2} n_{\alpha}^{2} & \text { if }\left|k_{j}^{2}\right|<n_{\alpha}^{2}\end{cases}
$$

Proof. Let $f, g \in H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{+}\right)$, then

$$
\begin{aligned}
T(f+g) & =\sum_{n \in Z} i \beta_{j}^{n}(f+g)^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x} \\
& =\sum_{n \in Z} i \beta_{j}^{n} f^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x}+\sum_{n \in Z} i \beta_{j}^{n} g^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x},
\end{aligned}
$$

since

$$
\begin{aligned}
(f+g)^{\left(n_{\alpha}\right)}(B) & =\frac{1}{d} \int_{0}^{d}(f+g)(x, B) e^{-i n_{\alpha} x} d x \\
& =\frac{1}{d} \int_{0}^{d}(f)(x, B) e^{-i n_{\alpha} x} d x+\frac{1}{d} \int_{0}^{d}(g)(x, B) e^{-i n_{\alpha} x} d x
\end{aligned}
$$

If $c \in \mathbb{C}$, then

$$
\begin{aligned}
T(c f) & =\sum_{n \in Z} i \beta_{j}^{n}(c f)^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x} \\
& =c T(f)
\end{aligned}
$$

To prove continuity, we note from equation (2.43) that

$$
\begin{aligned}
\left|\beta_{j}^{n}\right|^{2} & =\left|n_{\alpha}^{2}-k_{j}^{2}\right| \\
& =\left|n_{\alpha}^{2}\right|\left|1-\frac{k_{j}^{2}}{n_{\alpha}^{2}}\right| .
\end{aligned}
$$

If $n_{\alpha}^{2}>\left|k^{2}\right|$, we have

$$
1-\frac{k_{j}^{2}}{n_{\alpha}^{2}}=1-\frac{\mathfrak{R}\left(k_{j}^{2}\right)}{n_{\alpha}^{2}}-i \frac{\mathfrak{J}\left(k_{j}^{2}\right)}{n_{\alpha}^{2}}
$$

and since

$$
k_{j}^{2}=\mathfrak{R}^{2}\left(k_{j}\right)-\mathfrak{I}^{2}\left(k_{j}\right)+2 i \mathfrak{R}\left(k_{j}\right) \mathfrak{J}\left(k_{j}\right),
$$

then

$$
1-\frac{k_{j}^{2}}{n_{\alpha}^{2}}=1-\frac{\mathfrak{R}^{2}\left(k_{j}\right)}{n_{\alpha}^{2}}+\frac{\mathfrak{J}^{2}\left(k_{j}\right)}{n_{\alpha}^{2}}-i \frac{\mathfrak{J}\left(k_{j}^{2}\right)}{n_{\alpha}^{2}} .
$$

Therefore, we have

$$
\left|1-\frac{k_{j}^{2}}{n_{\alpha}^{2}}\right|<|2-i|
$$

and so if $\left|k_{j}^{2}\right|<n_{\alpha}^{2}$, there exists $2 \leq c_{1} \leq \sqrt{ } \overline{5}$ such that

$$
\begin{equation*}
\left|\left(\beta_{j}^{n}\right)^{2}\right| \leq c_{1} n_{\alpha}^{2} \tag{2.68}
\end{equation*}
$$

When, $\left|n_{\alpha}^{2}\right|<\left|k_{j}^{2}\right|$, we have

$$
\begin{aligned}
\left|1-\frac{n_{\alpha}^{2}}{k_{j}^{2}}\right| & =\left|1-\frac{n_{\alpha}^{2} \Re\left(k_{j}^{2}\right)}{\left|k_{j}^{2}\right|\left|k_{j}^{2}\right|}+i \frac{n_{\alpha}^{2} \mathfrak{J}\left(k_{j}^{2}\right)}{\left|k_{j}^{2}\right|\left|k_{j}^{2}\right|}\right| \\
& \leq|1+i| \\
& \leq \overline{2},
\end{aligned}
$$

so if $\left|k_{j}^{2}\right|>n_{\alpha}^{2}$, there exists $1 \leq c_{2} \leq{ }^{\sqrt{ }} \overline{2}$ such that

$$
\begin{equation*}
\left|\left(\beta_{j}^{n}\right)^{2}\right| \leq c_{2}\left|k_{j}^{2}\right| \tag{2.69}
\end{equation*}
$$

Hence, by setting $C^{2}=\sup \left(c_{1}, c_{2}\right)$, we have from Definition A-17

$$
\begin{aligned}
\mathbf{k} T f(x) \mathbf{k}_{H_{\alpha}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2}= & \sum_{n \in \mathbb{Z}}\left(n_{\alpha}^{2}+1\right)^{-\frac{1}{2}}\left|\beta_{j}^{n}\right|^{2}\left|f^{\left(n_{\alpha}\right)}\right|^{2}, \\
= & \sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|\right\}}\left(n_{\alpha}^{2}+1\right)^{-\frac{1}{2}}\left|\beta_{j}^{n}\right|^{2}\left|f^{\left(n_{\alpha}\right)}\right|^{2} \\
& +\sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|\right\}}\left(n_{\alpha}^{2}+1\right)^{-\frac{1}{2}}\left|\beta_{j}^{n}\right|^{2}\left|f^{\left(n_{\alpha}\right)}\right|^{2}, \\
\leq & \sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|\right\}} C^{2}\left(n_{\alpha}^{2}+1\right)^{-\frac{1}{2}}\left(\left|k_{j}^{2}\right|\right)\left|f^{\left(n_{\alpha}\right)}\right|^{2} \\
& +\sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|\right\}} C^{2}\left(n_{\alpha}^{2}+1\right)^{-\frac{1}{2}} n_{\alpha}^{2}\left|f^{\left(n_{\alpha}\right)}\right|^{2},
\end{aligned}
$$

by using equation (2.68) and equation (2.69). Therefore for $j=1,2$

$$
\begin{aligned}
& \mathrm{k} T f(x) \mathrm{k}_{H_{\alpha \# 1}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \leq C^{2}\left|k_{j}^{2}\right| \\
& \sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|\right\}}\left(n_{\alpha}^{2}+1\right)^{\frac{1}{2}}\left|f^{\left(n_{\alpha}\right)}\right|^{2} \\
&+C^{2} \sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|\right\}}\left(n_{\alpha}^{2}+1\right)^{\frac{1}{2}}\left|f^{\left(n_{\alpha}\right)}\right|^{2}
\end{aligned}
$$

since $n_{\alpha}^{2}<n_{\alpha}^{2}+1$. Finally, we have

$$
\begin{aligned}
\mathrm{k} T f(x) \mathrm{k}_{H_{\alpha \neq \#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} & \leq C^{2} \sup \left(\left|k_{j}^{2}\right|, 1\right) \sum_{n \in \mathbb{Z}}\left(n_{\alpha}^{2}+1\right)^{\frac{1}{2}}\left|f^{\left(n_{\alpha}\right)}\right|^{2} \\
& \leq C^{2} \sup \left(\left|k_{j}^{2}\right|, 1\right) \mathrm{k} f \mathrm{k}_{H_{\alpha \#}^{2}\left(\Gamma_{ \pm}\right)}^{2}
\end{aligned}
$$

by using Definition A-17, which finishes the proof.
We also note the following result which will be used later when we study the well-posedness of the scattering problems.

Lemma 7. The bilinear form

$$
T(f, g):(f, g) \nrightarrow(T f, g)_{\Gamma_{ \pm}}
$$

such that $(., .)_{\Gamma_{ \pm}}$is given by Lemma A-8 is continuous on $H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right) \times H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$, and there exists a constant positive $C \leq \overline{5}$ such that

$$
\left|(T f, g)_{\Gamma_{ \pm}}\right| \leq d^{\sqrt{ }} \bar{C} \sup \left(\left|k_{j}\right|, 1\right) \mathrm{k} f \mathrm{k}_{H_{\alpha \neq \#}^{\frac{1}{\alpha}}\left(\Gamma_{ \pm}\right)} \mathrm{kg}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}
$$

Proof. We have by using Definition 3

$$
\begin{align*}
T(f, g)_{\Gamma_{ \pm}} & =\int_{\Gamma_{ \pm}} \sum_{n \in \mathbb{Z}} i \beta_{j}^{n} f^{\left(n_{\alpha}\right)}( \pm B) e^{i n_{\alpha} x} \overline{g(x, \pm B)} d x \\
& =\sum_{n \in \mathbb{Z}} i \beta_{j}^{n} f^{\left(n_{\alpha}\right)}( \pm B) \int_{\Gamma_{ \pm}} e^{i n_{\alpha} x} \overline{g(x, \pm B)} d x \\
& =d \sum_{n \in \mathbb{Z}} i \beta_{j}^{n} f^{\left(n_{\alpha}\right)}( \pm B) \overline{g^{\left(n_{\alpha}\right)}( \pm B)} \tag{2.70}
\end{align*}
$$

from Lemma 2. We use the Schwarz inequality [22] to write

$$
\begin{aligned}
\left|(T g, f)_{\Gamma_{ \pm}}\right| & =\left|d \sum_{n \in \mathbb{Z}} i\left(\beta_{j}^{n}\right)^{1 / 2} f^{\left(n_{\alpha}\right)}( \pm B)\left(\beta_{j}^{n}\right)^{1 / 2} \overline{g^{\left(n_{\alpha}\right)}( \pm B)}\right| \\
& \leq d\left(\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}\right)^{\frac{1}{2}}\left(\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|\left|f^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}\right)^{\frac{1}{2}}
\end{aligned}
$$

We note that

$$
\begin{aligned}
\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \leq & \sqrt{ } \\
\bar{C}\left|k_{j}\right| & \sum_{\left\{n \in \mathbb{Z},\left|k_{j}^{2}\right|>n_{\alpha}^{2}\right\}}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& +\sqrt{C} \sum_{\left\{n \in \mathbb{Z},\left|k_{j}^{2}\right|<n_{\alpha}^{2}\right\}}\left|n_{\alpha}\right|\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2},
\end{aligned}
$$

by using equation (2.67) in Lemma 6. Since $\left|n_{\alpha}\right|=\left(n_{\alpha}^{2}\right)^{1 / 2}<\left(n_{\alpha}^{2}+1\right)^{1 / 2}$, we can write

$$
\begin{aligned}
\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \leq & \sqrt{ } \bar{C}\left|k_{j}\right| \\
& \sum_{\left\{n \in \mathbb{Z},\left|k_{j}^{2}\right|>n_{\alpha}^{2}\right\}}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& +\sqrt{C} \sum_{\left\{n \in \mathbb{Z},\left|k_{j}^{2}\right|<n_{\alpha}^{2}\right\}}\left(n_{\alpha}^{2}+1\right)^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}
\end{aligned}
$$

We also have $\left(n_{\alpha}^{2}+1\right)^{1 / 2}>1$,

$$
\begin{aligned}
\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \leq & \sqrt{ } \bar{C}\left|k_{j}\right| \sum_{\left\{n \in \mathbb{Z},\left|k_{j}^{2}\right|>n_{\alpha}^{2}\right\}}\left(n_{\alpha}^{2}+1\right)^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& +\sqrt{C} \sum_{\left\{n \in \mathbb{Z},\left|k_{j}^{2}\right|<n_{\alpha}^{2}\right\}}\left(n_{\alpha}^{2}+1\right)^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
\leq & \sqrt{C} \sup \left(\left|k_{j}\right|, 1\right) \sum_{n \in \mathbb{Z}}\left(n_{\alpha}^{2}+1\right)^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}, \\
\leq & \sqrt{C} \sup \left(\left|k_{j}\right|, 1\right) \mathrm{k}_{\mathrm{Z}} \mathrm{k}_{H_{\alpha \neq \#}^{2}}^{2}\left(\Gamma_{ \pm}\right)
\end{aligned}
$$

The last inequality is justified by using the definition of the $H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$norm given in Definition A-17. We apply the same arguments to show

$$
\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|\left|f^{\left(n_{\alpha}\right)}\right|^{2} \leq \sqrt{ }{ }^{\sqrt{C}} \sup \left(\left|k_{j}\right|, 1\right) \mathrm{k} f \mathrm{k}_{H_{\alpha \#}^{2}\left(\Gamma_{ \pm}\right)}^{2}
$$

Therefore,

$$
\left|(T g, f)_{\Gamma_{ \pm}}\right| \leq d^{\sqrt{ }} \bar{C} \sup \left(\left|k_{j}\right|, 1\right){\mathrm{k} g \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}} \mathrm{kf} \mathrm{k}_{H_{\alpha \#( }^{\frac{1}{\alpha}}\left(\Gamma_{ \pm}\right)}
$$

The lemma below has been shown to hold for $k_{j} \in \mathbb{R}[9,85]$. However, in the cases we shall consider $k_{j} \in \mathbb{C}$, and therefore we extend the result.

Lemma 8. The wavenumber $\beta_{j}^{n}$, of the scattered (or diffracted) wave in a direction given by $\alpha$, satisfies

$$
\left.\begin{array}{l}
\mathfrak{R}\left(\beta_{j}^{n}\right) \geq 0  \tag{2.71}\\
\mathfrak{J}\left(\beta_{j}^{n}\right) \geq 0
\end{array}\right\}
$$

The inner product of a function $g$ and its normal derivative on the boundary $\Gamma_{ \pm}$ satisfies

$$
\left.\begin{array}{ll}
\mathfrak{R}(T g, g)_{\Gamma_{ \pm}}=-d \sum_{n \in \mathbb{Z}} \sin \left(z_{n} / 2\right)\left|\beta_{j}^{n}\right|\left|g^{\left(n_{\alpha}\right)}\right|^{2} & \leq 0,  \tag{2.72}\\
\mathfrak{J}(T g, g)_{\Gamma_{ \pm}}=d \sum_{n \in \mathbb{Z}} \cos \left(z_{n} / 2\right)\left|\beta_{j}^{n}\right|\left|g^{\left(n_{\alpha}\right)}\right|^{2} & \geq 0,
\end{array}\right\}
$$

where $z_{n}$ is defined by equation (2.44).
Proof. We use the definition of $\beta_{j}^{n}$ given by equation (2.43)

$$
\beta_{j}^{n}=e^{i z_{n} / 2}\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2}
$$

for $j=1,2$. By using Euler's formula, we can write

$$
\begin{equation*}
\beta_{j}^{n}=\left(\cos z_{n} / 2+i \sin z_{n} / 2\right)\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2} \tag{2.73}
\end{equation*}
$$

Then, with $z_{n}$ as defined by equation (2.44)

$$
\sin z_{n}=\frac{\mathfrak{J}\left(k_{j}^{2}\right)}{\left|k_{j}^{2}-n_{\alpha}^{2}\right|}=\frac{2 \mathfrak{J}\left(k_{j}\right) \Re\left(k_{j}\right)}{\left|k_{j}^{2}-n_{\alpha}^{2}\right|} .
$$

From equation (2.21) and equation (2.23), $k_{j}=w^{\sqrt{ }} \overline{\varepsilon_{j}}>0$ (since the frequency and permittivity are positive). Hence $\mathfrak{R}\left(k_{j}\right)>0$. From equation (2.60), $\mathfrak{J}\left(k_{j}\right) \geq 0$ so that the scattered and diffracted waves are composed of bounded outgoing waves in the $y$ direction and we have

$$
\sin z_{n} \geq 0,
$$

and so $z_{n} \in[0, \pi]$. Therefore, $z_{n} / 2 \in[0, \pi / 2]$ and hence $\cos z_{n} / 2$ and $\sin z_{n} / 2$ are both positive which shows

$$
\begin{array}{ll}
\mathfrak{R}\left(\beta_{j}^{n}\right)=\cos z_{n} / 2\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2} & \geq 0, \\
\mathfrak{J}\left(\beta_{j}^{n}\right)=\sin z_{n} / 2\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2} \geq 0 .
\end{array}
$$

We have by using Definition 3

$$
T(g, g)_{\Gamma_{ \pm}}=d \sum_{n \in \mathbb{Z}} i \beta_{j}^{n}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}
$$

which can be written using equation (2.73)

$$
T(g, g)_{\Gamma_{ \pm}}=d \sum_{n \in \mathbb{Z}} i\left(\cos z_{n} / 2+i \sin z_{n} / 2\right)\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}
$$

Then,

$$
\mathfrak{R}\left(T(g, g)_{\Gamma_{ \pm}}\right)=-d \sum_{n \in \mathbb{Z}} \sin z_{n} / 2\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}
$$

and

$$
\mathfrak{J}\left(T(g, g)_{\Gamma_{ \pm}}\right)=d \sum_{n \in \mathbb{Z}} \cos z_{n} / 2\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}
$$

Since $d\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2}\left|g^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \geq 0$, we can use equation (2.71) and so

$$
\mathfrak{R}\left(T(g, g)_{\Gamma_{ \pm}}\right) \leq 0,
$$

and

$$
\mathfrak{I}\left(T(g, g)_{\Gamma_{ \pm}}\right) \geq 0 .
$$

The following lemma is a new result concerning the relation between the two operators $T$ and $T^{\alpha, 0}$ which we can use to derive some properties of $T^{\alpha, 0}$ using $T$ or vice versa.

Lemma 9. For $f \in L_{\#}^{1}([0, d])$

$$
T\left(e^{i \alpha x} f(x)\right)=e^{i \alpha x} T^{\alpha, 0} f(x)
$$

Proof. First let us consider the case where $y=B$. Using Definition 3

$$
T_{+} g(x)=\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} g^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x} .
$$

We take $g=e^{i \alpha x} f(x, B)$ and so

$$
T_{+}\left(e^{i \alpha x} f(x, B)\right)=\sum_{n \in \mathbb{Z}} i \beta_{1}^{n}\left(e^{i \alpha x} f(x, B)\right)^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x}
$$

By using equation (2.53)

$$
\begin{aligned}
\left(e^{i \alpha x} f(x, B)\right)^{\left(n_{\alpha}\right)}(B) & =\frac{1}{d} \int_{0}^{d} e^{i \alpha x} f(x, B) e^{-i n_{\alpha} x} d x \\
& =\frac{1}{d} \int_{0}^{d} e^{i \alpha x} f(x, B) e^{-i\left(\alpha+\frac{2 \pi n}{d}\right) x} d x \\
& =\frac{1}{d} \int_{0}^{d} f(x, B) e^{-i \frac{2 \pi n}{d} x} d x
\end{aligned}
$$

Hence

$$
\begin{equation*}
\left(e^{i \alpha x} f(x, B)\right)^{\left(n_{\alpha}\right)}(B)=f^{(n)}(B) \tag{2.74}
\end{equation*}
$$

Using equation (2.53) again gives

$$
\begin{aligned}
T_{+}\left(e^{i \alpha x} f(x, B)\right) & =\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} f^{(n)}(B) e^{i n_{\alpha} x}, \\
& =\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} f^{(n)}(B) e^{i\left(\alpha+\frac{2 \pi n}{d}\right) x}, \\
& =e^{i \alpha x} \sum_{n \in \mathbb{Z}} i \beta_{1}^{n} f^{(n)}(B) e^{i \frac{2 \pi n}{d} x},
\end{aligned}
$$

since $e^{i \alpha x}$ is independent of $n$. Hence

$$
T_{+}\left(e^{i \alpha x} f(x, B)\right)=e^{i \alpha x} T_{+}^{\alpha, 0} f(x)
$$

by using Definition 4.
When $y=-B$ a similar procedure can be followed to show that

$$
T_{-}\left(e^{i \alpha x} f(x)\right)=e^{i \alpha x} T_{-}^{\alpha} f(x)
$$

The different methods described in the next two chapters lead us to use the two operators $T^{\alpha, 0}$ and $T^{\alpha, \beta}$. In order to show the uniqueness of the solution of the scattering problem using these two different methods, we require the continuity of the operators $T^{\alpha, 0}$ and $T^{\alpha, \beta}$. We note by using Definition 5 , that the continuity of $T^{\alpha, \beta}$ follows from the continuity of $T^{\alpha, 0}$. In the literature [9, 14, 118], several authors state that $T^{\alpha, 0}$ is a continuous linear form but they don't calculate the constant of continuity. Below we calculate this constant explicitly.
Lemma 10. The operator $T^{\alpha, 0}: H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right) \rightarrow H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)$is a continuous linear form. There exist positive constants $C \leq \overline{5}$ and $C_{1}=\sup \left(3,1+\frac{\alpha d}{2 \pi}\right)^{2}$ such that

$$
\mathrm{k} T^{\alpha, 0} f \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \leq C \sup \left(\left|k_{j}^{2}\right|, C_{1}\right) \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2}
$$

for $j=1,2$ where $f \in H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$.
Proof. Let $f, g \in H_{\#}^{\frac{1}{2}}\left(\Gamma_{+}\right)$, then from Definition 4,

$$
\begin{aligned}
T^{\alpha, 0}(f+g) & =\sum_{n \in Z} i \beta_{j}^{n}(f+g)^{(n)}( \pm B) e^{i \frac{2 \pi n}{d} x} \\
& =\sum_{n \in Z} i \beta_{j}^{n} f^{(n)}( \pm B) e^{i \frac{2 \pi n}{d} x}+\sum_{n \in Z} i \beta_{j}^{n} g^{(n)}( \pm B) e^{i \frac{i \pi n}{d} x}
\end{aligned}
$$

because

$$
\begin{aligned}
(f+g)^{(n)}( \pm B) & =\frac{1}{d} \int_{0}^{d}(f+g)(x, \pm B) e^{-i \frac{2 \pi n}{d} x} d x \\
& =\frac{1}{d} \int_{0}^{d} f(x, \pm B) e^{-i \frac{2 \pi n}{d} x} d x+\frac{1}{d} \int_{0}^{d} g(x, \pm B) e^{-i \frac{2 \pi n}{d} x} d x
\end{aligned}
$$

If $c \in \mathbb{C}$, then

$$
\begin{aligned}
T^{\alpha, 0}(c f) & =\sum_{n \in Z} i \beta_{j}^{n}(c f)^{(n)}( \pm B) e^{i \frac{2 \pi n}{d} x} \\
& =c T^{\alpha, 0}(f)
\end{aligned}
$$

To prove continuity, we note that from Definition A-14

$$
\begin{aligned}
& \mathrm{k} T^{\alpha, 0} f(x) \mathrm{k}_{H_{\alpha \neq}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}= \sum_{n \in \mathbb{Z}}\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{-\frac{1}{2}} \\
&=\left|\beta_{j}^{n}\right|^{2}\left|f^{(n)}( \pm B)\right|^{2}, \\
&\left\{\sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|\right\}}\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{-\frac{1}{2}}\left|\beta_{j}^{n}\right|^{2}\left|f^{(n)}( \pm B)\right|^{2}\right. \\
&+\sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|\right\}}\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{-\frac{1}{2}}\left|\beta_{j}^{n}\right|^{2}\left|f^{(n)}( \pm B)\right|^{2},
\end{aligned}
$$

Then,

$$
\begin{aligned}
& \mathrm{k} T^{\alpha, 0} f(x) \mathrm{k}_{H_{\alpha \alpha}^{-\alpha}}^{2-\frac{1}{2}}\left(\Gamma_{ \pm}\right) \\
& \leq \sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|\right\}} C\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{-\frac{1}{2}}\left|k_{j}^{2}\right|\left|f^{(n)}( \pm B)\right|^{2} \\
&+\sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|\right\}} C\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{-\frac{1}{2}} n_{\alpha}^{2}\left|f^{(n)}( \pm B)\right|^{2}
\end{aligned}
$$

by using equation (2.67) in Lemma 6. Since $\left(\frac{2 \pi n}{d}\right)^{2}+1 \geq 1$, then

$$
\begin{align*}
\mathrm{k} T^{\alpha, 0} f(x) \mathrm{k}_{H_{\alpha \| \#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \leq & \sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|\right\}} C\left|k_{j}^{2}\right|\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2} \\
& +\sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|\right\}} C\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{-\frac{1}{2}} n_{\alpha}^{2}\left|f^{(n)}( \pm B)\right|^{2} . \tag{2.75}
\end{align*}
$$

We also note that when $n_{\alpha}^{2}>\left|k_{j}^{2}\right|$ but $\left(\frac{2 \pi n}{d}\right)^{2} \leq n_{\alpha}^{2}$ we have

$$
\begin{align*}
\frac{n_{\alpha}^{2}}{\left(\frac{2 \pi n}{d}\right)^{2}+1} & \leq \frac{n_{\alpha}^{2}}{\left(\frac{2 \pi n}{d}\right)^{2}} \\
& =\left(\frac{n_{\alpha}}{\frac{2 \pi n}{d}+\alpha-\alpha}\right)^{2} \\
& =\left(\frac{n_{\alpha}}{n_{\alpha}-\alpha}\right)^{2} \\
& =\left(\frac{1}{1-\frac{\alpha}{n_{\alpha}}}\right)^{2} \tag{2.76}
\end{align*}
$$

Since $\left(\frac{2 \pi n}{d}\right)^{2} \leq n_{\alpha}^{2}$ then $\left(\frac{2 \pi n}{d}\right)^{2} \leq\left(\frac{2 \pi n}{d}+\alpha\right)^{2}$. Therefore, we have $\frac{4 \pi n}{d} \alpha+\alpha^{2} \geq 0$. If $n<0$, then

$$
\begin{equation*}
\frac{4 \pi|n|}{d} \alpha \leq \alpha^{2} \tag{2.77}
\end{equation*}
$$

From equation (2.77), we have

$$
\begin{array}{rll}
-\frac{\alpha}{2} & \leq \frac{2 \pi n}{d} & \leq \frac{\alpha}{2} \\
-\frac{\alpha}{2}+\alpha & \leq \frac{2 \pi n}{d}+\alpha & \leq \frac{\alpha}{2}+\alpha, \\
\frac{\alpha}{2} & \leq n_{\alpha} & \leq \frac{3 \alpha}{2}
\end{array}
$$

and so

$$
\begin{aligned}
& \frac{2}{3 \alpha} \leq \frac{1}{n_{\alpha}} \quad \leq \frac{2}{\alpha}, \\
& -\frac{2}{\alpha} \leq-\frac{1}{n_{\alpha}} \quad \leq-\frac{2}{3 \alpha} \text {, } \\
& -2 \leq-\frac{\alpha}{n_{\alpha}} \leq-\frac{2}{3} \text {, } \\
& -1 \leq 1-\frac{\alpha}{n_{\alpha}} \leq \frac{1}{3} .
\end{aligned}
$$

Therefore, for $\left(\frac{2 \pi n}{d}\right)^{2} \leq n_{\alpha}^{2}$ and $n<0$ we have

$$
\begin{equation*}
1 \leq \frac{1}{\left(1-\frac{\alpha}{n_{\alpha}}\right)^{2}} \leq 9 \tag{2.78}
\end{equation*}
$$

If $\left(\frac{2 \pi n}{d}\right)^{2} \leq n_{\alpha}^{2}$ and $n>0$, we have

$$
\begin{aligned}
& \frac{2 \pi}{d}+\alpha \leq \frac{2 \pi n}{d}+\alpha \leq \infty, \\
& 0 \leq \frac{1}{n_{\alpha}} \leq \frac{1}{\frac{2 \pi}{d}+\alpha}, \\
& -\frac{1}{\frac{2 \pi}{d}+\alpha} \quad \leq-\frac{1}{n_{\alpha}} \quad \leq 0, \\
& -\frac{\alpha}{\frac{2 \pi}{d}+\alpha} \quad \leq-\frac{\alpha}{n_{\alpha}} \quad \leq 0, \\
& 1-\frac{\alpha}{\frac{2 \pi}{d}+\alpha} \leq 1-\frac{\alpha}{n_{\alpha}} \leq 1 .
\end{aligned}
$$

Therefore, for $\left(\frac{2 \pi n}{d}\right)^{2} \leq n_{\alpha}^{2}$ and $n>0$ we have

$$
\begin{equation*}
1 \leq \frac{1}{\left(1-\frac{\alpha}{n_{\alpha}}\right)^{2}} \leq\left(\frac{\frac{2 \pi}{d}+\alpha}{\frac{2 \pi}{d}}\right)^{2} \tag{2.79}
\end{equation*}
$$

We do not consider the case where $n=0$ since $n_{\alpha}^{2}>\left|k_{j}^{2}\right|$ and $\left|k_{j}\right| \geq \alpha$. Hence for $n_{\alpha}^{2}>\left|k_{j}^{2}\right|$, and $\left(\frac{2 \pi n}{d}\right)^{2} \leq n_{\alpha}^{2}$, we have from equations (2.76), (2.78) and (2.79)

$$
\frac{n_{\alpha}^{2}}{1+\left(\frac{2 \pi n}{d}\right)^{2}} \leq \sup \left(9,\left(\frac{\frac{2 \pi}{d}+\alpha}{\frac{2 \pi}{d}}\right)^{2}\right)
$$

and so if we note $C_{1}=\sup \left(9,\left(\frac{\frac{2 \pi}{d}+\alpha}{\frac{2 \pi}{d}}\right)^{2}\right)$

$$
\begin{equation*}
\frac{n_{\alpha}^{2}}{\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{\frac{1}{2}}} \leq C_{1}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{\frac{1}{2}} \tag{2.80}
\end{equation*}
$$

When $n_{\alpha}^{2}>\left|k_{j}^{2}\right|$, we note that if $\left(\frac{2 \pi n}{d}\right)^{2} \geq n_{\alpha}^{2}$ then

$$
\begin{align*}
\frac{n_{\alpha}^{2}}{\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{\frac{1}{2}}} & \leq \frac{\left(\frac{2 \pi n}{d}\right)^{2}}{\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{\frac{1}{2}}} \\
& \leq\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{\frac{1}{2}} \tag{2.81}
\end{align*}
$$

Now we can rewrite equation (2.75)

$$
\begin{aligned}
\mathrm{k} T^{\alpha, 0} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \leq & \sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|\right\}} C\left|k_{j}^{2}\right|\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2} \\
& +\sum_{\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|\right\}} C C_{1}\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2}
\end{aligned}
$$

Finally,

$$
\mathrm{k} T^{\alpha, 0} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \leq C \sup \left(\left|k_{j}^{2}\right|, C_{1}\right) \sum_{n \in \mathbb{Z}}\left(\left(\frac{2 \pi n}{d}\right)^{2}+1\right)^{\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2}
$$

We finish the proof by using Definition A-14 and so

$$
\mathrm{k} T^{\alpha, 0} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \leq C \sup \left(\left|k_{j}^{2}\right|, C_{1}\right) \mathrm{k} f \mathrm{k}_{H_{\#}^{2}\left(\Gamma_{ \pm}\right)}^{2} .
$$

We have the following relation between the two DtN maps $T$ and $T^{\alpha, 0}$.
Lemma 11. Let us denote $f_{\alpha}=e^{i \alpha x} f$ and $g_{\alpha}=e^{i \alpha x} g$ such that $f, g \in H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$.
Then

$$
\begin{equation*}
\left|T\left(f_{\alpha}, g_{\alpha}\right)\right|=\left|\left(T^{\alpha, 0} f, g\right)_{\Gamma_{ \pm}}\right| \tag{2.82}
\end{equation*}
$$

Proof. We note that $f_{\alpha}, g_{\alpha} \in H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$by using Lemma A-16 and so

$$
\begin{aligned}
\left|T\left(f_{\alpha}, g_{\alpha}\right)\right| & =\left|\left(T f_{\alpha}, g_{\alpha}\right)_{\Gamma_{ \pm}}\right|, \text {using Lemma } 7 \\
& =\left|\left(T\left(e^{i \alpha x} f\right), g_{\alpha}\right)_{\Gamma_{ \pm}}\right| \\
& =\left|\left(e^{i \alpha x} T^{\alpha, 0} f, g_{\alpha}\right)_{\Gamma_{ \pm}}\right|, \text {using Lemma } 9
\end{aligned}
$$

We note that

$$
\begin{aligned}
\left(e^{i \alpha x} T^{\alpha, 0} f, g_{\alpha}\right)_{\Gamma_{ \pm}} & =\int_{\Gamma_{ \pm}} e^{i \alpha x} T^{\alpha, 0} f \overline{g_{\alpha}} d x \\
& =\int_{\Gamma_{ \pm}} e^{i \alpha x} T^{\alpha, 0} f \overline{e^{i \alpha x} g} d x \\
& =\int_{\Gamma_{ \pm}} T^{\alpha, 0} f \bar{g} d x, \\
& =\left(T^{\alpha, 0} f, g\right)_{\Gamma_{ \pm}} .
\end{aligned}
$$

Hence

$$
|T(f, g)|=\left|\left(T^{\alpha, 0} f, g\right)_{\Gamma_{ \pm}}\right| .
$$

We also have the following new result which will be used to establish between the norms of $\alpha$-quasi periodic and periodic functions.

Lemma 12. Let $n \in \mathbb{Z}$ and let $n_{\alpha}$ satisfy equation (2.42). There exists constants $C_{\#}>\frac{2+\alpha^{2}+\sqrt{\alpha^{4}+4 \alpha^{2}}}{2}$ and

$$
C_{\alpha \#}> \begin{cases}\frac{-\left(2+\alpha^{2}\right)+\frac{\sqrt{ }}{\left(2+\alpha^{2}\right)^{2}+4\left(2 \alpha^{2}-1\right)}}{\mathcal{D}^{2}\left(\frac{\left.2 \alpha^{2}-1\right)}{\left(2+\alpha^{2}\right)^{2}+4\left(2 \alpha^{2}-1\right)}\right.}, & \text { if } \alpha^{2}>\frac{1}{2}  \tag{2.83}\\ \frac{-\left(2+\alpha^{2}\right)-}{2\left(2 \alpha^{2}-1\right)}, & \text { otherwise. }\end{cases}
$$

Then

$$
\begin{equation*}
1+\left(n_{\alpha}\right)^{2} \leq C_{\#}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right) \tag{2.84}
\end{equation*}
$$

and

$$
\begin{equation*}
1+\left(\frac{2 \pi n}{d}\right)^{2} \leq C_{\alpha \#}\left(1+\left(n_{\alpha}\right)^{2}\right) \tag{2.85}
\end{equation*}
$$

Proof. Let us denote $y=n_{\alpha}, x=\frac{2 \pi n}{d}$. We want to find a constant $C_{\#}$ such that

$$
\begin{equation*}
1+y^{2} \leq C_{\#}\left(1+x^{2}\right) \tag{2.86}
\end{equation*}
$$

where $y=\alpha+x$. When $C_{\#}\left(1+x^{2}\right) \geq 1+(x+\alpha)^{2}$ then $C_{\#}>K$ such that $K\left(1+x^{2}\right)=1+(x+\alpha)^{2}$ with one point at intersection. Expanding this,

$$
K\left(1+x^{2}\right)=1+x^{2}+\alpha^{2}+2 \alpha x
$$

i.e.

$$
(K-1) x^{2}-2 \alpha x+\left(K-1-\alpha^{2}\right)=0 .
$$

To intersect once this quadratic has only one solution in $x$ and so the discriminant must be zero. That is

$$
\begin{gathered}
\alpha^{2}-(K-1)\left(K-1-\alpha^{2}\right)=0, \\
\alpha^{2}-\left(K^{2}-2 K+\alpha^{2}-K \alpha^{2}+1\right)=0, \\
K^{2}-\left(2+\alpha^{2}\right) K+1=0,
\end{gathered}
$$

i.e.

$$
K=\frac{\left(2+\alpha^{2}\right) \pm \sqrt{(2+\alpha)^{2}-4}}{2}=\frac{\left(2+\alpha^{2}\right) \pm \sqrt{ } \overline{\alpha^{4}+4 \alpha^{2}}}{2}
$$

Taking the larger of these then for $C_{\#}>\frac{\left(2+\alpha^{2}\right)+\sqrt{\alpha^{4}+4 \alpha^{2}}}{2}$ we have

$$
1+n_{\alpha}^{2} \leq C_{\#}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)
$$

Let us denote $y=n_{\alpha}$ and $x=\frac{2 \pi n}{d}$, similarly we have $\left(1+x^{2}\right) \leq C_{\alpha \#}\left(1+y^{2}\right)$ then $C_{\alpha \#}>K^{*}$ such that $1+x^{2}=K^{*}\left(1+(x+\alpha)^{2}\right)$ and so

$$
\begin{aligned}
1+x^{2} & =K^{*}+K^{*}\left(x^{2}+2 \alpha x+\alpha^{2}\right) \\
\left(K^{*}-1\right) x^{2} & +2 K^{*} \alpha x+K^{*}+K^{*} \alpha^{2}-1=0
\end{aligned}
$$

To intersect once this quadratic has only one solution in $x$ and so the discriminant must be zero. That is

$$
\begin{gathered}
\left(K^{*}\right)^{2} \alpha^{2}-\left(K^{*}-1\right)\left(K^{*}+K^{*} \alpha^{2}-1\right)=0 \\
\left(K^{*}\right)^{2}\left(2 \alpha^{2}-1\right)+K^{*}\left(2+\alpha^{2}\right)-1=0
\end{gathered}
$$

i.e.

$$
K_{1}^{*}=\frac{-\left(2+\alpha^{2}\right)+\sqrt{\left(2+\alpha^{2}\right)^{2}+4\left(2 \alpha^{2}-1\right)}}{2\left(2 \alpha^{2}-1\right)}
$$

and

$$
K_{2}^{*}=\frac{-\left(2+\alpha^{2}\right)-\sqrt{\left(2+\alpha^{2}\right)^{2}+4\left(2 \alpha^{2}-1\right)}}{2\left(2 \alpha^{2}-1\right)}
$$

we see that if $\alpha^{2}>\frac{1}{2}, K_{2}^{*}<K_{1}^{*}$ so we choose $C_{\alpha \#}>K_{1}^{*}$ and vice versa when $\alpha^{2}<\frac{1}{2}$.

We also need the following result to establish relationship between the norms of $\alpha$-quasi periodic and periodic functions.

Lemma 13. Let $f \in L_{\#}^{2}\left(\Gamma_{ \pm}\right)$and let $f_{\alpha} \in L_{\alpha \#}^{2}\left(\Gamma_{ \pm}\right)$we have

$$
f^{(n)}(y)=f_{\alpha}^{\left(n_{\alpha}\right)}(y) .
$$

Proof. We have

$$
\begin{aligned}
f^{(n)}(y) & =\frac{1}{d} \int_{0}^{d} f(x, y) e^{-i \frac{2 \pi n}{d} x} d x \\
& =\frac{1}{d} \int_{0}^{d} e^{-i \alpha x} f_{\alpha}(x, y) e^{-i \frac{2 \pi n}{d} x} d x \\
& =\frac{1}{d} \int_{0}^{d} f_{\alpha}(x, y) e^{-i n_{\alpha} x} d x \\
& =f_{\alpha}^{\left(n_{\alpha}\right)}(y)
\end{aligned}
$$

Lemma 14. The bilinear form

$$
T^{\alpha, 0}(f, g):(f, g) \nrightarrow\left(T^{\alpha, 0} f, g\right)_{\Gamma_{ \pm}}
$$

is continuous on $H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right) \times H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$and

$$
\left|\left(T^{\alpha, 0} f, g\right)_{\Gamma_{ \pm}}\right| \leq d^{\sqrt{2}} \bar{C} \sup \left(\left|k_{j}\right|, 1\right) \mathrm{kfk}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \mathrm{k}_{\mathrm{g}} \mathrm{k}_{H_{\#}^{\frac{1}{\#}}\left(\Gamma_{ \pm}\right)},
$$

where $C \leq{ }^{\sqrt{ }} \overline{5}$ and $C_{\#}$ as defined in Lemma 12.
Proof. We can use Lemma 11 and we note that

$$
\left|T^{\alpha, 0}(f, g)_{\Gamma_{ \pm}}\right|=\left|\left(T f_{\alpha}, g_{\alpha}\right)_{\Gamma_{ \pm}}\right| .
$$

We know from Lemma 7 that $T(f, g)=(T f, g)_{\Gamma_{ \pm}}$is a continuous bilinear form and so

$$
\begin{align*}
\left|T^{\alpha, 0}(f, g)_{\Gamma_{ \pm}}\right| & =\left|\left(T f_{\alpha}, g_{\alpha}\right)_{\Gamma_{ \pm}}\right| \\
& \leq d \bar{C} \overline{\sup }\left(\left|k_{j}\right|, 1\right) \mathrm{k} f_{\alpha} \mathrm{k}_{H_{\alpha \# \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \mathrm{k} g_{\alpha} \mathrm{k}_{H_{\alpha \# \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}, \tag{2.87}
\end{align*}
$$

where $C \leq{ }^{\sqrt{ }} \overline{5}$. In order to prove the continuity of the bilinear form $T^{\alpha, 0}(f, g)$, we need to express $\mathbf{k} f_{\alpha} \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}$in terms of $\mathbf{k} f \mathbf{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}$.

Using Definition A-17, we have

$$
\begin{aligned}
\mathrm{k} f_{\alpha} \mathrm{k}_{H_{\alpha \#}^{2}\left(\Gamma_{ \pm}\right)}^{2} & =\sum_{n \in \mathbb{Z}}\left(1+\left(n_{\alpha}\right)^{2}\right)^{\frac{1}{2}}\left|f_{\alpha}^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& =\sum_{n \in \mathbb{Z}}\left(1+\left(n_{\alpha}\right)^{2}\right)^{\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2}, \quad \text { using Lemma 13 } \\
& \leq \sum_{n \in \mathbb{Z}} C_{\#}^{\frac{1}{2}}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2}
\end{aligned}
$$

from Lemma 12, equation (2.84) and so

$$
\begin{equation*}
\mathrm{k} f_{\alpha} \mathrm{k}_{H_{\alpha \#}^{2}\left(\Gamma_{ \pm}\right)}^{2} \leq C_{\#}^{\frac{1}{2}} \mathrm{k} f \mathrm{k}_{H_{\#}^{2}\left(\Gamma_{ \pm}\right)}^{2} \tag{2.88}
\end{equation*}
$$

using Definition A-14. We can show similarly that

$$
\begin{equation*}
\mathrm{k} g_{\alpha} \mathrm{k}_{H_{\alpha \#}^{2}}^{2}\left(\Gamma_{ \pm}\right) \leq C_{\#}^{\frac{1}{2}} \mathrm{k} g \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} . \tag{2.89}
\end{equation*}
$$

We combine equations (2.87), (2.88) and (2.89) and so

$$
\begin{equation*}
\left|T^{\alpha, 0}(f, g)_{\Gamma_{ \pm}}\right| \leq d^{\sqrt{2}} \bar{C} C_{\#} \sup \left(\left|k_{j}\right|, 1\right) \mathrm{k} f \mathrm{k}_{H_{\#}^{2}\left(\Gamma_{ \pm}\right)}^{2} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2} \tag{2.90}
\end{equation*}
$$

to finish the proof.

Lemma 15. Let $f(x) \in H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$, then $T_{\sqrt{\alpha, \beta}}^{\alpha}$ is a continuous operator in $H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$, and there exists a positive constant $\tilde{C}=\sqrt{ } \bar{C}+\sup \left(\frac{|1+2 i|}{\left(\left|R_{1}\right|-1\right)^{2}}, 1\right)$ such that

$$
\begin{equation*}
\mathrm{k} T^{\alpha, \beta} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \leq \tilde{C} \sup \left(\left|k_{j}\right|, \sqrt{C_{1}}\right) \mathrm{k} f(x) \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} . \tag{2.91}
\end{equation*}
$$

with $C \leq{ }^{\sqrt{ }} \overline{\overline{5}},\left|R_{1}\right| \leq 1$ and ${ }^{\sqrt{ }} \overline{C_{1}}=\sup \left(3,1+\frac{\alpha d}{2 \pi}\right)$.
Proof. We use the property of $T^{\alpha, \beta}$ in the remark following Definition 5,

$$
T^{\alpha, \beta}=\left.T^{\alpha, 0} \mp \frac{w^{\prime}(y)}{w(y)}\right|_{y= \pm B} I_{d}
$$

Note that the identity operator $I_{d}$, is a continuous operator in $H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)$. When $y=-B$ then, from equation (2.65),

$$
\left.\frac{w^{\prime}(y)}{w(y)}\right|_{y=-B}=\frac{-i \beta_{2}^{0} t_{2} e^{i \beta_{2}^{0} B}}{t_{2} e^{i \beta_{2}^{0} B}}=-i \beta_{2}^{0}
$$

and so

$$
\left|\frac{w^{\prime}(y)}{w(y)}\right|_{y=-B}\left|=\left|\beta_{2}^{0}\right| .\right.
$$

When $y=B$, we have

$$
\left.\frac{w^{\prime}(y)}{w(y)}\right|_{y=B}=i \beta_{1}^{0} \frac{r_{1} e^{i \beta_{1}^{0} B}-e^{-i \beta_{1}^{0} B}}{r_{1} e^{i \beta_{1}^{0} B}+e^{-i \beta_{1}^{0} B}},
$$

multiplying by the complex conjugate of $w(B)$, we get

$$
\left.\frac{w^{\prime}(y)}{w(y)}\right|_{y=B}=i \beta_{1}^{0} \frac{\left|r_{1}\right|^{2}-1+2 i \mathfrak{J}\left(r_{1} e^{2 i \beta_{1}^{0} B}\right)}{|w(B)|^{2}}
$$

This leads us to

$$
\begin{aligned}
\left.\left|\frac{w^{\prime}(y)}{w(y)}\right|_{y=B} \right\rvert\, & =\left|i \beta_{1}^{0} \frac{\left|r_{1}\right|^{2}-1+2 i \mathfrak{I}\left(r_{1} e^{2 i \beta_{1}^{0} B}\right)}{|w(B)|^{2}}\right|, \\
& \leq \beta_{1}^{0}\left|\frac{1+2 i}{\left|r_{1}\right|^{2}+1+2 \mathfrak{R}\left(r_{1} e^{2 i \beta_{1}^{0} B}\right)}\right| \\
& \leq \beta_{1}^{0}\left|\frac{1+2 i}{\left.\left|r_{1}\right|^{2}+1-2\left|r_{1}\right|\right)}\right|
\end{aligned}
$$

since $\left|r_{1}\right| \leq 1$, and $\left|\mathfrak{J}\left(r_{1} e^{2 i \beta_{1}^{0} B}\right)\right| \leq\left|r_{1} e^{2 i \beta_{1}^{0} B}\right|=\left|r_{1}\right|$. Let

$$
C_{w}\left(\alpha, \beta_{j}^{0}, B\right)=\sup \left(\beta_{1}^{0}\left|\frac{1+2 i}{\left(\left|r_{1}\right|-1\right)^{2}}\right|,\left|\beta_{2}^{0}\right|\right)
$$

then

$$
\begin{align*}
\mathrm{k} T^{\alpha, \beta} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)} & =\left.\mathrm{k} T^{\alpha, 0} f(x) \mp \frac{w^{\prime}(y)}{w(y)}\right|_{y= \pm B} I_{d} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \\
& \leq \sqrt{C \sup \left(\left|k_{j}^{2}\right|, C_{1}\right)} \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}+C_{w}\left(\alpha, \beta_{j}, B\right) \mathrm{k} f \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \tag{2.92}
\end{align*}
$$

by the triangle inequality and Lemma 10 . Note that

$$
\begin{align*}
C_{w}\left(\alpha, \beta_{j}^{0}, B\right) & =\sup \left(\beta_{1}^{0}\left|\frac{1+2 i}{\left(\left|r_{1}\right|-1\right)^{2}}\right|,\left|\beta_{2}^{0}\right|\right), \\
& \leq\left|k_{j}\right| \sup \left(\frac{|1+2 i|}{\left(\left|r_{1}\right|-1\right)^{2}}, 1\right), \tag{2.93}
\end{align*}
$$

since $\left|\beta_{j}^{0}\right| \leq\left|k_{j}\right|$ from equation (2.43). Using Definition A-14, we have for $s=-\frac{1}{2}$ that

$$
\begin{align*}
\mathrm{k} f \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)} & =\sum_{n \in \mathbb{Z}}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{-\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2} \\
& \leq \sum_{n \in \mathbb{Z}}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{\frac{1}{2}}\left|f^{(n)}( \pm B)\right|^{2} \\
& =\mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \tag{2.94}
\end{align*}
$$

from Definition A-14. Hence, using equations (2.93) and (2.94) we can rewrite (2.92) and

$$
\begin{aligned}
\mathrm{k} T^{\alpha, \beta} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \leq & \sqrt{C \sup \left(\left|k_{j}^{2}\right|, C_{1}\right)} \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \\
& +\left|k_{j}\right| \sup \left(\frac{|1+2 i|}{\left(\left|r_{1}\right|-1\right)^{2}}, 1\right) \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \\
\leq & \sup \left(\left|k_{j}\right|, \sqrt{C_{1}}\right)\left(\sqrt{ } \bar{C}+\sup \left(\frac{|1+2 i|}{\left(\left|r_{1}\right|-1\right)^{2}}, 1\right)\right) \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} .
\end{aligned}
$$

Let us denote by $\tilde{C}=\sqrt{ } \bar{C}+\sup \left(\frac{|1+2 i|}{\left(\left|r_{1}\right|-1\right)^{2}}, 1\right)$, then we have

$$
\mathrm{k} T^{\alpha, \beta} f(x) \mathrm{k}_{H_{\#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \leq \tilde{C} \sup \left(\left|k_{j}\right|, \sqrt{C_{1}}\right) \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} .
$$

### 2.4 Mathematical formulation of the truncated problem

Given the above DtN definitions we can now restate our problem as described in the following four lemmas. After stating all four lemmas, we outline a common method of proof.

## - Case 1A: Perfectly conducting grating: TE case

Lemma 16. $U$ is a solution of (2.25) if and only if it is a solution of the following problem set in the truncated domain $\Omega \backslash$ int $\Omega_{3}$ (see Figure 2.3)

$$
\begin{equation*}
\Delta U(x, y)+k_{1}^{2} U(x, y)=0, \quad(x, y) \in \Omega \backslash \Omega_{3} \tag{2.95}
\end{equation*}
$$

with the DtN map interface conditions at the boundaries of the truncated region given by

$$
\begin{array}{ll}
\left.\partial_{n} U(x, y)\right|_{\Gamma_{+}}=T_{+} U+g(x) & x \in \Gamma_{+}, \\
\left.\partial_{n} U(x, y)\right|_{\Gamma_{-}}=T_{-} U & x \in \Gamma_{-}, \tag{2.97}
\end{array}
$$

the Dirichlet boundary conditions at the surface of the diffraction grating

$$
U(x, y)=0, \quad(x, y) \in \partial \Omega_{3}
$$

and the $\alpha$-quasi-periodic condition

$$
U(d, y)=e^{i \alpha d} U(0, y), \quad y \in[-B, B] .
$$

Here

$$
\begin{align*}
T_{+} U(x) & =\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} U^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x},  \tag{2.98}\\
T_{-} U(x) & =\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} U^{\left(n_{\alpha}\right)}(-B) e^{i n_{\alpha} x},  \tag{2.99}\\
g(x) & =-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B+i \alpha x}, \tag{2.100}
\end{align*}
$$

with

$$
\begin{equation*}
U^{\left(n_{\alpha}\right)}(y)=\frac{1}{d} \int_{0}^{d} U(x, y) e^{-i n_{\alpha} x} d x \tag{2.101}
\end{equation*}
$$

and $n_{\alpha}=\frac{2 \pi n}{d}+\alpha$, with

$$
z_{n}=\arg \left(k_{1}^{2}-n_{\alpha}^{2}\right),
$$

and

$$
\beta_{j}^{n}=e^{i z_{n} / 2}\left(\left|k_{1}^{2}-n_{\alpha}^{2}\right|\right)^{1 / 2}
$$

## - Case 1B: Perfectly conducting grating: TM case

Lemma 17. $U$ is a solution of (2.27) if and only if it is a solution of the following problem set in the truncated domain $\Omega \backslash$ int $\Omega_{3}$

$$
\begin{equation*}
\Delta U(x, y)+k_{1}^{2} U(x, y)=0, \quad(x, y) \in \Omega \backslash \Omega_{3}, \tag{2.102}
\end{equation*}
$$

with the DtN map interface conditions at the boundaries of the truncated region given by

$$
\begin{array}{ll}
\left.\partial_{n} U(x, y)\right|_{\Gamma_{+}}=T_{+} U+g(x) & \\
\left.\partial_{n} U(x, y)\right|_{\Gamma_{-}}=T_{-} U & \\
x \in \Gamma_{+},
\end{array}
$$

the Neumann boundary conditions at the surface of the diffraction grating

$$
\begin{equation*}
\partial_{n} U(x, y)=0, \quad(x, y) \in \partial \Omega_{3}, \tag{2.103}
\end{equation*}
$$

and the $\alpha$-quasi-periodic condition

$$
U(d, y)=e^{i \alpha d} U(0, y), \quad y \in[-B, B] .
$$

Here $T_{ \pm}$and $g$ are given by equations (2.98), (2.100) and (2.99) in Lemma 16.

## - Case 2A: Transmitting dielectric grating: TE case

Lemma 18. $U$ is a solution of (2.29) if and only if it is a solution of the following problem set in the truncated domain $\Omega$

$$
\begin{equation*}
\Delta U(x, y)+k(x, y)^{2} U(x, y)=0, \quad(x, y) \in \Omega \tag{2.104}
\end{equation*}
$$

with the DtN map interface conditions at the boundaries of the truncated region given by

$$
\begin{aligned}
& \left.\partial_{n} U(x, y)\right|_{\Gamma_{+}}=T_{+} U+g(x) \quad x \in \Gamma_{+}, \\
& \left.\partial_{n} U(x, y)\right|_{\Gamma_{-}}=T_{-} U \quad x \in \Gamma_{-},
\end{aligned}
$$

and the $\alpha$-quasi-periodic condition

$$
U(d, y)=e^{i \alpha d} U(0, y), \quad y \in[-B, B] .
$$

Here $T_{ \pm}$and $g$ are given by equations (2.98), (2.100) and (2.99) in Lemma 16.

## - Case 2B: Transmitting dielectric grating: TM case

Lemma 19. $U$ is a solution of (2.31) if and only if it is a solution of the following problem set in the truncated domain $\Omega$

$$
\begin{equation*}
\nabla \cdot\left(\frac{1}{k^{2}(x, y)} \nabla U(x, y)\right)+U(x, y)=0, \quad(x, y) \in \Omega \tag{2.105}
\end{equation*}
$$

with the DtN map interface conditions at the boundaries of the truncated region given by

$$
\begin{aligned}
& \left.\partial_{n} U(x, y)\right|_{\Gamma_{+}}=T_{+} U+g(x) \quad x \in \Gamma_{+}, \\
& \left.\partial_{n} U(x, y)\right|_{\Gamma_{-}}=T_{-} U \quad x \in \Gamma_{-},
\end{aligned}
$$

and the $\alpha$-quasi-periodic condition

$$
U(d, y)=e^{i \alpha d} U(0, y), \quad y \in[-B, B],
$$

Here $T_{ \pm}$and $g$ are given by equations (2.98), (2.100) and (2.99) in Lemma 16.
In the following proof we will just use $\mathbb{R}^{2}$ and $\Omega$, the proof will be similar when we deal with $\mathbb{R}_{+}^{2}$ and $\Omega \backslash$ int $\Omega_{3}$.

Proof. The proofs of Lemma 16, Lemma 17, Lemma 18 and Lemma 19 are similar. The first part of the proof consists of showing that $U$ satisfies in $\mathbb{R}^{2}$ (or in $\mathbb{R}_{+}^{2}$ )

$$
\Delta U+k^{2} U=0
$$

if and only if $\left.U\right|_{\Omega}=\hat{U}$ (or $\left.U\right|_{\Omega \backslash \text { int } \Omega_{3}}=\hat{U}$ ) satisfies in $\Omega$ (or in $\Omega \backslash$ int $\Omega_{3}$ )

$$
\Delta \hat{U}+k^{2} \hat{U}=0
$$

In fact, if $\Delta U+k^{2} U=0$ for all $(x, y) \in \mathbb{R}^{2}$, in particular $\Omega \subset \mathbb{R}^{2}$ then $\Delta U+k^{2} U=0$ in $\Omega$ i.e.

$$
\Delta \hat{U}+k^{2} \hat{U}=0
$$

Now let us suppose that $\Delta \hat{U}+k^{2} \hat{U}=0$, by using the $\operatorname{DtN}$ maps we match our inner solution $\hat{U}$ with the outer solution denoted by $\hat{U}_{C}$. Hence $U$ is given by

$$
U= \begin{cases}\hat{U}, & \text { for }(x, y) \in \Omega  \tag{2.106}\\ \hat{U}_{C}, & \text { for }(x, y) \in[0, d] \times \mathbb{R} \backslash \Omega\end{cases}
$$

and because we have $\Delta \hat{U}_{C}+k^{2} \hat{U}_{C}=0$ in $[0, d] \times \mathbb{R} \backslash \Omega$. In addition the grating is periodic therefore

$$
\Delta \hat{U}_{C}+k^{2} \hat{U}_{C}=0
$$

in $\mathbb{R}^{2}$. The $\alpha$-quasi-periodic condition and the Dirichlet and Neumann boundaries for the perfectly conducting grating follow naturally from the boundary conditions of the original scattering problem in $\mathbb{R}_{+}^{2}$ or $\mathbb{R}^{2}$.

The second part of the proof below will consider the continuity and smoothness of the solution at the truncated region interfaces $\Gamma_{+}$and $\Gamma_{-}$. We notice on one hand that when $|y| \geq B$, the wavenumber $k$ is a constant, and so we can set $U$ equal to the fundamental solution of the Helmholtz equation as given by Lemma 1. When $y \geq B, k=k_{1}$ is constant, and we have

$$
\begin{equation*}
U(x, y)=\sum_{n \in \mathbb{Z}} r_{n} e^{i \beta_{1}^{0} y} e^{i n_{\alpha} x}+t_{n} e^{-i \beta_{1}^{0} y} e^{i n_{\alpha} x} \tag{2.107}
\end{equation*}
$$

We use the condition that $U$ is composed of outgoing waves apart from $U_{I}$ (the incident wave) which is the only incoming wave allowed in $\Omega_{1}$ (corresponding to $n=0$ ); so we have $t_{n}=0$ for all $n \in \mathbb{Z} \backslash\{0\}$. On the other hand, we know that a quasi-periodic function can be written as a Fourier series by using Lemma 2 and so

$$
\begin{equation*}
U(x, y)=\sum_{n \in \mathbb{Z}} U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} \tag{2.108}
\end{equation*}
$$

Equating series (2.108) and (2.107) we see that for $n 60$

$$
r_{n} e^{i \beta_{1}^{n} B}=U^{\left(n_{\alpha}\right)}(B),
$$

that is

$$
\begin{equation*}
r_{n}=U^{\left(n_{\alpha}\right)}(B) e^{-i \beta_{1}^{n} B} . \tag{2.109}
\end{equation*}
$$

For $n=0$, with the amplitude of the incident wave $t_{0}=1$,

$$
r_{0} e^{i \beta_{1} B}+e^{-i \beta_{1} B}=U^{\left(0_{\alpha}\right)}(B)
$$

that is

$$
\begin{equation*}
r_{0}=U^{\left(0_{\alpha}\right)}(B) e^{-i \beta_{1} B}-e^{-2 i \beta_{1} B} \tag{2.110}
\end{equation*}
$$

From Lemma 1, the analytical solution in the region $y \geq B, x \in \mathbb{R}$, is

$$
\begin{equation*}
\hat{U}_{C}(x, y)=\sum_{n \in \mathbb{Z}} U^{\left(n_{\alpha}\right)}(B) e^{i \beta_{1}^{n}(y-B)} e^{i n_{\alpha} x}-e^{i \beta_{1}(y-2 B)} e^{i \alpha x}+e^{-i \beta_{1} y} e^{i \alpha x} \tag{2.111}
\end{equation*}
$$

Taking the normal derivative of $\hat{U}$ on $\Gamma_{+}$, (that is in the $y$ direction) gives

$$
\begin{align*}
\frac{\partial \hat{U}}{\partial n} & =\frac{\partial \hat{U}_{C}}{\partial n} \\
& =\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} U^{\left(n_{\alpha}\right)}(B) e^{i \beta_{1}^{n}(y-B)} e^{i n_{\alpha} x}-i \beta_{1}^{0} e^{i \beta_{1}^{0}(y-2 B)} e^{i \alpha x}-i \beta_{1} e^{-i \beta_{1} y} e^{i \alpha x}( \tag{2.112}
\end{align*}
$$

and so

$$
\begin{equation*}
\left.\frac{\partial \hat{U}}{\partial n}\right|_{y=B}=\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} U^{\left(n_{\alpha}\right)}(B) e^{i n_{\alpha} x}-2 i \beta_{1}^{0} e^{i \beta_{1}(-B)} e^{i \alpha x} \tag{2.113}
\end{equation*}
$$

From the definition of $T_{+}$given by (3) and if we denote

$$
\begin{equation*}
g(x)=-2 i \beta_{1} e^{-i \beta_{1} B} e^{i \alpha x} . \tag{2.114}
\end{equation*}
$$

We have

$$
\left.\frac{\partial \hat{U}}{\partial n}\right|_{y=B}=T_{+} \hat{U}(x)+g(x) .
$$

This ends the proof for the boundary conditions on the top boundary.
Let us apply the same argument for the boundary condition on the bottom boundary. When $y \leq-B, k=k_{2}$ is constant, and so from Lemma 1 we can write

$$
\begin{equation*}
U(x, y)=\sum_{n \in \mathbb{Z}} r_{n} e^{i \beta_{2}^{n} y} e^{i n_{\alpha} x}+t_{n} e^{-i \beta_{2}^{n} y} e^{i n_{\alpha} x} . \tag{2.115}
\end{equation*}
$$

We use the condition that $U$ is composed of outgoing waves and so $r_{n}=0$ for all $n \in \mathbb{Z}$. On the other hand, for $(x, y) \in \Omega, U$ is quasi-periodic, therefore

$$
\begin{equation*}
U(x, y)=\sum_{n \in \mathbb{Z}} U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} . \tag{2.116}
\end{equation*}
$$

Equating series (2.115) and (2.116) we see that for all $n \in \mathbb{Z}$

$$
t_{n} e^{i \beta_{2}^{n} B}=U^{\left(n_{\alpha}\right)}(-B),
$$

that is

$$
\begin{equation*}
t_{n}=U^{\left(n_{\alpha}\right)}(-B) e^{-i \beta_{2}^{n} B} . \tag{2.117}
\end{equation*}
$$

From Lemma 1, the analytical solution in the region $y \leq-B, x \in \mathbb{R}$ is

$$
\begin{equation*}
\hat{U}_{C}(x, y)=\sum_{n \in \mathbb{Z}} U^{\left(n_{\alpha}\right)}(-B) e^{i \beta_{2}^{n}(y+B)} e^{i n_{\alpha} x} \tag{2.118}
\end{equation*}
$$

Taking the normal derivative of $\hat{U}$ on $\Gamma_{-}$, (that is, in the y direction) gives

$$
\begin{aligned}
\frac{\partial \hat{U}}{\partial n} & =\frac{\partial \hat{U}_{C}}{\partial n} \\
& =-\sum_{n \in \mathbb{Z}}-i \beta_{2}^{n} U^{\left(n_{\alpha}\right)}(-B) e^{-i \beta_{2}^{n}(y+B)} e^{i n_{\alpha} x}
\end{aligned}
$$

and so

$$
\begin{equation*}
\left.\frac{\partial \hat{U}}{\partial n}\right|_{y=-B}=\sum_{n \in \mathbb{Z}} i \beta_{2}^{n} U^{\left(n_{\alpha}\right)}(-B) e^{i n_{\alpha} x} \tag{2.119}
\end{equation*}
$$

From the definition of $T_{-}$given in (3), we have

$$
\left.\frac{\partial \hat{U}}{\partial n}\right|_{y=-B}=T_{-} \hat{U}(x) .
$$

which ends the proof on the bottom boundary. By noting that $U$ is given by (2.106), we can see that

$$
\lim _{y \rightarrow+\infty} \frac{\partial U}{\partial n}=\lim _{y \rightarrow+\infty} \frac{\partial \hat{U}_{C}}{\partial n}=0
$$

### 2.5 Summary

In this chapter, we have introduced the physical and mathematical description of the problem of diffraction when an electromagnetic wave interacts with a periodic grating. We have shown, by using Maxwell's equations, that our problem can be decomposed into two elementary mathematical problems which are the transverse magnetic and the transverse electric Helmholtz problems. For each problem, the grating can be perfectly conducting or transmitting, and so we study four cases. We have also introduced some transparent boundary conditions (DtN maps) in order to solve the problem numerically. When we formulated the problem in Chapter 2, the incident wave was included via the boundary conditions. In Chapter 3 we will consider an equivalent but alternative formulation that incorporates the incident wave via an inhomogeneous forcing term (with compact support) in the Helmholtz equation. We introduce the inhomogeneous Helmholtz problem to allow us to study the regularity of the solution in order to construct an a priori error estimate of the solution to the problem stated in this chapter, in Chapters 4 and 6.

## Chapter 3

## A regularity estimate for the inhomogeneous Helmholtz problem for periodic gratings

### 3.1 General case

To get to the heart of the matter, we will just focus on the perfectly conducting grating interacting with a TE wave (Case 1A) in this chapter. We also illustrate the transmitting dielectric grating interacting with a TE wave (Case 2A). This case is needed for a special case where the multiple scattering problem is reduced into a one dimensional single scattering problem. Hence, we can investigate the robustness of our regularity result using the literature. We will derive the regularity results corresponding to Case 1B and Case 2B in Appendix C.

### 3.1.1 Case 1A: Perfectly conducting grating: TE case

Let $f(x, y) \in L^{2}\left(\mathbb{R}_{+}^{2}\right)$ represent the forcing term in the inhomogeneous Helmholtz equation. In this chapter, we want to study the regularity of the solution $U(x, y)$ of the inhomogeneous Helmholtz problem depending on the function $f(x, y)$. The regularity of the solution $U(x, y)$ will enable us to study the a priori error estimation of the approximate solution when we solve the Helmholtz problem numerically.
We use the same notation as in Chapter 2 for the spatial domains as shown in Figure 2.1. We also assume that $f$ is local with respect to the $y$ direction which means that $\operatorname{supp} f \subset \mathbb{R} \times[-B, B]$ (see Figure 2.3).

### 3.1.1.1 The inhomogeneous Helmholtz equation

In the presence of surface current $\left(\underline{J}=\left(J_{x}, J_{y}, J_{z}\right) \boldsymbol{6} \underline{0}\right.$ ) or surface charge ( $\rho \mathbf{6}$ 0 ) (given in equations (2.2) to (2.4)), we can eliminate $\underline{H}$ and we can get the differential equation for $\underline{E}$ using equations (2.1) and (2.2) with the constitutive
relations given by equations (2.6) and (2.7). That is [64, p. 8], we obtain

$$
\begin{equation*}
\nabla \times \nabla \times \underline{E}-w^{2} \varepsilon \mu \underline{E}=-i w \mu \underline{J} . \tag{3.1}
\end{equation*}
$$

Similarly, we can eliminate $\underline{E}$ to get the equation for $\underline{H}$, giving

$$
\begin{equation*}
\nabla \times\left(\frac{1}{\varepsilon} \nabla \times \underline{H}\right)-w^{2} \mu \underline{H}=\nabla \times\left(\frac{1}{\varepsilon} \underline{J}\right) . \tag{3.2}
\end{equation*}
$$

Here we are solving for $U=E_{z}$ for a given function $f(x, y) \in L^{2}\left(\mathbb{R}_{+}^{2}\right)$. The inhomogeneous Helmholtz problem is to find $U(x, y) \in C^{2}\left(\mathbb{R}_{+}^{2}\right)$ such that

$$
\begin{array}{rrr}
\Delta U(x, y)+k_{1}^{2} U(x, y)= & f(x, y), & (x, y) \in \mathbb{R}_{+}^{2} \\
U(x, y) & = & 0,  \tag{3.4}\\
& (x, y) \in \partial \mathbb{R}_{-}^{2} .
\end{array}
$$

subject to the radiation condition

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{3.5}
\end{equation*}
$$

Putting $\underline{E}=\left(0,0, E_{z}\right)$ and $\underline{J}=\left(J_{x}, J_{y}, J_{z}\right)$ in equation (3.1) we find that $f(x, y)=i w \mu J_{z}$. We now utilize the periodicity of the grating and restrict our problem to the vertical single strip $S=[0, d] \times \mathbb{R}$ as shown in Figure 2.2.

On this vertical single strip $U$ and $f$ are $\alpha$-quasi-periodic with respect to $x$. In addition, we choose our solution $U$ to satisfy the upward propagating radiation condition [28]. In order to study the regularity of our solution $U$, we first need to study the $\alpha$-quasi-periodic fundamental solution of the problem (3.3).

### 3.1.1.2 The $\alpha$-quasi-periodic Green functions of the Helmholtz equation

Definition 20. The Green's function $G$ [106] is defined as

$$
\begin{equation*}
G(x, y)=-\frac{1}{2 d} \sum_{n \in \mathbb{Z}} \frac{e^{i n_{\alpha} x+i \beta_{j}^{n}|y|}}{-i \beta_{j}^{n}} \tag{3.6}
\end{equation*}
$$

for $(x, y) \in \mathbb{R}^{2}$, with $n_{\alpha}, \beta_{j}^{n}$ and $z_{n}$ defined by equations (2.42), (2.43) and (2.44) extended to be valid for $j \in\{0,1,2,3\}$.

We then have the following lemma.
Lemma 21. For any $p \in \mathbb{Z}$, if $G$ is $\alpha$-quasi-periodic with respect to $x$, then

$$
G(x, y)=e^{i \alpha p d} G(x-p d, y)
$$

for $(x, y) \in \mathbb{R}^{2}$.

Proof. We may write

$$
\begin{aligned}
G(x, y) & =-\frac{1}{2 d} \sum_{n \in \mathbb{Z}} \frac{e^{i n_{\alpha}(x-p d+p d)+i \beta_{j}^{n}|y|}}{-i \beta_{j}^{n}} \\
& =-\frac{1}{2 d} \sum_{n \in \mathbb{Z}} e^{i \alpha p d} \frac{e^{i n_{\alpha}(x-p d)+i \beta_{j}^{n}|y|}}{-i \beta_{j}^{n}}
\end{aligned}
$$

where the last line is justified by the fact that

$$
e^{i n_{\alpha} p d}=e^{i \frac{2 \pi n}{d} p d+i \alpha p d}=e^{i \alpha p d} .
$$

Then, $G$ is $\alpha$-quasi-periodic .
To ease the notation, let us define the following functions which will be used often in this section.

Definition 22. For any $\left\{y, y_{0}\right\} \in \mathbb{R}$, and for $n \in \mathbb{Z}$, define

$$
g_{n}\left(y, y_{0}\right)=\frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}
$$

and

$$
g_{-n}\left(y, y_{0}\right)=\frac{e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}
$$

such that $\beta_{j}^{n}$ is given by equation (2.43).
Let us also define for any $\{m, n\} \in \mathbb{Z},\left\{x, x_{0}\right\} \in[0, d] \subset \mathbb{R}$

$$
h_{m n}\left(x, x_{0}\right)=e^{i m_{\alpha} x_{0}} e^{i n_{\alpha}\left(x-x_{0}\right)},
$$

such that $m_{\alpha}$ and $n_{\alpha}$ are given by equation (2.43).
In the following four results, we derive an expression for the problem solution $U$ by using $G$ and we define $S_{y}$ where $[0, d] \times S_{y}=S \backslash \overline{\Omega_{3}}$.

Theorem 23. Case 1A: Let $f(x, y) \in L_{\alpha \#}^{2}\left(S \backslash \Omega_{3}\right)$ and let $U \in C^{2}\left(S \backslash \Omega_{3}\right)$ satisfy the inhomogeneous Helmholtz equation given by equation (3.3). Then, the solution is given by

$$
\begin{equation*}
U(x, y)=\int_{S \backslash \overline{\Omega_{3}}} G\left(x-x_{0}, y-y_{0}\right) f\left(x_{0}, y_{0}\right) d x_{0} d y_{0} \tag{3.7}
\end{equation*}
$$

for $(x, y) \in S \backslash \overline{\Omega_{3}}$ and $U=0$ elsewhere.

Proof. In order to prove Theorem 23, since we know that $G$ is $\alpha$-quasi-periodic, we can focus on one period. We use the property that $f$ is $\alpha$-quasi-periodic, to write $f$ in its Fourier form

$$
\begin{equation*}
f(x, y)=\sum_{m \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) e^{i m_{\alpha} x} \tag{3.8}
\end{equation*}
$$

where

$$
\begin{equation*}
f^{\left(m_{\alpha}\right)}(y)=\frac{1}{d} \int_{0}^{d} f(x, y) e^{-i m_{\alpha} x} d x \tag{3.9}
\end{equation*}
$$

and so, $U$, as given by equation (3.7), can be written in the following form

$$
\begin{align*}
U(x, y) & =\frac{1}{2 d} \int_{S} \sum_{m \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) e^{i m_{\alpha} x_{0}} \sum_{n \in \mathbb{Z}} \frac{e^{i n_{\alpha}\left(x-x_{0}\right)+i \beta_{j}^{n}\left|y-y_{0}\right|}}{-i \beta_{j}^{n}} d x_{0} d y_{0}, \\
& =\frac{1}{2 d} \int_{S} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) e^{i m_{\alpha} x_{0}} \frac{e^{i n_{\alpha}\left(x-x_{0}\right)+i \beta_{j}^{n}\left|y-y_{0}\right|}}{-i \beta_{j}^{n}} d x_{0} d y_{0} . \tag{3.10}
\end{align*}
$$

By using Definition 22, we note

$$
\partial_{y} g_{n}\left(y, y_{0}\right)= \begin{cases}-\frac{e^{i \beta_{j}^{n}\left(y-y_{0}\right)}}{}, & y \geq y_{0},  \tag{3.11}\\ \frac{e^{-i \beta_{j}^{n}\left(y-y_{0}\right)}}{2}, & y \leq y_{0} .\end{cases}
$$

Therefore, we have the following relation, which captures the jump of the partial derivative of $g_{n}$ at $y=y_{0}$

$$
\begin{align*}
\partial_{y} g_{n}\left(y^{-}, y\right)-\partial_{y} g_{n}\left(y^{+}, y\right) & =1 / 2-(-1 / 2), \\
& =1 \tag{3.12}
\end{align*}
$$

We use Definition 22 to rewrite the function $U$ as defined by equation (3.10) as follows

$$
U(x, y)=\frac{1}{d} \int_{S} \sum_{m, n \in \mathbb{Z}} h_{m n}\left(x, x_{0}\right) g_{n}\left(y, y_{0}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d x_{0} d y_{0}
$$

Denote by $S_{y}^{+}$the interval in $S_{y}$ such that $y_{0}>y$ and by $S_{y}^{-}$the interval in $S_{y}$ such that $y_{0}<y$. Then

$$
\begin{aligned}
U(x, y)= & \frac{1}{d} \int_{[0, d]}\left(\int_{S_{y}^{-} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) g_{n}\left(y, y_{0}\right) d y_{0}\right. \\
& \left.+\int_{S_{y}^{+} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) g_{n}\left(y, y_{0}\right) d y_{0}\right) d x_{0} .
\end{aligned}
$$

The partial derivative of the function $U(x, y)$ with respect to $y$ is then

$$
\begin{aligned}
\partial_{y} U(x, y)= & \frac{1}{d} \int_{[0, d]} \int_{S_{y}^{-} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) g_{n}\left(y^{-}, y\right) d x_{0} \\
& -\frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) g_{n}\left(y^{+}, y\right) d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{+} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} .
\end{aligned}
$$

Since, our function $g_{n}$ is continuous at $y_{0}$, we have

$$
g_{n}\left(y^{+}, y\right)-g_{n}\left(y^{-}, y\right)=0
$$

Consequently, the partial derivative of $U(x, y)$ with respect to $y$ can be written as follows

$$
\begin{align*}
\partial_{y} U(x, y)= & \frac{1}{d} \int_{[0, d]} \int_{S_{y}^{-} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{+} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} . \tag{3.13}
\end{align*}
$$

We differentiate one more time with respect to $y$ to give

$$
\begin{aligned}
\partial_{y}^{2} U(x, y)= & \frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) \partial_{y} g_{n}\left(y^{+}, y\right) d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{-} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y}^{2} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} \\
& -\frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) \partial_{y} g_{n}\left(y^{-}, y\right) d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{+} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y}^{2} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0}, \\
= & \frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right)\left[\partial_{y} g_{n}\left(y^{+}, y\right)-\partial_{y} g_{n}\left(y^{-}, y\right)\right] d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{-} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y}^{2} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{+} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y}^{2} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0}, \\
= & \frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{-} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y}^{2} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} \\
& +\frac{1}{d} \int_{[0, d]} \int_{S_{y}^{-} \cup\{y\}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y}^{2} g_{n}\left(y, y_{0}\right) d y_{0} d x_{0} .
\end{aligned}
$$

The last equality is justified by using equation (3.12). This leads us to the following result

$$
\begin{align*}
\partial_{y}^{2} U(x, y)= & \frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0} \\
& +\frac{1}{d} \int_{S} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) \partial_{y}^{2} g_{n}\left(y, y_{0}\right) d x_{0} d y_{0} \tag{3.14}
\end{align*}
$$

Definition 22 and equation (3.11) give

$$
\begin{equation*}
\partial_{y}^{2} g_{n}\left(y, y_{0}\right)=-\left(\beta_{j}^{n}\right)^{2} g_{n}\left(y, y_{0}\right), \tag{3.15}
\end{equation*}
$$

and hence

$$
\begin{align*}
\partial_{y}^{2} U(x, y)= & \frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0} \\
& +\frac{1}{d} \int_{S} \sum_{m, n \in \mathbb{Z}}-\left(\beta_{j}^{n}\right)^{2} g_{n}\left(y, y_{0}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) d x_{0} d y_{0} .(3 . \tag{3.16}
\end{align*}
$$

To calculate the partial derivative of second order of $U(x, y)$, with respect to $x$, we use equation (3.10) and we follow the same argument as above since $e^{i n_{\alpha}\left(x-x_{0}\right)} \in$ $C^{\infty}([0, d])$ and we have

$$
\begin{equation*}
\partial_{x}^{2} U(x, y)=\frac{1}{d} \int_{S} \sum_{m, n \in \mathbb{Z}}-n_{\alpha}^{2} h_{m n}\left(x, x_{0}\right) g_{n}\left(y, y_{0}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d x_{0} d y_{0} \tag{3.17}
\end{equation*}
$$

Now, we can combine equations (3.16) and (3.17) to compute the Laplacian of the function $U(x, y)$ and so

$$
\begin{aligned}
\Delta U(x, y)= & \partial_{x}^{2} U(x, y)+\partial_{y}^{2} U(x, y), \\
= & \frac{1}{d} \int_{S} \sum_{m, n \in \mathbb{Z}}\left(-n_{\alpha}^{2}+\left(\beta_{j}^{n}\right)^{2}\right) g_{n}\left(y, y_{0}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) d x_{0} d y_{0} \\
& +\frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0} \\
= & -k^{2} U(x, y)+\frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0} .
\end{aligned}
$$

Hence

$$
\begin{aligned}
\Delta U(x, y)+k_{1}^{2} U(x, y) & =\frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0}, \\
& =\sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) \frac{1}{d} \int_{[0, d]} e^{i m_{\alpha} x_{0}} e^{i n_{\alpha}\left(x-x_{0}\right)} d x_{0}, \\
& =\sum_{n \in \mathbb{Z}} f^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} .
\end{aligned}
$$

The last equation is justified by the fact that

$$
\begin{align*}
\frac{1}{d} \int_{0}^{d} e^{i\left(m_{\alpha}-n_{\alpha}\right) x_{0}} d x_{0} & =\frac{1}{d} \int_{0}^{d} e^{i \frac{2 \pi n}{d}(m-n) x_{0}} d x_{0},  \tag{3.18}\\
& =\delta_{m n}
\end{align*}
$$

Therefore, by equation (3.8), we have shown that

$$
\Delta U+k_{1}^{2} U=f(x, y), \forall(x, y) \in S
$$

In order for $U$ given by equation (3.7) to be the solution of equation (3.3), we need to show that $U$ satisfies all the boundary conditions. It follows from equation (3.7) that $U=0$ on $\partial \Omega_{3}$. We also note that $f(x, y)$ and $G(x, y)$ are both $\alpha$-quasiperiodic and therefore $U(x, y)$ is also $\alpha$-quasi-periodic. Also $G(x, y)$ is composed of bounded outgoing waves and $f(x, y)$ has a locally compact support with respect to $y$ therefore $U(x, y)$ satisfies

$$
\lim _{|y| \rightarrow+\infty} U(x, y)=0
$$

which finishes the proof.

### 3.1.1.3 Regularity of the solution of Helmholtz problem for periodic grating

In this section, we use the $\alpha$-quasi-periodic fundamental solution $G$ to establish the regularity of our solution which means that we will try to bound its norm and its partial derivative by using some constants multiplied by the norm of the forcing term.

Lemma 24. Let $U \in C^{2}\left(S \backslash \Omega_{3}\right)$ be the solution of equation (3.3). Then, for any $(x, y)$, $\left(x_{0}, y_{0}\right)$ in $S \backslash \Omega_{3}$, if we define

$$
\begin{equation*}
V(x, y)=\sup _{n \in \mathbb{Z}} \frac{1}{2 \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha} x} f^{\left(n_{\alpha}\right)}(y), \tag{3.19}
\end{equation*}
$$

then we have

$$
\mathrm{k} U(x, y) \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq \mathrm{k} V(x, y) \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} .
$$

Proof. We can use Theorem 23 with the definition of $G$ given by Definition 20 and equation (3.8) to get

$$
U(x, y)=\frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} \int_{S_{y}} \frac{e^{i \beta_{1}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{1}^{n}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0} d x_{0}
$$

in $S \backslash \Omega_{3}$ and $U=0$ elsewhere. We know that the profile of the grating $\partial \Omega_{3}$ can be described by $y=P(x)$ therefore if $y \boldsymbol{\sigma} y_{0}$, and if the profile $\partial \Omega_{3}$ is as shown as in Figure 3.2, we have

$$
\begin{aligned}
U(x, y) & =\frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} \int_{S_{y}} \frac{e^{i \beta_{1}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{1}^{n}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0} d x_{0} \\
& =\frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}}\left(F(y,+\infty)-F\left(y, P\left(x_{0}\right)\right)\right) d x_{0}
\end{aligned}
$$

such that $F\left(y, y_{0}\right)=\int \frac{e^{i \beta^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0}$. By the upward propagating radiation condition, we have $F(y,+\infty)=0$, and so for $y$ ध $y_{0}$

$$
\begin{aligned}
U(x, y) & =-\frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} F\left(y, P\left(x_{0}\right)\right) d x_{0} \\
& =0
\end{aligned}
$$

because $\sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} F\left(y, P\left(x_{0}\right)\right)$ is continuous and it is periodic of period $d$ with respect to $x_{0}$. If the profile is as shown as in Figure 3.1, we have

$$
\begin{aligned}
U(x, y) & =\frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} \int_{S_{y}} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{1}^{n}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0} d x_{0}, \\
& =\frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}}(F(y,+\infty)-F(y,-\infty)) d x_{0}
\end{aligned}
$$

because the profile $\partial \Omega_{3}$ is a closed curve and the function to integrate is analytic so its contribution is zero from Cauchy's theorem [88, p. 4]. From the upward propagating radiation condition, we have $F(y, \pm \infty)=0$, and so for $y \in y_{0} U(x, y)=0$. Hence, the only contribution comes from $y=y_{0}$ and so,

$$
U(x, y)=-\frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} \frac{1}{-2 i \beta_{1}^{n}} f^{\left(m_{\alpha}\right)}(y) d x_{0}
$$

This leads us to

$$
\mathrm{k} U(x, y) \mathrm{k}_{L_{\alpha \#}^{2}(S)}^{2} \leq \sup _{n \in \mathbb{Z}} \frac{1}{4\left\|\beta_{1}^{n}\right\|_{\infty}^{2}} \sum_{n \in \mathbb{Z}}\left|f^{\left(n_{\alpha}\right)}(y)\right|^{2}=\mathrm{k} V(x, y) \mathrm{k}_{L_{\alpha \#}^{2}(S)}^{2},
$$

using equation (3.18).
In order to study the regularity, we need an upper bound of the norm of the partial derivatives of $U$ in term of the norm of $U$ itself.

Lemma 25. For any $\gamma=\left(\gamma_{1}, \gamma_{2}\right)$ such that $\gamma_{j} \in \mathbb{N}$, for $j=1,2$, and for $x \in[0, d]$ and $y \in S_{y}$, then

$$
\begin{aligned}
\mathrm{k} \partial_{x}^{\gamma_{1}} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} & \leq \sup _{n \in \mathbb{Z}} \mathrm{k}_{\alpha} \mathrm{k}_{\infty}^{\gamma_{1}} \mathrm{k} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}, \\
\mathrm{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} & \leq \sup _{n \in \mathbb{Z}} \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{\gamma_{2}} \mathrm{k} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} .
\end{aligned}
$$

Proof. Taking the partial derivative of $U$ with respect to $x, y, \gamma_{1}$ times, and using equation (3.35) gives

$$
\partial_{x}^{\gamma_{1}} U=\sum_{m, n \in \mathbb{Z}} \int_{0}^{d}\left(i n_{\alpha}\right)^{\gamma_{1}} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \int_{S_{y}} \frac{e^{i \beta_{1}^{n} \mid y-y_{0}} \mid f^{\left(m_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{1}^{n}} d y_{0} d x_{0}
$$

and so

$$
\begin{aligned}
& \mathbf{k} \partial_{x}^{\gamma_{1}} U \mathbf{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq \\
& \sup _{n \in \mathbb{Z}} \mathbf{k} n_{\alpha} \mathbf{k}_{\infty}^{\gamma_{1}} \left\lvert\, \sum_{m, n \in \mathbb{Z}} \int_{0}^{d} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \int_{S_{y}} \frac{e^{i \beta_{1}^{n}\left|y-y_{0}\right|} f^{\left(m_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{1}^{n}} d y_{0} d x_{0}\right. \|_{L^{2}\left(S \backslash \Omega_{3}\right)}, \\
& \leq \sup _{n \in \mathbb{Z}} \mathbf{k} n_{\alpha} \mathbf{k}_{\infty}^{k_{1}} \mathbf{k} U \mathbf{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} .
\end{aligned}
$$

We can do exactly the same with the partial derivative of $U$ with respect to $y$,

$$
\partial_{y}^{\gamma_{2}} U=\sum_{m, n \in \mathbb{Z}} \int_{0}^{d} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \int_{S_{y}}\left( \pm i \beta_{1}^{n}\right)^{\gamma_{2}} \frac{e^{i \beta_{1}^{n}\left|y-y_{0}\right|} f^{\left(m_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{1}^{n}} d y_{0} d x_{0}
$$



Figure 3.1: A perfectly conducting grating over one period where the profile $\partial \Omega_{3}$ of the scatterer is a closed curve.


Figure 3.2: A perfectly conducting grating over one period where the profile $\partial \Omega_{3}$ of the scatterer is open.
and so

$$
\begin{aligned}
& \mathbf{k} \partial_{y}^{\gamma_{2}} U \mathbf{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq \\
& \sup _{n \in \mathbb{Z}} \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}^{\gamma_{2}}\left\|\sum_{m, n \in \mathbb{Z}} \int_{0}^{d} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \int_{S_{y}} \frac{e^{i \beta_{1}^{n}\left|y-y_{0}\right|} f^{\left(m_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{1}^{n}} d y_{0} d x_{0}\right\|_{L^{2}\left(S \backslash \Omega_{3}\right)}, \\
& \leq \sup _{n \in \mathbb{Z}} \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}^{\gamma_{2}} \mathbf{k} U \mathbf{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} .
\end{aligned}
$$

Next, we give an approximation of $U$ using $f$.
Lemma 26. For any $x \in[0, d]$ and for any $y \in S_{y}$, we have for any function $f \in L_{\alpha \#}^{2}(S)$ and for $U$ as given by equation (3.10)

$$
\mathbf{k} U(x, y) \mathbf{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq \sup _{n \in \mathbb{Z}} \frac{1}{2 \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}} \mathbf{k} f(x, y) \mathbf{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}
$$

Proof. Using Lemma 24 we have

$$
\begin{aligned}
\mathrm{k} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} & \leq \sup _{n \in \mathbb{Z}} \frac{1}{2 \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}} \mathrm{k} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha} x} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}, \\
& \leq \sup _{n \in \mathbb{Z}} \frac{1}{2 \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}} \mathrm{k} f(x, y) \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)},
\end{aligned}
$$

from equation (3.8).
This leads to the following regularity result for Case 1A.
Theorem 27. For any $\gamma=\left(\gamma_{1}, \gamma_{2}\right)$ such that $\gamma_{j} \in \mathbb{N}$, for $j=1,2$, and for $x \in$ $[0, d], y \in S_{y}$ there exists a constant $C_{\text {reg }}$ which is independent of the wavenumber $k$ such that the solution $U$ of equation (3.3) given by Theorem 23 satisfies

$$
\mathrm{k} D^{\gamma} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq C_{r e g} \mathrm{k} k \mathrm{k}_{\infty}^{|\gamma|-1} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)},
$$

where $\mathrm{k} D^{\gamma} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}$ is as given in Definition A-10.
Proof. We use Lemma 25 with Lemma 26, to give

$$
\begin{aligned}
& \mathrm{k} \partial_{x}^{\gamma_{1}} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq \sup _{n \in \mathbb{Z}} \frac{\mathrm{k} n_{\alpha} \mathrm{k}_{\infty}^{\gamma_{1}}}{2 \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \\
& \mathrm{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq \sup _{n \in \mathbb{Z}} \frac{\mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}^{\prime 2}}{2 \mathbf{k} \beta_{1}^{n} \mathbf{k}_{\infty}} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}
\end{aligned}
$$

$n_{\alpha}$, and $\beta_{1}^{n}$ are given by equations (2.42) and (2.43) and hence

$$
\begin{aligned}
n_{\alpha} & =\frac{2 \pi n}{d}+k_{1} \sin \theta, \\
& =k_{1}\left(\frac{2 \pi n}{d k_{1}}+\sin \theta\right), \\
& =k_{1} \sin \theta_{n},
\end{aligned}
$$

where we define $\theta_{n} \in \mathbb{C}$, and

$$
\begin{aligned}
\beta_{1}^{n} & =e^{i z_{n} / 2} \sqrt{\left|k_{1}^{2}-n_{\alpha}^{2}\right|}, \\
& =e^{i z_{n} / 2} \sqrt{\left|k_{1}^{2}\left(1-\sin \theta_{n}^{2}\right)\right|} \\
& =\left|k_{1}\right| e^{i z_{n} / 2} \sqrt{\left.\mid 1-\sin \theta_{n}^{2}\right) \mid} \\
& =\left|k_{1}\right| e^{i z_{n} / 2} \cos \theta_{n} .
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\mathrm{k}_{x}^{\gamma_{1}} U \mathrm{k}_{L^{2}(\Omega)} & \leq \sup _{n \in \mathbb{Z}} \mathrm{k} \frac{\left(k_{1} \sin \theta_{n}\right)^{\gamma_{1}}}{\left|k_{1}\right| e^{i z_{n} / 2} \cos \theta_{n}} \mathrm{k}_{\infty} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}, \\
& \leq \sup _{n \in \mathbb{Z}} \mathrm{k} k_{1} \mathrm{k}_{\infty}^{\gamma_{1}-1} \frac{\left|\sin \theta_{n}\right|}{\left|\cos \theta_{n}\right|} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} .
\end{aligned}
$$

In a similar fashion,

$$
\begin{aligned}
\mathbf{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} & \leq \sup _{n \in \mathbb{Z}}\left(\mathrm{k}_{1} \mathrm{k}_{\infty}\left|e^{i z_{n} / 2} \cos \theta_{n}\right|\right)^{\gamma_{2}-1} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}, \\
& \leq \sup _{n \in \mathbb{Z}} \mathrm{k} k_{1} \mathbf{k}_{\infty}^{\gamma_{2}-1}\left(\left|\cos \theta_{n}\right|\right)^{\gamma_{2}-1} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)},
\end{aligned}
$$

We finish by setting

$$
C_{\text {reg }}=2 \sup _{n \in \mathbb{Z}}\left(\left|\cos \theta_{n}\right|^{\gamma_{2}-1}, \frac{\left|\sin \theta_{n}\right|^{\gamma_{1}}}{\left|\cos \theta_{n}\right|}\right)
$$

which is well defined since $\beta_{1}^{n} \in 0$, and hence $\cos \theta_{n} \sigma 0$.

### 3.1.2 Case 2A: Transmitting dielectric grating: TE case

### 3.1.2.1 The inhomogeneous Helmholtz equation

Let $f(x, y) \in L^{2}\left(\mathbb{R}^{2}\right)$ represent the forcing term in the inhomogeneous Helmholtz equation for transmitting dielectric gratings. Similar to Case 1A, we want to study the regularity of the solution $U(x, y)$ of the inhomogeneous Helmholtz problem depending on the function $f(x, y)$ for Case 2 A . The derivation is similar to that used to derive equation (3.3) but we have the following boundary conditions on each interface separating two different media with different electric permittivities [64, p. 10]

$$
\begin{align*}
\underline{n} \times\left(\underline{E_{1}}-\underline{E_{2}}\right) & =0,  \tag{3.20}\\
\underline{n} \cdot\left(\underline{D_{1}}-\underline{D_{2}}\right) & =\rho,  \tag{3.21}\\
\underline{n} \times\left(\underline{H_{1}}-\underline{H_{2}}\right) & =\underline{J},  \tag{3.22}\\
\underline{n} \cdot\left(\underline{B_{1}}-\underline{B_{2}}\right) & =0, \tag{3.23}
\end{align*}
$$

where $E_{j}, D_{j}, H_{j}$ and $B_{j}$ represent respectively the electric field, the electric displacement, the magnetic field and the magnetic flux corresponding to the first medium for $j=1$ and to the other medium for $j=2$.

Here we are solving for $U=E_{z}$ for a given function $f(x, y) \in L^{2}\left(\mathbb{R}^{2}\right)$. The inhomogeneous Helmholtz problem is to find $U(x, y) \in C^{2}\left(\mathbb{R}^{2}\right)$ such that

$$
\begin{equation*}
\Delta U(x, y)+k^{2}(x, y) U(x, y)=f(x, y), \quad(x, y) \in \mathbb{R}^{2} \tag{3.24}
\end{equation*}
$$

subject to the radiation condition

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{3.26}
\end{equation*}
$$

and the interface conditions given by equations (3.20), (3.21), (3.22) and (3.23). Putting $\underline{E}=\left(0,0, H_{z}\right)$ and $\underline{H}=\left(H_{x}, H_{y}, 0\right)$, in equation (3.1) gives

$$
f(x, y)=i w \mu J_{z} .
$$

We now utilize the periodicity of the grating and restrict our problem to the vertical single strip $S=[0, d] \times \mathbb{R}$ as shown in Figure 2.2. Hence we define the wavenumber $k$ to be

$$
k(x, y)= \begin{cases}k_{1} \in \mathbb{R} & \text { for }(x, y) \in \Omega_{1},  \tag{3.27}\\ k_{0} \in \mathbb{C} & \text { for }(x, y) \in \Omega_{0} \backslash \Omega_{3}, \\ k_{3} \in \mathbb{C} & \text { for }(x, y) \in \Omega_{3}, \\ k_{2} \in \mathbb{C} & \text { for }(x, y) \in \Omega_{2}\end{cases}
$$

for Case 2A. In a similar way to Case $1 \mathrm{~A}, U$ and $f$ are $\alpha$-quasi-periodic with respect to $x$ on this vertical single strip. We again choose our solution $U$ to satisfy the upward propagating radiation condition [28]. To study the regularity of our solution $U$, we also need to study the $\alpha$-quasi-periodic fundamental solution of the problem (3.24).

### 3.1.2.2 The $\alpha$-quasi-periodic Green functions of the Helmholtz equation

The solution $U$ can be found as follows.
Theorem 28. Case 2A: Let $f(x, y) \in L_{\alpha \#}^{2}(S)$ and let $U \in C^{2}(S)$ satisfy the inhomogeneous Helmholtz equation given by equation (3.24) in $S$. Then, the solution $U$ of equation (3.24) is given by

$$
\begin{equation*}
U(x, y)=\int_{S} G_{j}\left(x-x_{0}, y-y_{0}\right) f\left(x_{0}, y_{0}\right) d x_{0} d y_{0} \tag{3.28}
\end{equation*}
$$

for $j \in\{0,1,2,3\}$ with

$$
\begin{equation*}
G_{j}(x, y)=-\frac{1}{2 d} \sum_{n \in \mathbb{Z}} c_{j}^{n} \frac{e^{i n_{\alpha} x+i \beta_{j}^{n}|y|}}{-i \beta_{j}^{n}}-\frac{1}{2 d} \sum_{n \in \mathbb{Z}} d_{j}^{n} \frac{e^{i n_{\alpha} x-i \beta_{j}^{n}|y|}}{-i \beta_{j}^{n}}, \tag{3.29}
\end{equation*}
$$

with $d_{j}^{n}=0$ for $j=1,2$ (upward propagating radiation condition). Each subdomain of $S$ where $k=k_{j}$ is constant is denoted by $S_{j}$, and for $l \in\{0,1,2,3\}$, the coefficients $c_{j}^{n}, c_{l}^{n}$ and $d_{l}^{n}$ are chosen such that the boundary conditions on the interface separating $S_{j}$ and $S_{l}$, given by equations (3.20), (3.21), (3.22) and (3.23) are satisfied.

Proof. We use the same argument as in Theorem 23 in each subdomain $S_{j}$ of $S$ and let

$$
\begin{align*}
& S_{j x}=[0, d], j \in\{0,1,2\}, \\
& S_{3 x} \times \mathbb{R} \backslash S_{y}=\Omega_{3} \tag{3.30}
\end{align*}
$$

For a given $y$ then $S_{j x}$ is the range of $x$ values in domain $S_{j}$ and so

$$
\begin{aligned}
\Delta U= & \frac{1}{d} \int_{S_{j}} \sum_{m, n \in \mathbb{Z}} k_{j}^{2}\left(c_{j}^{n} g_{n}\left(y, y_{0}\right)+d_{j}^{n} g_{-n}\left(y, y_{0}\right)\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) d x_{0} d y_{0} \\
& +\frac{1}{d} \int_{S_{j x}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0}
\end{aligned}
$$

for $(x, y) \in S_{j}$. Hence

$$
\begin{aligned}
\Delta U(x, y)= & \sum_{j} \frac{1}{d} \int_{S_{j}} \sum_{m, n \in \mathbb{Z}} k_{j}^{2}\left(c_{j}^{n} g_{n}\left(y, y_{0}\right)+d_{j}^{n} g_{-n}\left(y, y_{0}\right)\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) h_{m n}\left(x, x_{0}\right) d x_{0} d y_{0} \\
& +\sum_{j} \frac{1}{d} \int_{S_{j x}} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0}, \\
= & -k^{2} U(x, y)+\frac{1}{d} \int_{[0, d]} \sum_{m, n \in \mathbb{Z}} f^{\left(m_{\alpha}\right)}(y) h_{m n}\left(x, x_{0}\right) d x_{0} .
\end{aligned}
$$

We finish the proof by using equations (3.18) and (3.8) and we have

$$
\Delta U(x, y)+k_{j}^{2} U(x, y)=f(x, y) .
$$

The $\alpha$-quasi-periodicity of $U(x, y)$ follows from the $\alpha$-quasi-periodicity of $f(x, y)$ and $G_{j}(x, y)$. Because $f$ is locally compact with respect to $y$ and $G$ is composed of upward propagating radiation condition waves, we have

$$
\lim _{|y| \rightarrow+\infty} U(x, y)=0 .
$$

### 3.1.2.3 Regularity of the solution of Helmholtz problem for periodic grating

In this section, we use the $\alpha$-quasi-periodic fundamental solution $G$ (or $G_{j}$ ) to establish the regularity of each solution which means that we will try to bound the norm of each solution and its partial derivative by using some constants times the norm of the forcing term. Let us now study the regularity for Case 2A. Combining equations (3.28), (3.6) and (3.8) gives

$$
\begin{align*}
U(x, y)= & \frac{1}{d} \int_{[0, d]} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)} \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} \\
& \int_{\mathbb{R}}\left(c_{j}^{n} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}+d_{j}^{n} \frac{e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0} d x_{0} . \tag{3.31}
\end{align*}
$$

Before studying the regularity of our solution $U(x, y)$, let us state the following lemmas. We have the following result.

Lemma 29. For any $\left\{y, y_{0}\right\} \in \mathbb{R}$, if

$$
I_{m n}\left(y, y_{0}\right)=\int_{\mathbb{R}} \sum_{n \in \mathbb{Z}}\left(c_{j}^{n} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}+d_{j}^{n} \frac{e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0}
$$

such that $c_{j}^{n}$ and $d_{j}^{n}$ are the complex scalars defined in equation (3.29) then, we have

$$
\begin{align*}
I_{m n}\left(y, y_{0}\right) & =0 & & \text { if } y_{0} \in y  \tag{3.32}\\
I_{m n}(y, y) & =\sum_{n \in \mathbb{Z}} \frac{1}{\beta_{j}^{n}}\left(c_{j}^{n}+d_{j}^{n}\right) f^{\left(m_{\alpha}\right)}(y) & & \text { if } y_{0}=y . \tag{3.33}
\end{align*}
$$

Proof. When $y=y_{0}$, the result is immediate

$$
I_{m n}(y, y)=\sum_{n \in \mathbb{Z}}\left(c_{j}^{n}+d_{j}^{n}\right) \frac{1}{\beta_{j}^{n}} f^{\left(m_{\alpha}\right)}(y)
$$

Now, let us suppose that $y_{0} \sigma y$. Therefore, we have

$$
\begin{aligned}
I_{m n}\left(y, y_{0}\right)= & \int_{\mathbb{R}-\{y\}} \sum_{n \in \mathbb{Z}}\left(c_{j}^{n} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}+d_{j}^{n} \frac{e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0}, \\
= & \int_{y_{+}}^{+\infty} \sum_{n \in \mathbb{Z}}\left(\frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}+d_{j}^{n} \frac{e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0} \\
& +\int_{-\infty}^{y-} \sum_{n \in \mathbb{Z}}\left(\frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}+d_{j}^{n} \frac{e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0} .
\end{aligned}
$$

Since $\sum_{n \in \mathbb{Z}}\left(\frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}+d_{j}^{n} \frac{e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}}{\beta_{j}^{n}}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0}$ is continuous with respect to $y_{0}$ then $I_{m n}$ is continuous with respect to $y_{0}$. Therefore,

$$
I_{m n}\left(y, y_{0}\right)=I_{m n}(y,+\infty)-I_{m n}\left(y, y_{+}\right)+I_{m n}\left(y, y_{-}\right)-I_{m n}(y,-\infty)
$$

From the continuity of $I$,

$$
I_{m n}\left(y, y_{0}\right)=I_{m n}(y,+\infty)-I_{m n}(y,-\infty)
$$

The outgoing wave boundary conditions state that $f^{\left(m_{\alpha}\right)}\left(y_{0}\right)$ and $e^{i \beta_{j}^{n}\left|y-y_{0}\right|}$ tend to zero at $\pm \infty$, and so $I_{m n}=0$.

From now on, we denote $I_{m n}(y, y)$ by $I_{m n}(y)$. We have the following inequality.
Lemma 30. For any $(x, y)$, $\left(x_{0}, y_{0}\right)$ which belong to $[0, d] \times \mathbb{R}$, if we define by $V$ the function

$$
\begin{equation*}
V(x, y)=\sup _{n \in \mathbb{Z}, j} \frac{\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)}{\mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha} x} f^{\left(n_{\alpha}\right)}(y), \tag{3.34}
\end{equation*}
$$

for $n \in \mathbb{Z}$ and $j=0,1,2,3$ then we have

$$
\mathrm{k} U(x, y) \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq \mathrm{k} V(x, y) \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)},
$$

such that $U$ is given by equation (3.31).
Proof. We have

$$
\begin{gathered}
U(x, y)=\frac{1}{d} \int_{[0, d] \times \mathbb{R}} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha}\left(x-x_{0}\right)}\left(c_{j}^{n} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}+d_{j}^{n} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}\right) \\
\sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d x_{0} d y_{0} .
\end{gathered}
$$

Now $\sum_{n \in \mathbb{Z}}\left(c_{j}^{n} \frac{e^{i \beta^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}+d_{j}^{n} \frac{e^{i \beta^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}\right)$ and $f$ are continuous with compact support and the integral is well defined so we can use Fubini's theorem [104, p. 110] to interchange the order of summation and integration to get

$$
\begin{align*}
U(x, y)= & \frac{1}{d} \sum_{n \in \mathbb{Z}} \int_{[0, d] \times \mathbb{R}} e^{i n_{\alpha}\left(x-x_{0}\right)}\left(c_{j}^{n} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}+d_{j}^{n} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|}}{-2 i \beta_{j}^{n}}\right) \\
& \sum_{m \in \mathbb{Z}} e^{i m_{\alpha} x_{0}} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d x_{0} d y_{0}, \\
= & \frac{1}{d} \sum_{m, n \in \mathbb{Z}} \int_{[0, d]} \frac{e^{i n_{\alpha}\left(x-x_{0}\right)}}{-2 i} e^{i m_{\alpha} x_{0}} I_{m n}\left(y, y_{0}\right) d x_{0}, \tag{3.35}
\end{align*}
$$

Using equation (3.33) then

$$
\mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}}^{2}([0, d] \times \mathbb{R}) \leq \sup _{n \in \mathbb{Z}} \frac{\left|c_{j}^{n}+d_{j}^{n}\right|^{2}}{4 \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{2}} \frac{1}{d} \sum_{n \in \mathbb{Z}} \int_{[0, d]}\left|f^{\left(n_{\alpha}\right)}(y)\right|^{2} d x_{0},
$$

using equation (3.18) and so

$$
\mathbf{k} U \mathbf{k}_{L_{\alpha \neq}^{2}([0, d] \times \mathbb{R})}^{2} \leq \sup _{n \in \mathbb{Z}} \frac{\sup _{n, j}\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)^{2}}{\mathrm{k} \beta_{j}^{n} \mathbf{k}_{\infty}^{2}} \sum_{n \in \mathbb{Z}}\left|f^{\left(n_{\alpha}\right)}(y)\right|^{2}
$$

for $j=0,1,2,3$. This leads us to

$$
\begin{aligned}
\mathrm{k} U(x, y) \mathrm{k}_{L_{\alpha \#}^{2}([0, d] \times \mathbb{R})}^{2} & \leq \sup _{n \in \mathbb{Z}} \frac{\sup _{n, j}\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)^{2}}{\left\|\beta_{j}^{n}\right\|_{\infty}^{2}} \sum_{n \in \mathbb{Z}}\left|f^{\left(n_{\alpha}\right)}(y)\right|^{2} \\
& \leq \mathrm{k} V(x, y) \mathrm{k}_{L_{\alpha \#}^{2}}^{2}([0, d] \times \mathbb{R})
\end{aligned}
$$

using equation (3.18).
From now on, we use the function $V(x, y)$ given by equation (3.34) to study the regularity of the function $U(x, y)$ defined by equation (3.31). We have the following Lemma.

Lemma 31. For any $\gamma=\left(\gamma_{1}, \gamma_{2}\right)$ such that $\gamma_{q} \in \mathbb{N}$, and $x \in[0, d]$ and $y \in \mathbb{R}$, there exists $N_{0} \in \mathbb{N}$, with $c_{j}^{n}$ and $d_{j}^{n}$ equal to zero when $|n|>N_{0}$, and there exists $k_{\text {ref }}<\left|k_{j}\right|$ such that

$$
\begin{aligned}
& \mathrm{k} \partial_{x}^{\gamma_{1}} U \mathrm{k}_{L^{2}(S)} \leq \sup _{n \in \mathbb{Z}} \mathrm{k} n_{\alpha} \mathbf{k}_{\infty}^{\gamma_{1}} \mathbf{k} U \mathrm{k}_{L^{2}(S)}, \\
& \mathrm{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}(S)} \leq \sup _{n \in \mathbb{Z}, j} \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{\gamma_{2}} \mathrm{k} U \mathrm{k}_{L^{2}(S)}+C\left(k_{0}, k_{3}\right) C_{s} \sup _{n \in \mathbb{Z}, j}\left|\beta_{j}^{n}\right|^{\gamma_{2}-2} \mathbf{k} f \mathrm{k}_{L^{2}(S)}
\end{aligned}
$$

for $j \in\{0,1,2,3\}$ where

$$
\begin{align*}
C\left(k_{0}, k_{3}\right) & =\sup _{j \in\{0,3\}}\left(e^{\sin z_{n} / 2\left|k_{j}\right| \frac{\sup \left(\frac{2 \pi N_{0}}{k_{0}},|\alpha|\right)}{k_{r e f}} \sup _{\left\{y_{0}, y\right\} \in \partial \Omega_{3}}\left|y-y_{0}\right|}, 1\right),  \tag{3.36}\\
C_{s} & =\sup _{n \in \mathbb{Z}, j \in\{0,1,2,3\}}\left(\gamma_{2}-1\right)\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right) \tag{3.37}
\end{align*}
$$

with $\beta_{j}^{n}$ and $z_{n}$ as given by equations (2.43) and (2.44) (extended for $j \in\{0,1,2,3\}$ ).
Proof. Taking the partial derivative of $U$ with respect to $x, y, \gamma_{1}$ times, and using equation (3.35) gives

$$
\begin{aligned}
& \partial_{x}^{\gamma_{1}} U= \sum_{m, n \in \mathbb{Z}} \int_{0}^{d}\left(i n_{\alpha}\right)^{\gamma_{1}} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \\
& \int_{\mathbb{R}} \frac{c_{j}^{n} e^{i \beta_{j}^{n}}\left|y-y_{0}\right|}{+d_{j}^{n} e^{-i \beta_{j}^{n}}\left|y-y_{0}\right|} f^{\left(m_{\alpha}\right)}\left(y_{0}\right) \\
&-2 i \beta_{j}^{n}
\end{aligned} y_{0} d x_{0},
$$

and so

$$
\begin{aligned}
\mathbf{k}_{x}^{\gamma_{1}} U \mathbf{k}_{L^{2}(S)} \leq & \sup _{n \in \mathbb{Z}} \mathbf{k} n_{\alpha} \mathbf{k}_{\infty}^{\gamma_{1}} \\
& \| \sum_{m, n \in \mathbb{Z}} \int_{0}^{d} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \\
& \int_{\mathbb{R}} \frac{c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}+d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|} f^{\left(m_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{j}^{n}} d y_{0} d x_{0} \|_{L^{2}(S)}, \\
\leq & \sup _{n \in \mathbb{Z}} \mathbf{k} n_{\alpha} \mathbf{k}_{\infty}^{\gamma_{1}} \mathbf{k} U \mathbf{k}_{L^{2}(S)} .
\end{aligned}
$$

We can do exactly the same with the partial derivative of $U$ with respect to $y$, but we need to take into account that $\left.\left.\sum_{n \in \mathbb{Z}} \frac{\left(c_{j}^{n} e^{i \beta_{j}^{n}} \mid y-y_{0}\right.}{}\right|_{d_{j}^{n} e^{-i \beta_{j}^{n}}\left|y-y_{0}\right|} ^{-2 i \beta_{j}^{n}}\right) f^{\left(m_{\alpha}\right)}\left(y_{0}\right) \quad$ is no longer continuous for $\gamma_{2}>1$ on each interface separating each medium of constant wavenumber $k_{j}$. Therefore

$$
\begin{aligned}
& \partial_{y}^{\gamma_{2}} U=\sum_{m, n \in \mathbb{Z}} \int_{0}^{d} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \int_{\mathbb{R}}\left( \pm i \beta_{j}^{n}\right)^{\gamma_{2}} \frac{\left(c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}+d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right)}{-2 i \beta_{j}^{n}} \\
& f^{\left(m_{\alpha}\right)}\left(y_{0}\right) d y_{0} d x_{0}+\sum_{p=2}^{\gamma_{2}}\left[\sum_{n \in \mathbb{Z}} \frac{\left(c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}+d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right)}{-2 i \beta_{j}^{n}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right]_{S_{j} \cap S_{l}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right),
\end{aligned}
$$

such that $\left[\sum_{n \in \mathbb{Z}} \frac{\left(c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|} f^{\left(n n_{\alpha}\right)}\left(y_{0}\right)+d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right) f^{\left(n_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{j}^{n}}\right]_{S_{j} \cap S_{l}}$ denotes the jump at the interface separating $S_{j}$ with $S_{l}$ with $j, l \in 0,1,2,3$ and is given by

$$
\begin{align*}
& \sum_{n \in \mathbb{Z}}\left(\frac{\left(i \beta_{j}^{n}\right)^{p-1}\left(c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}+d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right)}{-2 i \beta_{j}^{n}}\right. \\
& \left.-\frac{\left(i \beta_{l}^{n}\right)^{p-1}\left(c_{j}^{n} e^{i \beta_{l}^{n}\left|y-y_{0}\right|}+d_{j}^{n} e^{-i \beta_{l}^{n}\left|y-y_{0}\right|}\right)}{-2 i \beta_{l}^{n}}\right) f^{\left(n_{\alpha}\right)}(y) \tag{3.38}
\end{align*}
$$

for $y, y_{0} \in S_{j} \cap S_{l}$. Hence,

$$
\begin{aligned}
& \left|\left[\sum_{n \in \mathbb{Z}} \frac{\left(c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}+d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right) f^{\left(n_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{j}^{n}}\right]_{S_{j} \cap S_{l}}\right| \leq \\
& \frac{1}{2} \sup _{n \in \mathbb{Z}, j \in\{0,1,2,3\}}\left|\beta_{j}^{n}\right|^{p-2} \sum_{n \in \mathbb{Z}}\left|c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}+c_{l}^{n} e^{i \beta_{l}^{n}\left|y-y_{0}\right|}\right|\left|f^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right| \\
& +\frac{1}{2} \sup _{n \in \mathbb{Z}, j \in\{0,1,2,3\}}\left|\beta_{j}^{n}\right|^{p-2} \sum_{n \in \mathbb{Z}}\left|d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}+d_{l}^{n} e^{-i \beta_{l}^{n}\left|y-y_{0}\right|}\right|\left|f^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right| \\
& \leq \sup _{n \in \mathbb{Z}, l \in\{0,1,2,3\}}\left(\left|\beta_{j}^{n}\right|^{p-2}\left|c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}\right|\right) \sum_{n \in \mathbb{Z}}\left|f^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right| \\
& \sup _{n \in \mathbb{Z}, l \in\{0,1,2,3\}}\left(\left|\beta_{j}^{n}\right|^{p-2}\left|d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right|\right) \sum_{n \in \mathbb{Z}}\left|f^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right| .
\end{aligned}
$$

First, we note that when $j \in\{1,2\}, S_{j} \cap S_{0}$ is given by $y= \pm b$ as shown in Figure 2.2, therefore $y-y_{0}=0$ on the interface and so

$$
\begin{equation*}
\left|e^{ \pm i \beta_{j}^{n} \mid y-y_{0}}\right|=1 . \tag{3.39}
\end{equation*}
$$

We also note that

$$
\begin{equation*}
\left|e^{i \beta_{j}^{n}\left|y-y_{0}\right|}\right| \leq 1 \tag{3.40}
\end{equation*}
$$

because $\mathfrak{I}\left(\beta_{j}^{n}\right)>0$ and

$$
\begin{equation*}
\left|e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right| \leq e^{\sin z_{n} / 2\left|k_{j}\right| \frac{\sup \left(\frac{2 \pi N_{0}}{d},|\alpha|\right.}{k_{r e f}}} \sup _{\left\{y_{0}, y\right\} \in \partial \Omega_{3}}\left|y-y_{0}\right| \tag{3.41}
\end{equation*}
$$

because $\mathfrak{J}\left(\overline{\beta_{j}^{n}}\right)=\sin z_{n} / 2\left|k_{j}^{2}-n_{\alpha}^{2}\right|^{1 / 2}$ as given by equations (2.43) and (2.44). We also note from the remark after Theorem C-2, there exists $N_{0}$ such that for $|n|>$ $N_{0}, d_{j}^{n}$ equals to zero. Therefore, $\left|k_{j}^{2}-n_{\alpha}^{2}\right| \leq\left|k_{j}^{2}\right|\left|1-\frac{n_{\alpha}^{2}}{k_{j}^{2}}\right| \leq 4\left|k_{j}^{2}\right| \frac{\sup \left(\left(\frac{2 \pi N_{0}}{d}\right)^{2},|\alpha|^{2}\right)}{k_{r e f}^{2}}$ such that $\left|k_{j}\right|>k_{\text {ref }}$ for $j \in\{0,1,2,3\}$. Combining equations (3.39), (3.40) and (3.41), we have

$$
\begin{aligned}
& \left|\left[\sum_{n \in \mathbb{Z}} \frac{\left(c_{j}^{n} e^{i \beta_{j}^{n}\left|y-y_{0}\right|}+d_{j}^{n} e^{-i \beta_{j}^{n}\left|y-y_{0}\right|}\right) f^{\left(n_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{j}^{n}}\right]_{S_{j} \cap S_{l}}\right| \leq \sup _{n \in \mathbb{Z}, j \in\{0,1,2,3\}} \\
& \left|\beta_{j}^{n}\right|^{p-2}\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right) \sup \left(e^{\sin z_{n} / 2\left|k_{j}\right| \frac{N_{0}}{k_{r e f}} \sup _{\left\{y_{0}, y\right\} \in S_{0} \cap S_{3}}\left|y-y_{0}\right|}, 1\right) \sum_{n \in \mathbb{Z}}\left|f^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right| \\
& \leq \sup _{n \in \mathbb{Z}, j \in\{0,1,2,3\}} C_{s 0}\left(e^{\sin z_{n} / 2\left|k_{j}\right| \frac{N_{0}}{k_{r e f}} \sup _{\left\{y_{0}, y\right\} \in S_{0} \cap S_{3}}\left|y-y_{0}\right|}, 1\right)\left|\beta_{j}^{n}\right|^{p-2} \mathbf{k} f \mathbf{k}_{L^{2}(S)} .
\end{aligned}
$$

with $C_{s 0}=\sup _{n \in \mathbb{Z}, j \in\{0,1,2,3\}}\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)$. Hence

$$
\begin{aligned}
& \mathbf{k} \partial_{y}^{\gamma_{2}} U \mathbf{k}_{L^{2}(S)} \leq \sup _{n \in \mathbb{Z}} \mathbf{k} \beta_{j}^{n} \mathbf{k}_{\infty}^{\gamma_{2}} \\
& \left\|\sum_{m, n \in \mathbb{Z}} \int_{0}^{d} e^{i n_{\alpha}\left(x-x_{0}\right)} e^{i m_{\alpha}\left(x_{0}\right)} \int_{\mathbb{R}} \frac{e^{i \beta_{j}^{n}\left|y-y_{0}\right|} f^{\left(m_{\alpha}\right)}\left(y_{0}\right)}{-2 i \beta_{j}^{n}} d y_{0} d x_{0}\right\|_{L^{2}(S)} \\
& +\sum_{p=2}^{\gamma_{2}}\left(C_{s 0}\left(e^{\sin z_{n} / 2\left|k_{j}\right| \frac{N_{0}}{k_{r e f}} \sup _{\left\{y_{0}, y\right\} \in S_{0} \cap S_{3}}\left|y-y_{0}\right|}, 1\right) \sup _{n \in Z, j \in\{0,1,2,3\}}\left|\beta_{j}^{n}\right|^{p-2}\right) \mathbf{k} f \mathbf{k}_{L^{2}(S)}, \\
& \leq \sup _{n \in \mathbb{Z}} \mathbf{k} \beta_{j}^{n} \mathbf{k}_{\infty}^{\gamma_{2}} \mathbf{k} U \mathbf{k}_{L^{2}(S)}+C_{s 0}\left(e^{\sin z_{n} / 2}\left|k_{j}\right| \frac{N_{0}}{k_{r e f}} \sup _{\left\{y_{0}, y\right\} \in S_{0} \cap S_{3}}\left|y-y_{0}\right|\right. \\
& \\
& \left(\sum_{p=2}^{\gamma_{2}} \sup _{n \in Z, j \in\{0,1,2,3\}}\left|\beta_{j}^{n}\right|^{p-2}\right) \mathbf{k} f \mathbf{k}_{L^{2}(S)}, \\
& \leq \sup _{n \in \mathbb{Z}} \mathbf{k} \beta_{j}^{n} \mathbf{k}_{\infty}^{\gamma_{2}^{2}} \mathbf{k} U \mathbf{k}_{L^{2}(S)}+C_{s 0}\left(e^{\sin z_{n} / 2}\left|k_{j}\right| \frac{N_{0}}{k_{r e f}} \sup _{\left\{y_{0}, y\right\} \in S_{0} \cap S_{3}}\left|y-y_{0}\right|\right. \\
& \left(\gamma_{2}-1\right) \\
& \sup _{n \in Z, j \in\{0,1,2,3\}}\left|\beta_{j}^{n}\right|^{\gamma_{2}-2} \mathbf{k} f \mathbf{k}_{L^{2}(S)} .
\end{aligned}
$$

We finish the proof by denoting $C_{s}=C_{s 0}\left(\gamma_{2}-1\right)$
We can now give an upper bound on the solution in terms of the forcing term.
Lemma 32. For any $x \in[0, d]$ and for any $y \in \mathbb{R}$, we have for any function $f \in L_{\alpha \#}^{2}([0, d] \times \mathbb{R})$ and for $U$ as given by equation (3.31)

$$
\mathrm{k} U(x, y) \mathrm{k}_{L^{2}(S)} \leq \sup _{n \in \mathbb{Z}, j} \frac{\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)}{\mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}} \mathrm{k} f(x, y) \mathrm{k}_{L^{2}(S)}
$$

for $j=0,1,2,3$
Proof. Using Lemma 30 we have

$$
\begin{aligned}
\mathbf{k} U \mathrm{k}_{L^{2}(S)} & \leq \sup _{n \in \mathbb{Z}, j} \frac{\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)}{\mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}} \mathrm{k} \sum_{n \in \mathbb{Z}} e^{i n_{\alpha} x} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}(S)}, \\
& \leq \sup _{j \in\{0,1,2,3\}, n \in \mathbb{Z}} \frac{\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)}{\mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}} \mathrm{k} f(x, y) \mathrm{k}_{L^{2}(S)},
\end{aligned}
$$

from equation (3.8) which finishes the proof.
We then have the following regularity result for Case 2A .
Theorem 33. For any $\gamma=\left(\gamma_{1}, \gamma_{2}\right)$ such that $\gamma_{j} \in \mathbb{N}$, for $j=1,2$, and for $x \in[0, d], y \in[-B, B] \subset \mathbb{R}$ there exists a constant $C_{\text {reg }}$ which is independent of
the wavenumber $k$ such that the solution $U$ of equation (3.24) given by Theorem 28 satisfies

$$
\mathrm{k} D^{\gamma} U \mathrm{k}_{L^{2}(\Omega)} \leq C_{\text {reg }}\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right) \mathrm{k} k \mathbf{k}_{\infty}^{|\gamma|-1} \mathrm{k} f \mathrm{k}_{L^{2}(\Omega)},
$$

with $C_{s}$ and $C\left(k_{0}, k_{3}\right)$ as given by equations (3.37) and (3.36) in Lemma 31 and $\mathrm{k} D^{\gamma} U \mathrm{k}_{L^{2}(\Omega)}$ is as given in Definition A-10.

Proof. We use Lemma 31 with Lemma 26, we have

$$
\begin{aligned}
\mathrm{k} \partial_{x}^{\gamma_{1}} U \mathrm{k}_{L^{2}(S)} & \leq \sup _{n \in \mathbb{Z}} \frac{\mathrm{k} n_{\alpha} \mathrm{k}_{\infty}^{\gamma_{1}}}{2 \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}} \mathrm{k} f \mathrm{k}_{L^{2}(S)}, \\
\mathrm{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}(S)} & \leq \sup _{n \in \mathbb{Z}} \frac{\mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{\gamma_{2}}}{2 \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}} \mathrm{k} f \mathrm{k}_{L^{2}(S)}+C\left(k_{0}, k_{3}\right) C_{s} \sup _{n \in \mathbb{Z}, j}\left|\beta_{j}^{n}\right|^{\gamma_{2}-2} \mathrm{k} f \mathrm{k}_{L^{2}(S)},
\end{aligned}
$$

with $C>1$ We note that $n_{\alpha}$, and $\beta_{j}^{n}$ are given by (2.43)

$$
\begin{aligned}
n_{\alpha} & =\frac{2 \pi n}{d}+k \sin \theta, \\
& =k\left(\frac{2 \pi n}{d k}+\sin \theta\right), \\
& =k \sin \theta_{n} .
\end{aligned}
$$

We proceed the same as with $n_{\alpha}$ for $\beta_{j}^{n}$, and we have

$$
\begin{aligned}
\beta_{j}^{n} & =e^{i z_{n} / 2} \sqrt{\left|k^{2}-n_{\alpha}^{2}\right|}, \\
& =e^{i z_{n} / 2} \sqrt{\left|k^{2}\left(1-\sin \theta_{n}^{2}\right)\right|} \\
& =|k| e^{i z_{n} / 2} \sqrt{\left.\mid 1-\sin \theta_{n}^{2}\right) \mid} \\
& =|k| e^{i z_{n} / 2} \cos \theta_{n},
\end{aligned}
$$

therefore, we have

$$
\begin{aligned}
& \mathrm{k} \partial_{x}^{\gamma_{1}} U \mathrm{k}_{L^{2}(S)} \leq \sup _{n \in \mathbb{Z}} \mathrm{k} \frac{\left(k \sin \theta_{n}\right)^{\gamma_{1}}}{|k| e^{i z_{n} / 2} \cos \theta_{n}} \mathrm{k}_{\infty} \mathrm{k} f \mathrm{k}_{L^{2}(S)} \\
& \leq \sup _{n \in \mathbb{Z}} \mathrm{k} k \mathrm{k}_{\infty}^{\gamma_{1}-1} \frac{\left|\sin \theta_{n}\right|}{\left|\cos \theta_{n}\right|}{ }^{\gamma_{1}} \\
& \mathrm{k} f \mathrm{k}_{L^{2}(S)}
\end{aligned}
$$

In a similar fashion,

$$
\begin{aligned}
& \mathrm{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}(S)} \leq \sup _{n \in \mathbb{Z}}\left(\mathrm{k} k \mathbf{k}_{\infty}\left|e^{i z_{n} / 2} \cos \theta_{n}\right|\right)^{\gamma_{2}-1} \mathrm{k} f \mathrm{k}_{L^{2}(S)} \\
& +C\left(k_{0}, k_{3}\right) C_{s} \sup _{n \in \mathbb{Z}, j}\left|\beta_{j}^{n}\right|^{\gamma_{2}-2} \mathrm{k} f \mathrm{k}_{L^{2}(S)}, \\
& \leq \sup _{n \in \mathbb{Z}}\left(\mathrm{k} k \mathbf{k}_{\infty}\left|e^{i z_{n} / 2} \cos \theta_{n}\right|\right)^{\gamma_{2}-1} \mathrm{k} f \mathrm{k}_{L^{2}(S)}+C\left(k_{0}, k_{3}\right) C_{s} \\
& \sup _{n \in \mathbb{Z}, j} \frac{\left(\mathrm{k} k \mathbf{k}_{\infty}\left|e^{i z_{n} / 2} \cos \theta_{n}\right|\right)^{\gamma_{2}-1}}{\mathrm{k} k \mathrm{k}_{\infty}\left|e^{i z_{n} / 2} \cos \theta_{n}\right|} \mathrm{k} f \mathrm{k}_{L^{2}(S)}, \\
& \leq \sup _{n \in \mathbb{Z}} \mathrm{k} k \mathbf{k}_{\infty}^{\gamma_{2}-1}\left(\left|\cos \theta_{n}\right|\right)^{\gamma_{2}-1} \mathrm{k} f \mathrm{k}_{L^{2}(S)} \\
& +C\left(k_{0}, k_{3}\right) C_{s} \sup _{n \in \mathbb{Z}, j} \frac{\left(\mathrm{k} k \mathbf{k}_{\infty}\left|e^{i z_{n} / 2} \cos \theta_{n}\right|\right)^{\gamma_{2}-1}}{k_{r e f}\left|\cos \theta_{n}\right|} \mathrm{k} f \mathrm{k}_{L^{2}(S)} .
\end{aligned}
$$

If we denote

$$
C_{r e g}=\sup _{n \in \mathbb{Z}}\left(\frac{\left|\sin \theta_{n}\right|^{\gamma_{1}}}{\left|\cos \theta_{n}\right|},\left|\cos \theta_{n}\right|^{\gamma_{2}-1}, \frac{1}{k_{r e f}\left|\cos \theta_{n}\right|}\right)
$$

which is well defined because $\beta_{j}^{n} \in 0$, therefore $\cos \theta_{n} \sigma 0$. Hence,

$$
\begin{aligned}
\mathbf{k} \partial_{y}^{\gamma_{2}} U \mathbf{k}_{L^{2}(S)} & \leq\left(C_{r e g} \mathbf{k} k \mathbf{k}_{\infty}^{\gamma_{2}-1}+C_{r e g} \mathbf{k} k \mathbf{k}_{\infty}^{\gamma_{2}-1} C\left(k_{0}, k_{3}\right) C_{s}\right) \mathrm{k} f \mathrm{k}_{L^{2}(S)}, \\
& \leq C_{r e g}\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right) \mathbf{k} k \mathbf{k}_{\infty}^{\gamma_{2}-1} \mathrm{k} f \mathrm{k}_{L^{2}(S)} .
\end{aligned}
$$

Since $\operatorname{supp} f \subset \Omega$, then

$$
\begin{aligned}
\mathrm{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}(\Omega)} & \leq\left(C_{r e g} \mathrm{k} k \mathbf{k}_{\infty}^{\gamma_{2}-1}+C_{r e g} \mathrm{k} k \mathbf{k}_{\infty}^{\gamma_{2}-1} C\left(k_{0}, k_{3}\right) C_{s}\right) \mathbf{k} f \mathrm{k}_{L^{2}(\Omega)} \\
& \leq C_{r e g}\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right) \mathbf{k} k \mathbf{k}_{\infty}^{\gamma_{2}-1} \mathbf{k} f \mathbf{k}_{L^{2}(\Omega)}
\end{aligned}
$$

### 3.2 A special case

In this section we want to check the accuracy of our regularity results. We study a special case of equation (3.24), when the wavenumber $k$ is independent of the $x$ direction and $\Omega_{3}$ does not exist; therefore there are three layers with different wavenumbers. In this case, by using the method of separation of variables, each of the Fourier coefficients $U^{\left(n_{\alpha}\right)}(y)$ of our solution satisfies the one-dimensional Helmholtz equation for $n_{\alpha}^{2}<k^{2}$. For the case $n_{\alpha}^{2}>k^{2}$ the field equation becomes transformed Poisson's equation and a detailed analysis of this case is also performed here. The regularity bound for the one-dimensional Helmholtz equation was studied in [85] for Case 2A. Hence, we can use a similar analysis to produce a regularity result for this special case here. In this section, we will show that the regularity result for this special case precisely matches that in Theorem 33.

### 3.2.1 Solution operator $N_{\beta_{j}^{n}}$

Let $f \in L_{\alpha \#}^{2}(\Omega)$ and $U \in C^{2}(\Omega)$. We have the Helmholtz equation for this special case

$$
\begin{align*}
\Delta U+k^{2} U & =f \quad \text { in } \Omega,  \tag{3.42}\\
\left.U(d, y)\right|_{\Gamma_{R}} & =\left.e^{i \alpha x} U(0, y)\right|_{\Gamma_{L}},  \tag{3.43}\\
\left.\frac{\partial U}{\partial n}\right|_{\Gamma_{+}} & =T_{+} U,  \tag{3.44}\\
\left.\frac{\partial U}{\partial n}\right|_{\Gamma_{-}} & =T_{-} U, \tag{3.45}
\end{align*}
$$

such that

$$
k(y)= \begin{cases}k_{1} & \text { in } \Omega_{1},  \tag{3.46}\\ k_{0} & \text { in } \Omega_{0}, \\ k_{2} & \text { in } \Omega_{2},\end{cases}
$$

where $k_{j} \in \mathbb{R}, j=0,1,2$ (so that we fulfill the conditions in [85]) and $\Omega$ is as given in Figure 2.3 in Chapter 2 and

$$
\begin{align*}
& \Gamma_{L}=\{(0, y) \in \Omega\},  \tag{3.47}\\
& \Gamma_{R}=\{(d, y) \in \Omega\} \tag{3.48}
\end{align*}
$$

The boundary condition is defined using the Dirichlet to Neumann map $T$ as given by Definition 3. Since $U$ and $f$ are quasi-periodic

$$
\begin{equation*}
U(x, y)=\sum_{n \in \mathbb{Z}} U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}, \tag{3.49}
\end{equation*}
$$

and

$$
\begin{equation*}
f(x, y)=\sum_{n \in \mathbb{Z}} f^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} \tag{3.50}
\end{equation*}
$$

Now, we can substitute equations (3.49) and (3.50) inside equation (3.42) to give

$$
\begin{equation*}
-\left(\frac{d^{2}}{d y^{2}}-\left(\beta_{j}^{n}\right)^{2}\right) U^{\left(n_{\alpha}\right)}(y)=f^{\left(n_{\alpha}\right)}(y), \text { for } y \in[-B, B], \forall n \in \mathbb{Z} \tag{3.51}
\end{equation*}
$$

and in equations (3.44) and (3.45) to give

$$
\begin{equation*}
\left.\frac{\partial}{\partial_{n}} U^{\left(n_{\alpha}\right)}(y)\right|_{y= \pm B}=i \beta_{j}^{n} U^{\left(n_{\alpha}\right)}( \pm B), j=1,2 . \tag{3.52}
\end{equation*}
$$

To study the regularity of $U$, we use equation (3.51). We start by approximating the partial derivatives of $U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}$.

Lemma 34. For all $n \in \mathbb{Z}$, and for all $p_{1} \in\{0,1,2\}$, there exists a constant $C_{1} \leq \sup \left(1, k^{-2}\right)$ such that if $\alpha=0$

$$
\begin{equation*}
\mathrm{k} \partial_{x}^{p_{1}}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathbf{k}_{L^{2}([-B, B])} \leq C_{1} \mathbf{k} \partial_{x}^{2}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathbf{k}_{L^{2}([-B, B])} \tag{3.53}
\end{equation*}
$$

If $\alpha=0$, there exists a constant $C_{2} \leq \sup \left(1, \frac{d^{2}}{4 \pi^{2}}\right)$ such that
$\mathbf{k} \partial_{x}^{p_{1}}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathbf{k}_{L^{2}([-B, B])} \leq C_{2} \mathbf{k} \partial_{x}^{2}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathrm{k}_{L^{2}([-B, B])}, \quad \forall n \in \mathbb{Z} \backslash\{0\}$.

Proof. If $n_{\alpha} \in 0$, and $\alpha \in 0$, then

$$
\begin{aligned}
\partial_{x}^{p_{1}}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) & =\left(i n_{\alpha}\right)^{p_{1}} U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} \\
& =\frac{\left(i n_{\alpha}\right)^{p_{1}}}{-n_{\alpha}^{2}} \partial_{x}^{2}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right)
\end{aligned}
$$

We note that

$$
\left|\frac{\left(i n_{\alpha}\right)^{p_{1}}}{-n_{\alpha}^{2}}\right| \leq 1
$$

because $p_{1} \leq 2$. If $n \boldsymbol{\sigma} 0$ and $\alpha=0$, since $\left|n_{\alpha}\right| \geq \frac{2 \pi}{d}$ then

$$
\left|\frac{\left(i n_{\alpha}\right)^{p_{1}}}{-n_{\alpha}^{2}}\right| \leq \sup \left(\frac{d^{2}}{4 \pi^{2}}, 1\right)
$$

Next we study the properties of the functions $U^{\left(n_{\alpha}\right)}(y)$ which correspond to the 1D case. We consider the two cases where $k^{2}>n_{\alpha}^{2}$ and where $k^{2}<n_{\alpha}^{2}$ separately. The case where $k^{2}=n_{\alpha}^{2}$ corresponds to the resonance phenomenon and is not considered here [95].

### 3.2.2 First case: $k^{2}>n_{\alpha}^{2}$

In this case, for each $n \in \mathbb{Z}$, equation (3.51) corresponds to the general Helmholtz equation in one dimension [85].

Definition 35. Let $N_{\beta_{j}^{n}}: L^{2}([-B, B]) \rightarrow L^{2}([-B, B])$ be an operator defined by

$$
N_{\beta_{j}^{n}} f^{\left(n_{\alpha}\right)}(y)=\int_{[-B, B]} G_{\beta_{j}^{n}}\left(y-y_{0}\right) f^{\left(n_{\alpha}\right)}\left(y_{0}\right) d y_{0},
$$

where $G_{\beta_{j}^{n}}: \mathbb{R} \backslash\{0\} \rightarrow \mathbb{C}$ corresponds to the Green's function of the operator $\chi_{\beta_{j}^{n}}:=\frac{d^{2}}{d y^{2}}-\left(\beta_{j}^{n}\right)^{2}$, with $\left(\beta_{j}^{n}\right)^{2}<0$. As a consequence we have

$$
G_{\beta_{j}^{n}}(y)=g_{\beta_{j}^{n}}(|y|),
$$

where

$$
\begin{equation*}
g_{\beta_{j}^{n}}(r):=-\frac{e^{i \beta_{j}^{n} r}}{2 i \beta_{j}^{n}} . \tag{3.55}
\end{equation*}
$$

This leads to the following result.
Lemma 36. Let $v^{\left(n_{\alpha}\right)}(y)=N_{\beta_{j}^{n}} f^{\left(n_{\alpha}\right)}(y)$ where $N_{\beta_{j}^{n}}$ is given by Definition 35. If $\left(\beta_{j}^{n}\right)^{2}<0$, then $v^{\left(n_{\alpha}\right)}$ is a solution of equation (3.51), and we have

$$
U^{\left(n_{\alpha}\right)}(y)=v^{\left(n_{\alpha}\right)}(y)=N_{\beta_{j}^{n}} f^{\left(n_{\alpha}\right)}(y)=\int_{[-B, B]} G_{\beta_{j}^{n}}\left(y-y_{0}\right) f^{\left(n_{\alpha}\right)}\left(y_{0}\right) d y_{0}
$$

where $U^{\left(n_{\alpha}\right)}(y)$ satisfies equation (3.51) [85].
Proof. The proof can be seen in detail in [85].

### 3.2.2.1 Properties of $U^{\left(n_{\alpha}\right)}$

We have shown that $U^{\left(n_{\alpha}\right)}(y)$ can be written in term of the solution operator $N_{\beta_{j}^{n}}$, when $\left(\beta_{j}^{n}\right)^{2}<0$. We also have $U^{\left(n_{\alpha}\right)}(y)=v^{\left(n_{\alpha}\right)}(y)$ given by Lemma 36. Let us investigate some properties of the $v^{\left(n_{\alpha}\right)}(y)$. From Lemma 36, we can apply the results given in [85].

Lemma 37 (Decomposition Lemma). Let $B 1 \geq B$. Then, there exists a constant $C>0$ depending only on $B 1$ and $\beta_{\text {ref }}$ such that $\mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}>\beta_{\text {ref }}>0$, and for any $f^{\left(n_{\alpha}\right)} \in L^{2}([-B, B])$, the function $v^{\left(n_{\alpha}\right)}$ given by

$$
v^{\left(n_{\alpha}\right)}(y)=N_{\beta_{j}^{n}} f^{\left(n_{\alpha}\right)}(y)=\int_{[-B, B]} G_{\beta_{j}^{n}}\left(y-y_{0}\right) f^{\left(n_{\alpha}\right)}\left(y_{0}\right) d y_{0}, \quad y \in[-B, B],
$$

satisfies

$$
\begin{align*}
& \left(\beta_{j}^{n}\right)^{-1} \mathrm{k} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{H^{2}([-B, B])}+\mathrm{k} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{H^{1}([-B, B])}+\beta_{j}^{n} \mathrm{k} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \\
& \leq C \mathrm{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} . \tag{3.56}
\end{align*}
$$

Proof. The proof of Lemma 37 was already given and can be seen in detail in [85].

We then have the following results.
Lemma 38. For all $p_{2} \in\{0,1,2\}$, we have

$$
\begin{equation*}
\mathrm{k} \partial_{y}^{p_{2}} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \leq \mathrm{k} \frac{\left(i \beta_{j}^{n}\right)^{p_{2}}}{\left(\beta_{j}^{n}\right)^{2}} \mathrm{k}_{\infty} \mathrm{k} \partial_{y}^{2} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \tag{3.57}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathrm{k} \partial_{y}^{2} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \leq C \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty} \mathrm{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \tag{3.58}
\end{equation*}
$$

such that $C$ is the constant given in equation (3.56) in Lemma 37.

Proof. We show Lemma 38, by noting that

$$
\begin{aligned}
v^{\left(n_{\alpha}\right)}(y) & =\int_{[-B, B]} \frac{-e^{i \beta_{j}^{n}\left(y-y_{0}\right)}}{2 i \beta_{j}^{n}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right) d y_{0} \\
\partial_{y}^{p_{2}} v^{\left(n_{\alpha}\right)}(y) & =\int_{[-B, B]}\left(i \beta_{j}^{n}\right)^{p_{2}} \frac{-e^{i \beta_{j}^{n}\left(y-y_{0}\right)}}{2 i \beta_{j}^{n}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right) d y_{0} \\
& =\int_{[-B, B]} \frac{\left(i \beta_{j}^{n}\right)^{p_{2}}}{\left(i \beta_{j}^{n}\right)^{2}} \frac{\left(\beta_{j}^{n}\right)^{2} e^{i \beta_{j}^{n}\left(y-y_{0}\right)}}{2 i \beta_{j}^{n}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right) d y_{0}
\end{aligned}
$$

we take the norm and we have

$$
\begin{aligned}
\mathrm{k} \partial_{y}^{p_{2}} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} & \leq \mathrm{k} \frac{\left(i \beta_{j}^{n}\right)^{p_{2}}}{\left(i \beta_{j}^{n}\right)^{2}} \mathrm{k}_{\infty} \mathrm{k} \int_{[-B, B]} \frac{-\left(i \beta_{j}^{n}\right)^{2} e^{i \beta_{j}^{n}\left(y-y_{0}\right)}}{2 i \beta_{j}^{n}} \mathrm{k}_{L^{2}([-B, B])} \\
& =\mathrm{k} \frac{\left(i \beta_{j}^{n}\right)^{p_{2}}}{\left(i \beta_{j}^{n}\right)^{2}} \mathrm{k}_{\infty} \mathrm{k} \partial_{y}^{2} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])}
\end{aligned}
$$

which ends the proof of equation (3.57). We also note that

$$
\begin{aligned}
\mathrm{k} \partial_{y}^{2} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} & =\int_{[-B, B]}\left|\frac{\left(\beta_{j}^{n}\right)^{2} e^{i \beta_{j}^{n}\left(y-y_{0}\right)}}{2 i \beta_{j}^{n}} f^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right|^{2} d y_{0} \\
& \leq \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{4} \mathrm{k} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])}^{2}, \\
& \leq \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{2} \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{2} \mathrm{k} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{H^{1}([-B, B])}^{2}
\end{aligned}
$$

from Definition A-10 with $l=1$ and $s=2$ and so

$$
\mathbf{k} \partial_{y}^{2} v^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \leq C^{2} \mathbf{k} \beta_{j}^{n} \mathbf{k}_{\infty}^{2} \mathbf{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])}^{2},
$$

from equation (3.56) which finishes the proof.
It then follows that
Lemma 39. Let $N_{-}=\left\{n \in \mathbb{Z}: n_{\alpha}^{2}<k^{2}\right\}$, there exists a constant $\epsilon>0$ and a constant $C$ depending on $\lambda, \beta_{\text {ref }}$ as defined in Lemma 37 such that $\forall p=\left(p_{1}, p_{2}\right)$ with $|p| \leq 2$, we have

$$
\sum_{n \in N_{-}} \mathbf{k} D^{p}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathbf{k}_{L^{2}(\Omega)} \leq \sqrt{C(\epsilon)} k^{|p|-1} \sum_{n \in N_{-}} \mathbf{k} f^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} \mathbf{k}_{L^{2}(\Omega)}
$$

where $C(\epsilon)=C^{2} \sup _{n \in N_{-}}\left(\frac{C_{3}}{\epsilon_{n}}+1\right)$ such that $C$ is the constant defined in equation (3.56), $C_{3}=\sup \left(C_{1}, C_{2}\right)$ which are given in Lemma 34 and $\epsilon_{n} \geq\left(\frac{k^{2}}{\left(n_{\alpha}\right)^{2}-1}\right)^{2 p_{2}}$.

We also have
Lemma 41. For $n \in \mathbb{Z}$ such that $n_{\alpha}^{2}>k^{2}$, and for $p_{2} \in\{0,1,2\}$ we have

$$
\mathrm{k} \partial_{y}^{p_{2}} U^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \leq \mathrm{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} .
$$

Proof. We know that $U^{\left(n_{\alpha}\right)}$ satisfies

$$
-\partial_{y}^{2} U^{\left(n_{\alpha}\right)}+\left(n_{\alpha}^{2}-k^{2}\right) U^{\left(n_{\alpha}\right)}=f^{\left(n_{\alpha}\right)}(y) .
$$

Using the Fourier transform of $U^{\left(n_{\alpha}\right)}$, we have

$$
\sum_{m \in \mathbb{Z}}\left(m^{2}+\left(n_{\alpha}^{2}-k^{2}\right)\right) U_{m}^{\left(n_{\alpha}\right)} e^{i m y}=f^{\left(n_{\alpha}\right)}(y)
$$

which can be written as

$$
\sum_{m \in \mathbb{Z}}\left(\frac{m^{2}+\left(n_{\alpha}^{2}-k^{2}\right)}{m^{2}}\right) \partial_{y}^{2}\left(U_{m}^{\left(n_{\alpha}\right)} e^{i m y}\right)=f^{\left(n_{\alpha}\right)}(y)
$$

Furthermore we know that $n_{\alpha}^{2}-k^{2}>0$, which means

$$
\begin{aligned}
& \mathrm{k} \sum_{m \in \mathbb{Z}}\left(\frac{m^{2}+\left(n_{\alpha}^{2}-k^{2}\right)}{m^{2}}\right) \partial_{y}^{2}\left(U_{m}^{\left(n_{\alpha}\right)} e^{i m y}\right) \mathrm{k}_{L^{2}([-B, B])} \\
& =\mathrm{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \geq \mathrm{k} \sum_{m \in \mathbb{Z}} \partial_{y}^{2}\left(U_{m}^{\left(n_{\alpha}\right)} e^{i m y}\right) \mathrm{k}_{L^{2}([-B, B])} .
\end{aligned}
$$

We apply Lemma 40 to finish the proof of Lemma 41.
We also need the following property
Lemma 42. For $\left(n_{\alpha}\right)^{2}>k^{2}$, we have the following result

$$
\begin{aligned}
\mathrm{k} U^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])} \leq & \left(n_{\alpha}^{2}-k^{2}\right)^{-1}\left(\mathbf{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])}\right. \\
& \left.+\mathbf{k} \partial_{y}^{2} U^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}([-B, B])}\right) .
\end{aligned}
$$

Proof. The result is straightforward by noting that $U^{\left(n_{\alpha}\right)}$ satisfies

$$
-\partial_{y}^{2} U^{\left(n_{\alpha}\right)}+\left(n_{\alpha}^{2}-k^{2}\right) U^{\left(n_{\alpha}\right)}=f^{\left(n_{\alpha}\right)}
$$

and

$$
\begin{gathered}
\left(n_{\alpha}^{2}-k^{2}\right) U^{\left(n_{\alpha}\right)}=f^{\left(n_{\alpha}\right)}+\partial_{y}^{2} U^{\left(n_{\alpha}\right)} \\
U^{\left(n_{\alpha}\right)}=\left(n_{\alpha}^{2}-k^{2}\right)^{-1}\left(f^{\left(n_{\alpha}\right)}+\partial_{y}^{2} U^{\left(n_{\alpha}\right)}\right)
\end{gathered}
$$

We take the norm to finish the proof.
Now let us denote by $N_{+}=\left\{n \in \mathbb{Z}: n_{\alpha}^{2}>k^{2}\right\}$. Then, we have the following result

Lemma 43. For all $n \in N_{+}$, and for $p=\left(p_{1}, p_{2}\right)$ with $|p| \leq 2$, there exists a constant $\epsilon^{+}>0$ such that

$$
\sum_{n \in N_{+}} \mathrm{k} D^{p}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathrm{k}_{L^{2}(\Omega)}^{2} \leq\left(\frac{1}{\epsilon^{+}}+2\right) \sum_{n \in N_{+}} \mathrm{k} f^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x} \mathbf{k}_{L^{2}(\Omega)}^{2}
$$

Proof. We have for $p=\left(p_{1}, p_{2}\right), p_{1}, p_{2} \in\{0,1,2\}$

$$
\begin{aligned}
\mathrm{k} D^{p} U^{n_{\alpha}}(y) e^{i n_{\alpha} x} \mathrm{k}_{L^{2}(\Omega)}^{2} & =\mathrm{k} \partial_{x}^{p_{1}}\left(U^{n_{\alpha}}(y) e^{i n_{\alpha} x}\right) \mathrm{k}_{L^{2}(\Omega)}^{2}+\mathrm{k} \partial_{y}^{p_{2}}\left(U^{n_{\alpha}}(y) e^{i n_{\alpha} x}\right) \mathrm{k}_{L^{2}(\Omega)}^{2} \\
& \leq \mathrm{k}-n_{\alpha}^{2} U^{n_{\alpha}}(y) e^{i n_{\alpha} x} \mathrm{k}_{L^{2}(\Omega)}^{2}+\mathrm{k} \partial_{y}^{2} U^{n_{\alpha}}(y) e^{i n_{\alpha} x} \mathbf{k}_{L^{2}(\Omega)}^{2},
\end{aligned}
$$

by using Lemma 40. We use Lemma 41 and Lemma 42 to give

$$
\begin{aligned}
\mathrm{k} D^{p} U^{n_{\alpha}}(y) e^{i n_{\alpha} x} \mathrm{k}_{L^{2}(\Omega)}^{2} & \leq\left(\frac{n_{\alpha}^{4}}{\left(n_{\alpha}^{2}-k^{2}\right)^{2}}+1\right) \mathrm{k} f^{n_{\alpha}}(y) \mathrm{k}_{L^{2}(\Omega)}^{2} \\
& =\left(\frac{1}{\left(1-\frac{k^{2}}{n_{\alpha}^{2}}\right)^{2}}+1\right) \mathrm{k} f^{n_{\alpha}}(y) \mathrm{k}_{L^{2}(\Omega)}^{2}
\end{aligned}
$$

Furthermore, we have $\frac{k^{2}}{n_{\alpha}^{2}}<1$, then if we denote $\epsilon_{n}^{+}=\left(1-\frac{k^{2}}{n_{\alpha}^{2}}\right)^{2}$, we have

$$
\mathrm{k} D^{p} U^{n_{\alpha}}(y) e^{i n_{\alpha} x} \mathrm{k}_{L^{2}(\Omega)}^{2} \leq\left(\frac{1}{\epsilon_{n}^{+}}+1\right) \mathrm{k} f^{n_{\alpha}}(y) \mathrm{k}_{L^{2}(\Omega)}^{2}
$$

We take the sum over $n \in N_{+}$and taking the $\epsilon^{+}=\inf _{n \in N_{+}} \epsilon_{n}^{+}$to end the proof

$$
\sum_{n \in N_{+}} \mathrm{k} D^{p}\left(U^{n_{\alpha}}(y) e^{i n_{\alpha} x}\right) \mathrm{k}_{L^{2}(\Omega)}^{2} \leq\left(\frac{1}{\epsilon^{+}}+1\right) \sum_{n \in N_{+}} \mathrm{k} f^{n_{\alpha}}(y) \mathrm{k}_{L^{2}(\Omega)}^{2} .
$$

### 3.2.4 Regularity of the solution corresponding to the special case

Finally, we can use Lemma 39 and Lemma 43 to show the regular estimate corresponding to equations (3.42) to (3.45).

Theorem 44. If $U(x, y)$ satisfies equations (3.42) to (3.45), there exists $k_{r e f}<k$ such that we have the following regularity estimate

$$
\mathrm{k} D^{p} U \mathrm{k}_{L^{2}(\Omega)} \leq \sqrt{C\left(\epsilon, k_{r e f}, \epsilon_{+}\right)} k^{|p|-1} \mathrm{k} f \mathrm{k}_{L^{2}(\Omega)}
$$

where $\epsilon$ and $\epsilon_{+}$are given respectively in Lemma 39 and Lemma 43.

Proof. We have from Lemma 43 and Lemma 39 that

$$
\begin{aligned}
\sum_{n \in \mathbb{Z}} \mathrm{k} D^{p}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathrm{k}^{2}= & \sum_{n \in N_{-}} \mathrm{k} D^{p}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathrm{k}^{2} \\
& +\sum_{n \in N_{+}} \mathrm{k} D^{p}\left(U^{\left(n_{\alpha}\right)}(y) e^{i n_{\alpha} x}\right) \mathrm{k}_{L^{2}(\Omega)}^{2}, \\
\leq & C(\epsilon) k^{2(|p|-1)} \sum_{n \in N_{-}} \mathrm{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}(\Omega)}^{2} \\
& +\left(\frac{1}{\epsilon^{+}}+1\right) \sum_{n \in N_{+}} \mathrm{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}(\Omega)}^{2}, \\
\leq & C\left(\epsilon, k_{r e f}, \epsilon_{+}\right) k^{2(|p|-1)} \sum_{n \in \mathbb{Z}} \mathrm{k} f^{\left(n_{\alpha}\right)}(y) \mathrm{k}_{L^{2}(\Omega)}^{2}
\end{aligned}
$$

with

$$
C(\epsilon)+\frac{1}{}
$$

with a TE wave (Case 1A) and the transmitting dielectric interacting with a TE wave (Case 2A). To keep this chapter to a manageable size Cases 1B and 2B are relegated to Appendix C. We investigated the robustness of our regularity result for the special case where the multiple scattering problem is reduced into a one dimensional single scattering problem. We have given the solution in terms of the Green's functions so that we can examine the regularity of our solution in terms of the forcing term (incident wave). We derived a result for multiple scattering and this showed an explicit dependence on the wavenumber $k$ and the forcing term $f$. This regularity result will give us a hold on the convergence and the stability of the solution when we later solve numerically the scattering problems using finite element methods. In fact, if we let $h$ denote the maximum mesh size of our elements, and $p$ be the highest order of the finite element basis, since we know explicitly the dependence of the regularity result on the wavenumber $k$, then the a priori error estimate will present a power factor of $k h / p$. This shows that when we have a high wavenumber, we need a finer mesh or a higher order for the polynomial basis. Hence, when we solve numerically, we can use this information to balance the computational time and the accuracy of our approximated solution.

We also examined a special case that reduced the multiple scattering problem into a one dimensional single scattering problem. In fact, when $k^{2}>\left(n_{\alpha}\right)^{2}$, we have shown that each $U^{\left(n_{\alpha}\right)}(y)$ satisfies a one dimensional Helmholtz problem. Hence, we can use the results given in [85] which studied recently the Helmholtz problem for single scattering for the TE, transmitting dielectric gratings case. In the other case, where $k^{2}<\left(n_{\alpha}\right)^{2}$, we cannot use the results from [85] since $U^{\left(n_{\alpha}\right)}(y)$ satisfies transformed Poisson's equation, so we need to use the properties of the Fourier transform to study the regularity of $U^{\left(n_{\alpha}\right)}(y)$. We have shown that the regularity result for this special case, which is given by Theorem 44, agrees with the regularity result in Theorem 33. So, the regularity result given for the general case are the best approximation and the only results to date for a periodic grating with respect to one direction, which show an explicit dependence on the wavenumber $k$, for perfectly conducting gratings. Regularity results in $H^{1}\left(\mathbb{R}^{2}\right)$ for general unbounded penetrable rough layers have been studied in [74]. Since the authors in [74] did not focus on the special case where the grating is periodic, their regularity constant depends on the wavenumber $k(x, y)$ (if $p=1$ in Theorem 33 then we just have $k_{0}$ dependency).

In the following chapters, we will study the homogeneous problem corresponding to equations (2.95) and (2.104). The regularity of the solution of the inhomogeneous Helmholtz problem will be used to perform an a priori error estimate of the solution corresponding to the homogenous problems.

## Chapter 4

## A priori error estimates using the $\alpha, 0$-quasi periodic transformation

In this chapter, we introduce the approximation method that we will use to solve the diffraction problems described in Chapters 2 and 3. For clarity, we mainly focus on Case 1A and we relegate Cases 1B, 2A and 2B to Appendix D. Since $U$ is $\alpha$-quasi periodic then we can use the $\alpha, 0$-quasi periodic transformation by defining a function denoted by $U_{\alpha, 0}$ which is periodic with respect to $x$ where $U=e^{i \alpha x} U_{\alpha, 0}$.

Note that it is easier to study analytically the scattering problem using $U$ since it lends itself more readily to a variational formulation. It will transpire however that the numerical implementation of the finite element method is computationally less expensive and less complicated if we base it on $U_{\alpha, 0}$. Hence, instead of approximating numerically the scattering problem and looking directly for $U$, we look for the wave equations satisfied by $U_{\alpha, 0}$. We start by studying the continuity properties corresponding to $U_{\alpha, 0}$, and then examine the variational formulation. When we adopt such an approach, we need to show that the problem corresponding to the variational formulation is well-posed; that is, the solution exists, is unique and depends continuously on the data. We will show that the problem is $H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$-coercive. We then use the finite element method to discretise the problem, and provide a rigorous study of the a priori error estimation. In order to do so, we first derive a regularity result for the scattering problem in periodic space $H_{\#}^{l}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ for $l \geq 1$. We will show that solving the variational formulation for $U$ is equivalent to solving the variational formulation for $U_{\alpha, 0}$. Since we have already derived regularity results in the quasi-periodic case in Chapter 3 and since the variational formulation for $U$ is much simpler than that of $U_{\alpha, 0}$, we investigate the a priori error estimation in $H_{\alpha \neq}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$. This will then allow us to prove a new error estimate that differs from those in the literature since it provides an explicit dependence on the wavenumber. This provides a clearer insight into the convergence of the solution and will help in our numerical implementation to balance the accuracy against the computational cost.

### 4.1 Restatement of the boundary value problems for the periodic solution

We are going to use the $\alpha$-quasi periodicity of the solution $U$. Hence, from Lemma A-16 there exists $U_{\alpha, 0}$ which is periodic, of period $d$ with respect to $x$, such that

$$
\begin{equation*}
U(x, y)=e^{i \alpha x} U_{\alpha, 0}(x, y) \tag{4.1}
\end{equation*}
$$

Therefore the propagating equation is changed, and we have the following lemma.
Lemma 45. Let $U_{\alpha, 0} \in C^{2}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ satisfy equation (4.1), then $U_{\alpha, 0}$ is the solution of the following problem in the truncated domain $\Omega \backslash$ int $\Omega_{3}$ (see Figure 2.3)

$$
\begin{equation*}
\Delta U_{\alpha, 0}+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0}=0 \tag{4.2}
\end{equation*}
$$

with the DtN map interface conditions at the boundaries of the truncated region given by

$$
\begin{array}{ll}
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}, & \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=0, & \text { on } \Gamma_{-} . \tag{4.4}
\end{array}
$$

The Dirichlet boundary condition at the surface of the diffraction grating is

$$
\begin{equation*}
U_{\alpha, 0}(x, y)=0, \quad \text { on } \partial \Omega_{3} \tag{4.5}
\end{equation*}
$$

and the periodic condition

$$
\begin{equation*}
U_{\alpha, 0}(d, y)=U_{\alpha, 0}(0, y), \quad y \in[-B, B], \tag{4.6}
\end{equation*}
$$

holds where $U(x, y)$ is the solution of the original Helmholtz problem given by equation (2.95) and

$$
\begin{equation*}
T_{ \pm}^{\alpha, 0} U_{\alpha, 0}(x)=\sum_{n \in \mathbb{Z}} i \beta_{1}^{n} U_{\alpha, 0}^{(n)}( \pm B) e^{i \frac{2 \pi n}{d} x} \tag{4.7}
\end{equation*}
$$

using Definition 4.
Proof. We have

$$
\begin{align*}
\nabla U & =\nabla\left(e^{i \alpha x} U_{\alpha, 0}\right), \\
& =\nabla\left(e^{i \alpha x}\right) U_{\alpha, 0}+e^{i \alpha x} \nabla U_{\alpha, 0}, \\
& =\left[\begin{array}{c}
i \alpha e^{i \alpha x} \\
0
\end{array}\right] U_{\alpha, 0}+e^{i \alpha x} \nabla U_{\alpha, 0} . \tag{4.8}
\end{align*}
$$

We also have

$$
\begin{aligned}
\nabla . \nabla U & =-\alpha^{2} e^{i \alpha x} U_{\alpha, 0}+2 i \alpha e^{i \alpha x} \nabla U_{\alpha, 0}+e^{i \alpha x} \Delta U_{\alpha, 0}, \\
\Delta U & =e^{i \alpha x}\left(-\alpha^{2} U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0}+\Delta U_{\alpha, 0}\right),
\end{aligned}
$$

and since $e^{i \alpha x} 60$, then equation (2.95) implies

$$
\Delta U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0}+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}=0
$$

which is equation (4.2). For the boundary conditions, since $U$ is $\alpha$-quasi periodic with respect to $x$ then $U_{\alpha, 0}$ is periodic by using Lemma A-16 and similarly if $U_{\alpha, 0}$ is periodic with respect to $x$ then $U$ is $\alpha$-quasi periodic. On the top boundary $\Gamma_{+}=\{(x, y) \in \Omega: y=B\}$, we have from equation (2.96),

$$
\left.\partial_{n}(U)\right|_{\Gamma_{+}}=\partial_{n}\left(e^{i \alpha x} U_{\alpha, 0}\right)_{\Gamma_{+}}=T_{+}\left(e^{i \alpha x} U_{\alpha, 0}\right)-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} e^{i \alpha x} .
$$

By using Lemma 9 this becomes

$$
\begin{equation*}
\left.e^{i \alpha x} \partial_{n} U_{\alpha, 0}\right|_{\Gamma_{+}}=e^{i \alpha x}\left(T_{+}^{\alpha} U_{\alpha, 0}-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}\right) . \tag{4.9}
\end{equation*}
$$

Since $e^{i \alpha x} 60$ then

$$
\left.\partial_{n} U_{\alpha, 0}\right|_{\Gamma_{+}}=T_{+}^{\alpha, 0} U_{\alpha, 0}-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} .
$$

On the bottom boundary $\Gamma_{-}=\{(x, y) \in \Omega: y=-B\}$, we have from equation (2.97)

$$
\left.\partial_{n}(U)\right|_{\Gamma_{-}}=\partial_{n}\left(e^{i \alpha x} U_{\alpha, 0}\right)_{\Gamma_{-}}=T_{-}\left(e^{i \alpha x} U_{\alpha, 0}\right) .
$$

By using Lemma 9, we can write

$$
\begin{equation*}
\left.e^{i \alpha x} \partial_{n} U_{\alpha, 0}\right|_{\Gamma_{-}}=e^{i \alpha x} T_{-}^{\alpha, 0} U_{\alpha, 0} . \tag{4.10}
\end{equation*}
$$

Since $e^{i \alpha x} 60$ then

$$
\left.\partial_{n} U_{\alpha, 0}\right|_{\Gamma_{-}}=T_{-}^{\alpha, 0} U_{\alpha, 0} .
$$

Since $e^{i \alpha x} G 0$, we finish the proof by noting that $U=0$ on $\partial \Omega_{3}$ if and only if $U_{\alpha, 0}=0$ on $\partial \Omega_{3}$.

### 4.1.1 Variational formulation

To obtain a numerical method for computing an approximation to $U_{\alpha, 0}$ we start by deriving a variational statement of our scattering problem.

Lemma 46. The variational form of the boundary value problem given by equation (4.2) to equation (4.6) is given by the following statement. Find $U_{\alpha, 0} \in$ $H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, for all $v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and $\left.v\right|_{\partial \Omega_{3}}=0$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v\right)=(f, v)_{\Gamma_{+}}, \tag{4.11}
\end{equation*}
$$

and

$$
U_{\alpha, 0}=0, \quad \text { on } \partial \Omega_{3}
$$

where

$$
\begin{align*}
a(w, v)= & \int_{\Omega \backslash \operatorname{int} \Omega_{3}} \nabla w \cdot \nabla \bar{v}-\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) w \bar{v}-2 i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x} w\right) \bar{v} \\
& -\int_{\Gamma_{+}} T_{+}^{\alpha, 0} w \bar{v}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} w \bar{v} \tag{4.12}
\end{align*}
$$

and

$$
\begin{equation*}
(f, v)_{\Gamma_{+}}=-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \bar{v} d x \tag{4.13}
\end{equation*}
$$

for $w \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and $\left.w\right|_{\partial \Omega_{3}}=0$.
Proof. Multiplying both sides of equation (4.2) by $\bar{v}$ and integrating gives

$$
\int_{\Omega \backslash \text { int } \Omega_{3}} \Delta U_{\alpha, 0} \bar{v}+\int_{\Omega \backslash \text { int } \Omega_{3}}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0} \bar{v}+2 i \alpha \int_{\Omega \backslash \text { int } \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \bar{v}=0,
$$

for all $v \in H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$. We integrate by parts to get

$$
\begin{aligned}
& \int_{\Omega \backslash \mathrm{int} \Omega_{3}} \nabla U_{\alpha, 0} \cdot \nabla \bar{v}-\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0} \bar{v}-2 i \alpha \int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \bar{v} \\
&-\int_{\partial\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}=0,
\end{aligned}
$$

for all $v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and $\left.v\right|_{\partial \Omega_{3}}=0$. Then,

$$
\begin{align*}
& \int_{\Omega \backslash \operatorname{int} \Omega_{3}} \nabla U_{\alpha, 0} . \nabla \bar{v}-\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0} \bar{v}-2 i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \bar{v} \\
&-\int_{\partial \Omega_{3}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}-\int_{\Gamma_{L} \cup \Gamma_{R} \cup \Gamma_{ \pm}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}=0 . \tag{4.14}
\end{align*}
$$

with $\Gamma_{L}, \Gamma_{R}$ as defined in equations (3.47) and (3.48). Let $n_{\Gamma_{L}}\left(n_{\Gamma_{R}}\right)$ denote the exterior unit normal vector on $\Gamma_{L}$ (the exterior unit normal vector on $\Gamma_{R}$ ), and note that $n_{\Gamma_{L}}=-n_{\Gamma_{R}}$. Since $v$ and $U_{\alpha, 0}$ are periodic then

$$
\begin{aligned}
\int_{\Gamma_{L} \cup \Gamma_{R}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v} & =\int_{\Gamma_{L}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}+\int_{\Gamma_{R}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v} \\
& =\int_{\Gamma_{L}}\left(\frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}-\frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}\right) \\
& =0 .
\end{aligned}
$$

We also have $\int_{\partial \Omega_{3}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}=0$ from the Dirichlet boundary condition on $v$ and so equation (4.14) becomes

$$
\begin{aligned}
\int_{\Omega \backslash i n t} \Omega_{3} & \nabla U_{\alpha, 0} \cdot \nabla \bar{v}-\int_{\Omega \backslash i n t}\left(\Omega_{3}\right. \\
& \left.-k_{\Gamma_{+}} T_{+}^{\alpha, 0} U_{\alpha, 0}\right) U_{\alpha, 0} \bar{v}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} U_{\alpha, 0} \bar{v}=-\int_{\Gamma_{+}} 2 i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x}^{0} e^{-i \beta_{1}^{0} B} \bar{v}\right.
\end{aligned}
$$

using equations (4.3) and (4.4) and this finishes the proof.

### 4.1.2 Equivalence of the variational forms for the periodic and $\alpha$-quasi periodic problems

We want to show that the periodic problem is well posed. We also want to establish an upper bound on the error that arises when we solve the scattering problem numerically. For these reasons, we need to study the equivalence of the variational form for the periodic and $\alpha$-quasi periodic problem and this can be described as follows.

Similar to Lemma 46, for the periodic function $U_{\alpha, 0}$ and $v_{\alpha} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, let

$$
\begin{align*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)= & \left(\nabla U_{\alpha, 0}, \nabla v_{\alpha}\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}-2 i \alpha\left(\partial_{x} U_{\alpha, 0}, v_{\alpha}\right)_{\Omega \backslash \mathrm{int} \Omega_{3}} \\
& -\left(\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}, v_{\alpha}\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}-\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, v_{\alpha}\right)_{\Gamma_{ \pm}},  \tag{4.15}\\
\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}}= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{\alpha}} .
\end{align*}
$$

We have the Dirichlet boundary condition that

$$
\begin{equation*}
U_{\alpha, 0}=0, \text { on } \partial \Omega_{3}, \tag{4.16}
\end{equation*}
$$

and $T_{ \pm}^{\alpha, 0}$ are given by Definition 4. The variational problem is to find $U_{\alpha, 0} \in$ $H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ for all $v_{\alpha}$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)=\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}} . \tag{4.17}
\end{equation*}
$$

Similarly, for the $\alpha$-quasi periodic function $U$ let

$$
\begin{align*}
a(U, v) & =(\nabla U, \nabla v)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} U, v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm} U, v\right)_{\Gamma_{ \pm}}  \tag{4.18}\\
(f, v)_{\Gamma_{+}} & =-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{v}
\end{align*}
$$

where $T_{ \pm}$are given by Definition 3 and $U$ satisfies the Dirichlet boundary condition

$$
\begin{equation*}
U(x, y)=0, \text { on } \partial \Omega_{3} . \tag{4.19}
\end{equation*}
$$

The variational problem is to find $U \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ such that

$$
\begin{equation*}
a(U, v)=(f, v)_{\Gamma_{+}} . \tag{4.20}
\end{equation*}
$$

We have the following result.
Lemma 47. Finding $U_{\alpha, 0} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ for all $v_{\alpha} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ such that $a\left(U_{\alpha, 0}, v_{\alpha}\right)=\left(f_{\alpha}, v_{\alpha}\right)$ as given in equation (4.17) is equivalent to finding $U \in$ $H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ such that $a(U, v)=(f, v)_{\Gamma_{+}}$using equation (4.20).

Proof. Let $v=e^{i \alpha x} v_{\alpha}, w=e^{i \alpha x} w_{\alpha}$, such that $w_{\alpha} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, then we have

$$
\begin{aligned}
\left(\nabla v_{\alpha}, \nabla w_{\alpha}\right)_{\Omega \backslash \text { int } \Omega_{3}} & =\left(\nabla\left(e^{-i \alpha x} v\right), \nabla\left(e^{-i \alpha x} w\right)\right)_{\Omega \backslash \text { int } \Omega_{3}} \\
& =\int_{\Omega \backslash \text { int } \Omega_{3}}\left[\begin{array}{l}
-i \alpha e^{-i \alpha x} v+e^{-i \alpha x} \partial_{x} v \\
e^{-i \alpha x} \partial_{y} v
\end{array}\right] \cdot\left[\begin{array}{l}
i \alpha e^{i \alpha x} \bar{w}+e^{i \alpha x} \partial_{x} \bar{w} \\
e^{i \alpha x} \partial_{y} \bar{w}
\end{array}\right]
\end{aligned}
$$

and we have

$$
\begin{equation*}
\left(\nabla v_{\alpha}, \nabla w_{\alpha}\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}=\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\alpha^{2} v \bar{w}-i \alpha v \partial_{x} \bar{w}+i \alpha \bar{w} \partial_{x} v+\nabla v \cdot \nabla \bar{w}\right) . \tag{4.21}
\end{equation*}
$$

Since $v \bar{w}$ is periodic with respect to $x$, then $\int_{0}^{d} \partial_{x}(v \bar{w})=0$ and imposing Dirichlet boundary condition on $\partial \Omega_{3}$ we also get $\int_{\Omega \backslash \text { int } \Omega_{3}} \partial_{x}(v \bar{w})=0$. So,

$$
\int_{\Omega \backslash \text { int } \Omega_{3}}\left(\partial_{x} v\right) \bar{w}=-\int_{\Omega \backslash \text { int } \Omega_{3}} v\left(\partial_{x} \bar{w}\right)
$$

and hence equation (4.21) becomes

$$
\begin{equation*}
\left(\nabla v_{\alpha}, \nabla w_{\alpha}\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}=\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\alpha^{2} v \bar{w}+2 i \alpha \bar{w} \partial_{x} v+\nabla v \cdot \nabla \bar{w}\right) . \tag{4.22}
\end{equation*}
$$

Examining the next term in the sesquilinear form given by equation (4.15) we see

$$
\begin{align*}
-i \alpha\left(\partial_{x} v_{\alpha}, w_{\alpha}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} & =-i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(-i \alpha e^{-i \alpha x} v+e^{-i \alpha x} \partial_{x} v\right) e^{i \alpha x} \bar{w} \\
& =\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(-\alpha^{2} v \bar{w}-i \alpha \bar{w} \partial_{x} v\right) . \tag{4.23}
\end{align*}
$$

Next we have the following

$$
\begin{align*}
\left(\left(k^{2}-\alpha^{2}\right) v_{\alpha}, w_{\alpha}\right)_{\Omega \backslash \text { int } \Omega_{3}} & =\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) e^{-i \alpha x} v e^{i \alpha x} \bar{w} \\
& =\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) v \bar{w} . \tag{4.24}
\end{align*}
$$

Finally, the last term of the sesquilinear form in equation (4.15) can be written as

$$
\begin{equation*}
\left(T_{ \pm}^{\alpha, 0} v_{\alpha}, w_{\alpha}\right)_{\Gamma_{ \pm}}=\left(T_{ \pm}^{\alpha, 0} e^{-i \alpha x} v, e^{i \alpha x} \bar{w}\right)_{\Gamma_{ \pm}}=\left(T_{ \pm} v, \bar{w}\right)_{\Gamma_{ \pm}} . \tag{4.25}
\end{equation*}
$$

Substituting equations (4.22),(4.23),(4.24) and (4.25) into equation (4.15) we get

$$
\begin{aligned}
a\left(v, w_{\alpha}\right) & =\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\nabla v \cdot \nabla \bar{w}-k^{2} v \bar{w}\right)-\int_{\Gamma_{ \pm}} T_{ \pm} v \bar{w} \\
& =a(v, w) .
\end{aligned}
$$

We finish the proof of Lemma 47 by noting that

$$
\begin{align*}
\left(f_{\alpha}, w_{\alpha}\right)_{\Gamma_{+}} & =-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{w_{\alpha}} \\
& =-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} e^{i \alpha x} \bar{w}, \\
& =(f, w)_{\Gamma_{+}} . \tag{4.26}
\end{align*}
$$

### 4.1.3 Well posedness of the variational problem

Before solving the variational formulation numerically, we first show that our problem is well posed which means that a solution exists, it is unique and the solution depends continuously on the data (also called regularity of the solution) [54, 37]. We differ from other previous published work [9] since we show rigorously this well posedness of the boundary value problem. The variational form associated with $U$ is easier to study analytically so we are going to show that the $\alpha$-quasi periodic problem is well posed and from that result we then show that the variational formulation corresponding to $U_{\alpha, 0}$ is also well posed.

### 4.1.3.1 Existence and uniqueness of the solution

Lemma 48. For all $v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, the solution $U_{\alpha, 0} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ which satisfies equation (4.11) exists and is unique.

Proof. We note by using the Cauchy-Schwarz inequality [22, p. 50] that

$$
\begin{align*}
\left|(\nabla U, \nabla v)_{\Omega \backslash \operatorname{int} \Omega_{3}}\right| & =\left|\int_{\Omega \backslash \operatorname{int} \Omega_{3}} \nabla U . \nabla \bar{v} d x d y\right|  \tag{4.27}\\
& \leq \int_{\Omega \backslash \operatorname{int} \Omega_{3}}|\nabla U . \nabla \bar{v}| d x d y  \tag{4.28}\\
& \leq \mathrm{k} \nabla U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} \nabla v \mathrm{k}_{L_{\alpha \neq}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \tag{4.29}
\end{align*}
$$

and also that

$$
\begin{align*}
\left|\left(k^{2} U, v\right)_{\Omega \backslash \text { int } \Omega_{3}}\right| & =\left|\int_{\Omega \backslash \operatorname{int} \Omega_{3}} k^{2} U \bar{v} d x d y\right|  \tag{4.30}\\
& \leq \mathrm{k} k^{2} \mathbf{k}_{\infty} \int_{\Omega \backslash \operatorname{int} \Omega_{3}}|U \bar{v}| d x d y  \tag{4.31}\\
& \leq \mathrm{k} k^{2} \mathbf{k}_{\infty} \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \tag{4.32}
\end{align*}
$$

Definition 3 and Lemma 2 with $f=v$ give

$$
\begin{aligned}
\left|\int_{\Gamma_{ \pm}} T_{ \pm} U \bar{v} d x\right|= & d\left|\sum_{n \in \mathbb{Z}} i \beta_{j}^{n} U^{\left(n_{\alpha}\right)}( \pm B) v^{\left(n_{\alpha}\right)}( \pm B)\right| \\
= & d\left|\sum_{n \in \mathbb{Z}} i \beta_{j}^{n}\left(1+n_{\alpha}^{2}\right)^{-1 / 4}\left(1+n_{\alpha}^{2}\right)^{1 / 4} U^{\left(n_{\alpha}\right)}( \pm B) v^{\left(n_{\alpha}\right)}( \pm B)\right| \\
\leq & d\left(\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|^{2}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}\right)^{1 / 2} \\
& \left(\sum_{n \in \mathbb{Z}}\left(1+n_{\alpha}^{2}\right)^{1 / 2}\left|v^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}\right)^{1 / 2}
\end{aligned}
$$

We use equations (2.68) and (2.69)

$$
\begin{aligned}
\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|^{2}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \leq & c_{2} \sum_{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|}\left|k_{j}^{2}\right|\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& +c_{1} \sum_{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|} n_{\alpha}^{2}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
\leq & c_{2}\left|k_{j}^{2}\right| \sum_{n \in \mathbb{Z}}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& +c_{1} \sum_{n \in \mathbb{Z}}\left(1+n_{\alpha}^{2}\right)^{1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
\leq & c_{2}\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{H_{\alpha}^{-\frac{1}{\#}}\left(\Gamma_{ \pm}\right)}^{2}+c_{1} \mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{2}}^{\frac{1}{2}\left(\Gamma_{ \pm}\right)},
\end{aligned}
$$

from Definition A-17. Hence

$$
\begin{aligned}
\left|\int_{\Gamma_{ \pm}} T_{ \pm} U \bar{v} d x\right|^{2} \leq & C^{2} d^{2}\left(\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{H_{\alpha \# \#}^{-\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2}\right. \\
& \left.\left.+\mathrm{k} U \mathrm{k}_{H_{\alpha \# \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2}\right) \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{1}\left(\Gamma_{ \pm}\right)}^{2}\right)
\end{aligned}
$$

where $C^{2}=\sup \left(c_{1}, c_{2}\right)$ and we have

$$
\begin{gather*}
\left|\int_{\Gamma_{ \pm}} T_{ \pm} U \bar{v} d x\right|^{2} \leq C^{2} d^{2}\left(\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+\mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}\right) \\
\mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}, \tag{4.33}
\end{gather*}
$$

from Theorem A-13. Hence, we have from equation (4.18)

$$
\begin{aligned}
|a(U, v)| \leq & |U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash i n t \Omega_{3}\right)}+k^{2} \mathbf{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}{\mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}} \quad+C d\left(\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{L_{\alpha \# \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+\mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2}\right) \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2},
\end{aligned}
$$

and so

$$
\begin{equation*}
|a(U, v)| \leq C_{0} \sup \left(1, k^{2}\right){\mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{1}(\Omega \backslash i n t} \Omega_{3} \mathrm{k} \mathrm{k}_{\mathrm{H}_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} .} \tag{4.34}
\end{equation*}
$$

Hence, $a(U, v)$ is continuous using Definition A-6. Taking the real part of $a(U, U)$ and from equation (2.72), we get

$$
\begin{equation*}
\mathfrak{R}(a(U, U)) \geq|U|_{H_{\alpha \neq}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-\Re\left(k^{2}\right) \boldsymbol{k} U \mathbf{k}_{L_{\alpha \neq 1}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} . \tag{4.35}
\end{equation*}
$$

Hence,

$$
\left.\mathfrak{R}(a(U, U))+\Re\left(k^{2}\right) \mathbf{k} U \mathbf{k}_{L_{\alpha \#}^{2}(\Omega \backslash i n t}^{2} \Omega_{3}\right) \geq|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}
$$

and

$$
\begin{align*}
& \mid a(U, U)+\mathfrak{R}\left(k^{2}\right){\mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \mid}^{\geq} \mathfrak{R}(a(U, U))+\mathfrak{R}\left(k^{2}\right) \mathbf{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}, \\
& \geq|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}^{2}, \\
& \geq M_{1} \mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{1}}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3} .\right. \tag{4.36}
\end{align*}
$$

Then, $a(U, U)$ is $H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$-coercive from Definition B-4. We can then use Lemma B- 5 to show the existence of a solution from its uniqueness. Let us suppose that we have two solutions $U_{1}$ and $U_{2}$ and let us denote $w=U_{1}-U_{2}$. We have from equation (4.18) that

$$
a(w, w)=|w|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2}-\int_{\Omega \backslash \mathrm{int} \Omega_{3}} k^{2}|w|^{2}-\int_{\Gamma_{ \pm}}\left(T_{ \pm} w\right) \bar{w}=0
$$

and so

$$
\begin{aligned}
& |w|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-\int_{\Omega \backslash \operatorname{int} \Omega_{3}} \mathfrak{R}\left(k^{2}\right)|w|^{2}-\mathfrak{R}\left(\int_{\Gamma_{ \pm}}\left(T_{ \pm} w\right) \bar{w}\right) \\
& -i\left(\mathfrak{J}\left(\int_{\Gamma_{ \pm}}\left(T_{ \pm} w\right) \bar{w}\right)+\int_{\Omega \backslash \operatorname{int} \Omega_{3}} \mathfrak{J}\left(k^{2}\right)|w|^{2}\right)=0 .
\end{aligned}
$$

Hence,

$$
|w|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2}-\int_{\Omega \backslash \text { int } \Omega_{3}} \mathfrak{R}\left(k^{2}\right)|w|^{2}-\mathfrak{R}\left(\int_{\Gamma_{ \pm}}\left(T_{ \pm} w\right) \bar{w}\right)=0
$$

and

$$
\left(\mathfrak{I}\left(\int_{\Gamma_{ \pm}}\left(T_{ \pm} w\right) \bar{w}\right)+\int_{\Omega \backslash \mathrm{int} \Omega_{3}} \mathfrak{J}\left(k^{2}\right)|w|^{2}\right)=0 .
$$

We note that $\left(\mathfrak{I}\left(\int_{\Gamma_{ \pm}}\left(T_{ \pm} U\right) \bar{U}\right)+\int_{\Omega \backslash \text { int } \Omega_{3}} \mathfrak{J}\left(k^{2}\right)|U|^{2}\right) \geq 0$ from equation (2.72) and $\mathfrak{J}\left(k^{2}\right)=2 \mathfrak{J}(k) \mathfrak{R}(k) \geq 0$. Therefore $w=0$ and hence $U_{1}=U_{2}$. Since $U_{1}=e^{i \alpha x} U_{\alpha, 0_{1}}$ and $U_{2}=e^{i \alpha x} U_{\alpha, 0_{2}}$ then $U_{\alpha, 0_{1}}=U_{\alpha, 0_{2}}$ which finishes the proof.

In order for a variational formulation to depend continuously on the data, it is necessary to show that the variational formulation satisfies a regularity estimate which will be studied in the following section.

### 4.1.3.2 Regularity estimate of the exact solution

This problem was studied in [9] for the transmitting dielectric grating for the TE case (Case 2A ). They derived, using Cauchy's inequality, an a priori error estimate. However, this estimate did not show an explicit dependence on the wavenumber $k$ and the degree of the polynomial basis $p$. Furthermore, the regularity of the solution (which is required to show that the problem is well posed and is a prerequisite for deriving an a priori error estimate) was simply assumed and not proven. In our study, we will derive an explicit dependency on $k$ in the proof of the regularity of the solution. In the a priori error estimate, we will derive an explicit dependency on the wavenumber $k$ and the degree of the polynomial basis $p$. To simplify the algebra, let us define the following norm.

Definition 49. Let $F \subset \mathbb{R}^{2}$ and $v \in H^{1}(\digamma)$ (see Definition A-10) then we define [85]

$$
\begin{equation*}
\mathrm{k} v \mathrm{k}_{\mathcal{H}}^{2}=|v|_{H^{1}(\digamma)}^{2}+\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} v \mathrm{k}_{L^{2}(\digamma)}^{2} . \tag{4.37}
\end{equation*}
$$

Note that $\mathbf{k} v \mathbf{k}_{\mathcal{H}}$ is equivalent to $\mathbf{k} v \mathbf{k}_{H^{1}(\digamma)}$ since

$$
\inf \left(1, \frac{1}{\mathrm{k} k \mathrm{k}_{\infty}}\right) \mathbf{k} v \mathbf{k}_{\mathcal{H}} \leq \mathrm{k} v \mathrm{k}_{H^{1}(\digamma)} \leq \sup \left(1, \frac{1}{\mathrm{k} k \mathrm{k}_{\infty}}\right) \mathbf{k} v \mathbf{k}_{\mathcal{H}}
$$

 $H_{\alpha \#}^{l}(\digamma) \subset H^{l}(\digamma)$ and $H_{\#}^{l}(\digamma) \subset H^{l}(\digamma)$ for any $l \geq 0$, then for any $v \in H_{\alpha \#}^{1}(\digamma)$ $\left(v \in H_{\#}^{1}(\digamma)\right) \mathbf{k}^{v} \mathbf{k}_{\mathcal{H}}$ is well defined.

Also we have the following theorem.
Theorem 50. Let $v \in H_{\alpha \#}^{l}(\digamma)$ and let $v_{\alpha, 0} \in H_{\#}^{l}(\digamma)$ such that

$$
\begin{equation*}
v=e^{i \alpha x} v_{\alpha, 0} . \tag{4.38}
\end{equation*}
$$

Let $V \subset H_{\alpha \#}^{l}(\digamma)$ be a finite element subspace of order $p$ with $l \geq 1$ as described in Section B. 3 and let us denote with h the maximum mesh size after partitioning $\digamma$. If we make the following standard assumption on the subspace $V$ [35]

$$
\begin{align*}
& \inf _{\psi \in V}\left\{\mathrm{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}+\frac{h}{p} \mathrm{k} \nabla v-\nabla \psi\right. \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}+\left(\frac{h}{p}\right)^{\frac{1}{2}} \mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \neq \#}^{2}\left(\Gamma_{ \pm}\right)} \\
&\left.+\frac{h}{p} \mathrm{k} v-\psi \mathbf{k}_{H_{\alpha \#}^{\frac{1}{\alpha}}\left(\Gamma_{ \pm}\right)}\right\} \leq C\left(\frac{h}{p}\right)^{l} \mathrm{~K}^{l} \mathbf{k}_{H_{\alpha \#}^{l}(\digamma)} \tag{4.39}
\end{align*}
$$

and $\mathrm{k} k \mathrm{k}_{\infty} \frac{h}{p}<1$, then, we have

$$
\begin{equation*}
\mathbf{k} v_{\alpha, 0} \mathbf{k}_{H_{\#}^{l}(\digamma)} \leq l 2^{l} \mathbf{k} v \mathrm{k}_{H_{\alpha \# \#}^{l}(\digamma)} \tag{4.40}
\end{equation*}
$$

for $l \in 0$ and

$$
\begin{equation*}
\mathbf{k} v_{\alpha, 0} \mathbf{k}_{L_{\#}^{2}(\digamma)}=\mathbf{k} v \mathrm{k}_{L_{\alpha \#}^{2}(\digamma)} . \tag{4.41}
\end{equation*}
$$

If $\mathrm{I} \subset \mathbb{R}$, where $y=y_{0}$ is constant, then we also have

$$
\begin{equation*}
\mathrm{k} v_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}(\mathrm{I})} \leq C_{\alpha \#}^{1 / 4} \mathrm{k}^{2} \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}(\mathrm{I})}, \tag{4.42}
\end{equation*}
$$

where $C_{\alpha \#}$ is given by equation (2.83).
Proof. First, we note that

$$
\begin{equation*}
\mathbf{k} v_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{L_{\#}^{2}(\digamma)}=\left|e^{-i \alpha x}\right| \mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}=\mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)} . \tag{4.43}
\end{equation*}
$$

Differentiating equation (4.38) gives

$$
\begin{aligned}
\frac{h}{p} \mathrm{k} \nabla v_{\alpha, 0}-\nabla \psi_{\alpha, 0} \mathbf{k}_{L_{\#}^{2}(\digamma)} & =\frac{h}{p} \mathrm{k}-i \alpha(v-\psi)+(\nabla v-\nabla \psi) \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}, \\
& \leq \frac{h}{p}|\alpha| \mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}+\frac{h}{p} \mathbf{k} \nabla v-\nabla \psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}
\end{aligned}
$$

by the triangle inequality. Since $|i \alpha|<\boldsymbol{k} k \mathbf{k}_{\infty}$ then

$$
\begin{align*}
\frac{h}{p} \mathrm{k} \nabla v_{\alpha, 0}-\nabla \psi_{\alpha, 0} \mathbf{k}_{L_{\#}^{2}(\digamma)} & \leq \frac{\mathbf{k} k \mathbf{k}_{\infty} h}{p} \mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}+\frac{h}{p} \mathbf{k} \nabla v-\nabla \psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}, \\
& \leq \mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)}+\frac{h}{p} \mathbf{k} \nabla v-\nabla \psi \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)} \tag{4.44}
\end{align*}
$$

since $\mathbf{k} k \mathbf{k}_{\infty} \frac{h}{p}<1$. We also note that

$$
\begin{equation*}
\mathbf{k} v_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{L^{2}\left(\Gamma_{ \pm}\right)}=\mathbf{k} v-\psi \mathbf{k}_{L^{2}\left(\Gamma_{ \pm}\right)} \tag{4.45}
\end{equation*}
$$

by Definition A-7. Using Definition A-14, we have

$$
\begin{aligned}
\mathbf{k} v_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{\#}}(\mathrm{I})}^{2} & =\sum_{n \in \mathbb{Z}}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{\frac{1}{2}}\left|v_{\alpha, 0}^{(n)}\left(y_{0}\right)-\psi_{\alpha, 0}^{(n)}\left(y_{0}\right)\right|^{2}, \\
& =\sum_{n \in \mathbb{Z}}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{\frac{1}{2}}\left|v^{\left(n_{\alpha}\right)}\left(y_{0}\right)-\psi^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right|^{2},
\end{aligned}
$$

using Lemma 13 and so from Lemma 12

$$
\mathbf{k} v_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{2}}(\mathrm{I})}^{2} \leq \sum_{n \in \mathbb{Z}} C_{\alpha \#}^{\frac{1}{2}}\left(1+\left(n_{\alpha}\right)^{2}\right)^{\frac{1}{2}}\left|v^{\left(n_{\alpha}\right)}\left(y_{0}\right)-\psi^{\left(n_{\alpha}\right)}\left(y_{0}\right)\right|^{2} .
$$

Hence, using Definition A-17

$$
\begin{equation*}
\mathbf{k} v_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{2}}(1)}^{2} \leq C_{\alpha \#}^{\frac{1}{2}} \mathbf{k} v-\psi \mathbf{k}_{H_{\alpha \#}^{2}(1)}^{2} . \tag{4.46}
\end{equation*}
$$

and we obtain equation (4.42). Now from equation (4.38) and using Definition A10

$$
\begin{aligned}
\mathrm{k} v_{\alpha, 0} \mathrm{k}_{H_{\#}^{l}(\digamma)} & =\mathrm{k} e^{-i \alpha x} v \mathrm{k}_{H_{\alpha \#( }^{l}(\digamma)} \\
& =\sum_{n=0}^{l} \mathrm{k} D^{n} e^{-i \alpha x} v \mathrm{k}_{L_{\alpha \#}^{2}(\digamma)} .
\end{aligned}
$$

Hence, we differentiate to get

$$
\begin{aligned}
& =\sum_{n=0}^{l} \mathrm{k} \sum_{r=0}^{n} C(n, r)(-i \alpha)^{r} D^{n-r} v \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)} \\
& \leq \sum_{n=0}^{l} \sum_{r=0}^{n} C(n, r) \mathbf{k} k \mathbf{k}_{\infty}^{r} \mathrm{k} D^{n-r} v \mathbf{k}_{L_{\alpha \#}^{2}(\digamma)} .
\end{aligned}
$$

Since $D^{n-r} v \in H_{\alpha \#}^{r}(\digamma)$, we can use the first term in equation (4.39) and we replace $v$ by $D^{n-r} v$ to get

$$
\begin{aligned}
& \leq \sum_{n=0}^{l} \sum_{r=0}^{n} C(n, r)\left(\mathrm{k} k \mathbf{k}_{\infty} \frac{h}{p}\right)^{r}{\mathrm{k} v \mathbf{k}_{H_{\alpha \#}^{n-r+r}(\digamma)},}_{\leq} \leq \sum_{n=0}^{l} \sum_{r=0}^{n} C(n, r) \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{n}(\digamma)}
\end{aligned}
$$

since we assume $\mathbf{k} k \mathbf{k}_{\infty} \frac{h}{p}<1$ and so

$$
\begin{aligned}
\mathrm{k}_{\alpha, 0} \mathrm{k}_{H_{\#}^{l}(\digamma)} & \leq \sum_{n=0}^{l} 2^{n} \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{n}(\digamma)} \\
& \leq 2^{l} \sum_{n=0}^{l} \mathrm{k}^{l} \mathrm{k}_{H_{\alpha \#}^{n}(\digamma)} \\
& \leq l 2^{l} \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{l}(\digamma)}
\end{aligned}
$$

Hence we obtain equation (4.40) which finishes the proof.
In the following theorem, we obtain a new result on the regularity estimate for the solution $U_{\alpha, 0}$.

Theorem 51. Let $f_{\alpha} \in H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and let $U_{\alpha, 0} \in H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be the solution of

$$
\begin{align*}
\Delta U_{\alpha, 0}+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0} & =f_{\alpha}, \text { in } \Omega \backslash \text { int } \Omega_{3},  \tag{4.47}\\
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, \text { on } \Gamma_{-}, \\
U_{\alpha, 0} & =0, \text { on } \partial \Omega_{3} . \tag{4.48}
\end{align*}
$$

Then there exists a constant $C_{\text {stab }}$ which is independent of the wavenumber $k$ such that

$$
\mathbf{k} U_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \leq C_{\text {stab }} \mathbf{k} f_{\alpha} \mathbf{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}
$$

Proof. $U \in H_{\alpha \#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ is the solution of equation (3.3) and from Definition 49

$$
\mathbf{k} U_{\alpha, 0} \mathbf{k}_{\mathcal{H}}^{2}=\left|U_{\alpha, 0}\right|_{H_{\#}^{1}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}+\mathbf{k} k \mathbf{k}_{\infty}^{2} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}
$$

Since $U_{\alpha, 0}$ satisfies equation (4.47), then equation (4.38) holds for $v_{\alpha}=U$ and $v=U_{\alpha, 0}$ and for $v_{\alpha}=f$ and $v=f_{\alpha}$ and so from equation (4.40)

$$
\mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}}^{2} \leq 2^{2} \mathrm{k} U \mathrm{k}_{H_{\alpha \neq}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash i n t \Omega_{3}\right)}^{2}
$$

By using the regularity estimate of the $U$ as given in Theorem 27 we have

$$
\begin{aligned}
\mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}}^{2} & \leq 2^{2} C_{r e g}^{2} \mathrm{k} f \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+\mathrm{k} k \mathrm{k}_{\infty}^{2-2} C_{r e g}^{2} \mathrm{k} f \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}, \\
& =\left(2^{2}+1\right) C_{r e g}^{2} \mathrm{k} f_{\alpha} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}^{2}
\end{aligned}
$$

using equation (4.45) and we denote $C_{\text {stab }}={ }^{\sqrt{\prime}} \overline{5} C_{\text {reg }}$ which finishes the proof.
We have shown that the problem is well posed by proving that its solution exists, is unique and satisfies a regularity result. Now, let us study the efficiency of the diffraction grating which presents the main interest for most applications.

### 4.1.4 Efficiency of the diffraction grating

For most applications, one is more interested in how efficient a given diffraction grating is in reflecting or transmitting electromagnetic waves, rather than just studying the magnetic field $U_{\alpha, 0}$ for the TM case, or the electric field for the TE case.

Definition 52. The diffraction efficiency is the physical quantity that characterizes how the incident field power is distributed between the different orders. It is given by the ratio between the energy flow of a particular order in a direction perpendicular to the grating surface and the corresponding flow of the incident wave through the same surface [79][p. 35].

Hence, for a chosen diffraction order $m$, the diffraction efficiency for the reflected order $m$ (transmitted order $m$ ) is given by $R_{m}$ (respectively $T_{m}$ ) and can be computed as follows

$$
\begin{align*}
R_{m} & =\frac{\beta_{1}^{m}}{\beta_{1}^{0}}\left|r_{1}^{m}\right|^{2}  \tag{4.49}\\
T_{m} & =\frac{\beta_{2}^{m}}{\beta_{1}^{0}}\left|t_{2}^{m}\right|^{2} \tag{4.50}
\end{align*}
$$

with $r_{1}^{m}$ and $t_{2}^{m}$ given by equations (2.58) and (2.59) such that $\left(\frac{2 \pi m}{d}+\alpha\right)^{2}<k_{j}^{2}$, for $j=1,2$. Note that we can compute the diffraction efficiency for all $m$, however only the propagating modes can carry energy away from the grating since they do not have purely imaginary wavevector component $\beta_{1}^{m}\left(\beta_{2}^{m}\right)$ in the vertical direction [79][p. 36]. Recall that in our numerical computation we are looking for $U_{\alpha, 0}$ where $U=U_{\alpha, 0} e^{i \alpha x}$ since implementing the periodicity condition for $U_{\alpha, 0}$ is easier and computationally more efficient than the $\alpha$-quasi periodicity condition. Hence, to study the efficiency of the diffraction grating we need to establish a formula which enables us to compute the efficiency of $U$ from knowledge of $U_{\alpha, 0}$. From
equation (2.55), we note that on $\Gamma_{ \pm}$, we have

$$
\begin{aligned}
U(x, y) & =\sum_{m \in \mathbb{Z}} r_{1}^{m} e^{i \beta_{1}^{m} y+i\left(\alpha+\frac{2 \pi n}{d}\right) x}+t_{2}^{m} e^{-i \beta_{2}^{m} y+i\left(\alpha+\frac{2 \pi n}{d}\right) x}, \\
& =e^{i \alpha x} \sum_{m \in \mathbb{Z}} r_{1}^{m} e^{i \beta_{1}^{m} y+i \frac{2 \pi n}{d} x}+t_{2}^{m} e^{-i \beta_{2}^{m} y+i \frac{2 \pi n}{d} x}, \\
& =e^{i \alpha x} U_{\alpha, 0}(x, y)
\end{aligned}
$$

and so we can compute the efficiency from the Rayleigh expansion of $U_{\alpha, 0}$ on $\Gamma_{ \pm}$using equations (2.74), (2.58) and (2.59). We can also use the variational formulation given by equation (4.12) to show that the energy is conserved. By calculating the energy we can assess the presence or otherwise of the numerical inaccuracies in our numerical method. Before doing so, we need the following results.

Lemma 53. We have

$$
\begin{equation*}
\mathfrak{R}\left(\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}}\right)=0 \tag{4.51}
\end{equation*}
$$

Proof. Let us define $\Omega_{x}$ where $\Omega_{x} \times[-B, B]=\Omega \backslash \operatorname{int} \Omega_{3}$. We have

$$
\int_{\Omega \backslash \text { int } \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=\int_{-B}^{B}\left(\int_{\Omega_{x}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x\right) d y
$$

We note by integrating by parts that

$$
\int_{\Omega_{x}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x=\left[U_{\alpha, 0} \overline{U_{\alpha, 0}}\right]_{\partial \Omega_{x}}-\int_{\Omega_{x}} U_{\alpha, 0} \overline{\partial_{x} U_{\alpha, 0}} d x
$$

and so

$$
\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=\int_{-B}^{B}\left(\left[U_{\alpha, 0} \overline{U_{\alpha, 0}}\right]_{\partial \Omega_{x}}-\int_{\Omega_{x}} U_{\alpha, 0} \overline{\partial_{x} U_{\alpha, 0}} d x\right) d y
$$

For fixed $y$, either $\partial \Omega_{x} \subset \partial \Omega_{3}$ or $\partial \Omega_{x}=0$ or $d$. In the first case $U_{\alpha, 0}=0$ on $\partial \Omega_{3}$ and in the second case, since $U_{\alpha, 0}$ is periodic, then

$$
I=\int_{-B}^{B}\left[U_{\alpha, 0} \overline{U_{\alpha, 0}}\right]_{\partial \Omega_{x}} d y=0 .
$$

Hence,

$$
\begin{equation*}
\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=-\int_{\Omega \backslash \operatorname{int} \Omega_{3}} U_{\alpha, 0} \overline{\partial_{x} U_{\alpha, 0}} d x d y . \tag{4.52}
\end{equation*}
$$

For any $c \in \mathbb{C}$, if $c=-\bar{c}$ then $\mathfrak{R}(c)=0$ and the result follows.

### 4.1.5 Conservation of the energy or energy balance

Let us denote respectively by $E_{r}, E_{t}$ and $E_{a b s}$, the refracted energy, the transmitted energy and the absorbed energy by the grating [79, p. 36].

Definition 54. We have

$$
\begin{aligned}
E_{r} & =\sum_{\left(\frac{2 \pi m}{d}+\alpha\right)^{2}<k_{1}^{2}} R_{m} \\
E_{t} & =\sum_{\left(\frac{2 \pi m}{d}+\alpha\right)^{2}<k_{2}^{2}} T_{m} \\
E_{a b s} & =\frac{1}{d \beta_{1}^{0}} \int_{\Omega} \mathfrak{J}\left(k^{2}\right)\left|U_{\alpha, 0}\right|^{2}
\end{aligned}
$$

such that $R_{m}$ and $T_{m}$ are as given by equations (4.49) and (4.50).
By using the relation between the Rayleigh coefficients (see equation (2.55)) and the Fourier coefficients of the scattering field on $\Gamma_{ \pm}$as given by the system of equations (2.57), we have the following result.

Lemma 55. The square of the absolute value of the Rayleigh coefficients can be obtained from the Fourier coefficients of the solution on the top boundary $y=B$ and the bottom boundary $y=-B$ as follows

$$
\begin{array}{lr}
\left|r_{1}^{n}\right|^{2}=\left|U_{\alpha, 0}^{(n)}(B)\right|^{2}, & n \in \mathbb{Z} \backslash\{0\} \\
\left|r_{1}^{0}\right|^{2}=\left|U_{\alpha, 0}{ }^{(0)}(B)\right|^{2}+1-2 \mathfrak{R}\left(\overline{U_{\alpha, 0}(0)}(B) e^{-i \beta_{1}^{0} B}\right) & \\
\left|t_{2}^{n}\right|^{2} \leq\left|U_{\alpha, 0}^{(n)}(-B)\right|^{2} e^{2 \Im\left(\beta_{2}^{n}\right) B}, & n \in \mathbb{Z}
\end{array}
$$

Proof. The proof for the first and the last lines are straightforward using the system of equations (2.57). So, we will just show the proof of the second line. From Lemma 13 we note that $U^{(0)}(B)=U_{\alpha, 0}{ }^{(0)}(B)$. We then use equation (2.58) to get

$$
\begin{aligned}
\left|r^{0}\right|^{2} & =\left|U_{\alpha, 0}{ }^{(0)}(B) e^{-i \beta_{1}^{0} B}-e^{-2 i \beta_{1}^{0} B}\right|^{2} \\
& =\left|U_{\alpha, 0}{ }^{(0)}(B)-e^{-i \beta_{1}^{0} B}\right|^{2}
\end{aligned}
$$

because we have $\left|e^{-i \beta_{1}^{0} B}\right|=1$.
Therefore,

$$
\left|r^{0}\right|^{2}=\left|U_{\alpha, 0}{ }^{(0)}(B)\right|^{2}+1-2 \mathfrak{R}\left(\overline{U_{\alpha, 0}{ }^{(0)}(B)} e^{-i \beta_{1}^{0} B}\right) .
$$

Theorem 56 (Conservation of energy). Let $E_{t}, E_{r}$, and $E_{a b s}$ be defined as in Definition 54. Then, we have the energy balance

$$
E_{t}+E_{r}=1
$$

Proof. Since $k_{1} \in \mathbb{R}$ then for perfectly conducting $k=k_{1}$. Hence by using Definition 54, we have $E_{a b s}=0$. We use equation (4.15) with $w=v=U_{\alpha, 0}$ to get

$$
\begin{gather*}
\int_{\Omega \backslash \text { int } \Omega_{3}}\left|\nabla U_{\alpha, 0}\right|^{2}-\int_{\Omega \backslash \Omega_{3}}\left(k^{2}-\alpha^{2}\right)\left|U_{\alpha, 0}\right|^{2}-2 i \alpha \int_{\Omega \backslash \text { int } \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} \\
-\int_{\Gamma_{ \pm}}\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}}=2 i \beta_{1}^{0} \int_{\Gamma_{+}} e^{-i \beta_{1}^{0} B} \overline{U_{\alpha, 0}} \tag{4.53}
\end{gather*}
$$

Taking the imaginary part of equation (4.53), and using Lemma 53 with Lemma 8 leaves us with

$$
\begin{equation*}
\mathfrak{J}\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, U_{\alpha, 0}\right)_{\Gamma_{ \pm}}+2 \beta_{1}^{0} \Re\left(e^{-i \beta_{1}^{0} B}, U_{\alpha, 0}\right)_{\Gamma_{+}}=0 . \tag{4.54}
\end{equation*}
$$

Then we use Lemma 55 with equation (2.72) in Lemma 8 to get

$$
\begin{align*}
& \mathfrak{I}\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, U_{\alpha, 0}\right)_{\Gamma_{ \pm}}+2 \beta_{1}^{0} \Re\left(e^{-i \beta_{1}^{0} B}, U_{\alpha, 0}\right)_{\Gamma_{+}} \\
& =d \sum_{n \in \mathbb{Z}} \Re \beta_{j}^{n}\left|U_{\alpha, 0}{ }^{(n)}\right|^{2}+2 d \beta_{1}^{0} \Re\left(e^{-i \beta_{1}^{0} B}{\overline{U_{\alpha, 0}}}^{(0)}\right) \text {, } \\
& =d \sum_{n \in \mathbb{Z}} \mathfrak{R} \beta_{j}^{n}\left|U_{\alpha, 0}{ }^{(n)}\right|^{2}+d \beta_{1}^{0}\left(2 \mathfrak{R}\left(e^{-i \beta_{1}^{0} B}{\overline{U_{\alpha, 0}}}^{(0)}\right)+1\right)-d \beta_{1}^{0} \text {, } \\
& =d \beta_{1}^{0} \sum_{n_{\alpha}^{2}>k^{2}} R_{m}+T_{m}-d \beta_{1}^{0} \text {. } \tag{4.55}
\end{align*}
$$

We use equations (4.54) with (4.55) to get

$$
\begin{equation*}
d \beta_{1}^{0}\left(\sum_{n_{\alpha}^{2}>k^{2}} R_{m}+T_{m}-1\right)=0 \tag{4.56}
\end{equation*}
$$

which finishes the proof.

### 4.2 The discrete problem

In order to solve numerically the scattering problem, we need to discretise the variational formulation corresponding to the continuous problem.

### 4.2.1 Variational formulation

Let $X \subset H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be a finite element space with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, . ., N$, be a basis of $X$. We discretise the variational form given by the equation (4.11) and this leads us to find $U_{\alpha, 0_{h}} \in X$, for all $v_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0_{h}}, v_{h}\right)=\left(f, v_{h}\right), \tag{4.57}
\end{equation*}
$$

and subject to the constraint

$$
U_{\alpha, 0_{h}}=0, \text { on } \partial \Omega_{3}
$$

where

$$
\begin{align*}
a\left(w_{h}, v_{h}\right)= & \int_{\Omega \backslash \operatorname{int} \Omega_{3}} \nabla w_{h} \cdot \nabla \overline{v_{h}}-\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) w_{h} \overline{v_{h}}-2 i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& -\int_{\Gamma_{+}} T_{+}^{\alpha, 0} w_{h} \overline{v_{h}}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} w_{h} \overline{v_{h}}  \tag{4.58}\\
\left(f, v_{h}\right)= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{h}}, \tag{4.59}
\end{align*}
$$

for $w_{h} \in X$. Since $U_{\alpha, 0_{h}} \in X$, there exists $U_{j}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, 0_{h}}=$ $\sum_{j=1}^{N} U_{j} \psi_{j}(x, y)$. Hence, the discrete problem given by equation (4.57) is equivalent to the following linear algebraic system

$$
\begin{equation*}
A U=L \tag{4.60}
\end{equation*}
$$

with $U=U_{j}$ for $j=, \cdots, N$,

$$
A=a\left(\psi_{i}, \psi_{j}\right)
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

### 4.2.2 Truncation of the DtN map

The DtN operators that we use as transparent boundary conditions when we truncate the domain involve an infinite sum. For computational purposes we also need to truncate the summation in equation (4.7) to a finite sum. Let $M \in \mathbb{N}$ and $M<\infty$, then we approximate $T_{ \pm}^{\alpha, 0}$ with

$$
\begin{equation*}
T_{ \pm}^{\alpha, 0^{M}}\left(U_{\alpha, 0_{h}}\right)(x)=\sum_{n=-M}^{M} i \beta_{j}^{n} U_{\alpha, 0_{h}}^{(n)}( \pm B) e^{i \frac{2 \pi n}{d} x} \tag{4.61}
\end{equation*}
$$

At the boundary of the truncated domain we introduced an upward propagating radiation condition (UPRC) in Section 2.1. Therefore, we will also need $U_{\alpha, 0_{h}}$ to satisfy the UPRC. First, we expand $U_{\alpha, 0_{h}}$ in a truncated Fourier series

$$
\begin{equation*}
U_{\alpha, 0_{h}}(x, y)=\sum_{n=-M}^{M} U_{\alpha, 0_{h}}^{(n)}(y) e^{i \frac{2 \pi n}{d} x}, \tag{4.62}
\end{equation*}
$$

where

$$
U_{\alpha, 0_{h}}^{(n)}(y)=\frac{1}{d} \int_{0}^{d} U_{\alpha, 0_{h}}(x, y) e^{-i \frac{2 \pi n}{d} x} d x
$$

By truncating the fundamental solution in equation (2.55), and using equation (4.1), the solution is truncated as

$$
\begin{equation*}
U_{\alpha, 0_{h}}=\sum_{n=-M}^{M} r_{j}^{n, M} e^{i \beta_{j}^{n} y+i \frac{2 \pi n}{d} x}+t_{j}^{n, M} e^{-i \beta_{j}^{n} y+i \frac{2 \pi n}{d} x}, j=1,2 \tag{4.63}
\end{equation*}
$$

where the unknowns $r_{j}^{n, M}$ and $t_{j}^{n, M}$ are complex scalars. $U_{\alpha, 0_{h}}$ contains the incident wave and must satisfy the UPRC condition in $\Omega_{1}$. It follows that $t_{j}^{n, M}=0$, for $n \sigma$ 0 and $t_{1}^{0, M}=1$ in $\Omega_{1}$ and in $\Omega_{2}$, all the $r_{j}^{n, M}=0$ (UPRC). We proceed exactly as we have done for the continuous problem but the solution is now truncated as given by equations (4.62) and (4.63).The coefficients $\beta_{j}^{n}$ are given by equation (2.43) and so, similar to equation (2.57),

$$
\begin{array}{ll}
U_{\alpha, 0_{h}}^{(n)}(y)=r_{1}^{n, M} e^{i \beta_{1}^{n} y}+\delta_{n 0} e^{-i \beta_{1}^{n} y}, & \text { in } \Omega_{1}  \tag{4.64}\\
U_{\alpha, 0_{h}}^{(n)}(y)=t_{2}^{n, M} e^{-i \beta_{2}^{n} y}, & \text { in } \Omega_{2}
\end{array}
$$

At $y= \pm B$, equations (2.58) and (2.59) become

$$
\begin{align*}
r_{1}^{n, M} & =U_{\alpha, 0_{h}}^{(n)}(B) e^{-i \beta_{1}^{n} B}-\delta_{n 0} e^{-2 i \beta_{1}^{0} B},  \tag{4.65}\\
t_{2}^{n, M} & =U_{\alpha, 0_{h}}^{(n)}(-B) e^{-i \beta_{2}^{n} B} .
\end{align*}
$$

Similarly to the continuous problem, but now we use equations (4.61) and (4.65) so that the boundary conditions for the discrete problem are given by

$$
\begin{array}{ll}
\left(T_{+}^{\alpha, 0^{M}}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0_{h}}=2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}, & \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, 0^{M}}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0_{h}}=0, & \text { on } \Gamma_{-} . \tag{4.67}
\end{array}
$$

Therefore, instead of solving directly equation (4.57), we approximate $U_{\alpha, 0}$ by $U_{\alpha, 0_{h}}^{M}$ and we solve numerically the following problem. Let $X \subset H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ be a finite element space with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, \ldots, N$, be a basis of $X$. We want to find $U_{\alpha, 0_{h}}^{M} \in X$ such that for all $v_{h} \in X$ we have

$$
\begin{equation*}
a^{M}\left(U_{\alpha, 0_{h}}^{M}, v_{h}\right)=\left(f, v_{h}\right), \tag{4.68}
\end{equation*}
$$

and subject to the constraint

$$
U_{\alpha, 0_{h}}^{M}=0, \text { on } \partial \Omega_{3}
$$

where

$$
\begin{align*}
a^{M}\left(w_{h}, v_{h}\right)= & \int_{\Omega \backslash \operatorname{int} \Omega_{3}} \nabla w_{h} . \nabla \overline{v_{h}}-\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) w_{h} \overline{v_{h}}-2 i \alpha \int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& -\int_{\Gamma_{ \pm}} T_{ \pm}^{\alpha, 0^{M}} w_{h} \overline{v_{h}} \\
\left(f, v_{h}\right)= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{h}} \tag{4.69}
\end{align*}
$$

for $w_{h} \in X$. This leads to a linear algebraic system

$$
\begin{equation*}
A^{M} U^{M}=L \tag{4.70}
\end{equation*}
$$

with $U^{M}=U_{j}^{M}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, 0_{h}}^{M}=\sum_{j=1}^{N} U_{j}^{M} \psi_{j}(x, y)$.

$$
A^{M}=a^{M}\left(\psi_{i}, \psi_{j}\right),
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

### 4.2.3 Efficiency of the grating

The efficiency of the grating with respect to each diffraction order $n$ can be computed by using equations (2.55), (4.49) and (4.63), to give

$$
\begin{align*}
R_{n}^{M} & =\frac{\beta_{1}^{n}}{\beta_{1}^{0}}\left|r_{j}^{n, M}\right|^{2}  \tag{4.71}\\
T_{n}^{M} & =\frac{\beta_{2}^{n}}{\beta_{1}^{0}}\left|t_{j}^{n, M}\right|^{2} .
\end{align*}
$$

When we replace $R^{M}$ with $R_{n}^{M}$, and $T^{M}$ with $T_{n}^{M}$, in Definitions 54 and D-19 then $E_{t}, E_{r}$ and $E_{a b s}$ will be called respectively $E_{t}^{M}, E_{r}^{M}$ and $E_{a b s}^{M}$.

### 4.2.4 Checking the energy balance using the truncated DtN map

By truncating the DTN operator and by discretising the scattering problem, Lemma 55 becomes

$$
\begin{align*}
\left|r_{1}^{n, M}\right|^{2} & =\left|U_{\alpha, 0_{h}}^{M}{ }^{(n)}(B)\right|^{2}, & n \in \mathbb{Z} \backslash\{0\} \\
\left|r_{M}^{0}\right|^{2} & =\left|U_{\alpha, 0_{h}}^{M}(B)\right|^{(0)}+1-2 \mathcal{R}\left(\overline{U_{\alpha, 0_{h}}^{M}\left({ }^{(0)}(B)\right.} e^{-i \beta_{1}^{0} B}\right) & \\
\left|t_{2}^{n, M}\right|^{2} & =\left|U_{\alpha, 0_{h}}^{M}{ }^{(n)}(-B)\right|^{2}, & n \in \mathbb{Z} . \tag{4.72}
\end{align*}
$$

We note that when $k_{2} \in \mathbb{R}, \beta_{j}^{n}$ is real or purely imaginary. Hence, we can use the DtN operator $T_{ \pm}^{\alpha, 0^{M}}$ given by equation (4.61) to check the energy balance. Let us call

$$
T^{M}=\left(T_{ \pm}^{\alpha, 0 M} \psi_{l}, \psi_{j}\right)_{\Gamma_{ \pm}}
$$

for $\{l, j\} \in\{1, \ldots N\}$. We have

$$
\begin{equation*}
\left(T_{ \pm}^{\alpha, 0^{M}} U_{\alpha, 0_{h}}^{M}, U_{\alpha, 0_{h}}^{M}\right)_{\Gamma_{ \pm}}=d \sum_{n \in \mathbb{Z}} i \beta_{j}^{n}\left|U_{\alpha, 0_{h}}^{M(n)}( \pm B)\right|^{2} \tag{4.73}
\end{equation*}
$$

by following the same process to get equation (2.70) but using Definition 4. Since $U_{\alpha, 0_{h}}^{M}=\sum_{j=1}^{N} U_{j}^{M} \psi_{j}(x, y)$ then $\left(T_{ \pm}^{\alpha, 0^{M}} U_{\alpha, 0_{h}}^{M}, U_{\alpha, 0_{h}}^{M}\right)_{\Gamma_{ \pm}}=\left(U^{M} T^{M}{\overline{U^{M}}}^{T}\right)_{\Gamma_{ \pm}}$with $U^{M}=U_{j}^{M}$ for $j \in\{1, ., N\}$. Hence, from equations (4.71) and (4.72)

$$
E_{t}^{M}+E_{r}^{M}=\frac{1}{d \beta_{1}^{0}} \mathfrak{J}\left(U^{M} T^{M}{\overline{U^{M}}}^{T}\right)-2 \Re\left(e^{-i \beta_{1}^{0} B} \overline{U_{\alpha, 0_{h}}^{M}{ }^{(0)}}\right)+1=1 .
$$

### 4.3 A priori error estimates for the exact solution

The well-posedness of each problem that we derived in Section 4.1.3 will now allow us to derive an a priori error estimate for $U_{\alpha, 0}$. However, since the sesquilinear form corresponding to $U$ is simpler to study than $U_{\alpha, 0}$, and we have already studied the regularity of $U$ in Section 3.1, then we first study an a priori error estimate of $U$ with the norm $\mathrm{k} . \mathrm{k}_{\mathcal{H}}$. This will then allow us to derive an a priori error estimate for $U_{\alpha, 0}$ and we will show an explicit dependence of the result on the wavenumber $k$ and the order of the polynomial basis $p$. We have the following three key results

- an estimate of the error from the discretisation of the continuous problem.
- an estimate of the error from truncating the DtN operator corresponding to the continuous problem.
- an estimate of the total error.


### 4.3.1 A priori error estimation of the discretised problem

In this section we will derive an upper bound on the error between the exact periodic solution $U_{\alpha, 0}$ and that found numerically by discretising the problem $U_{\alpha, 0_{h}}$. In each case we will state the discretised periodic problem in its variational form, find a regularity bound for the $\alpha$-quasi periodic exact solution $U$ in terms of the norm in Definition 49, examine the discretisation error for $U$, and then use this to derive an a priori bound on the discretisation error for the periodic solution $U_{\alpha, 0}$. Let $X \subset H_{\alpha \#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be a finite element subspace of order $p$ with $l \geq 1$, and let $\zeta_{h}$ be any regular partition of $X$ as described in Section B.3. We denote by $h$ the maximum mesh size after partitioning $\Omega \backslash \operatorname{int} \Omega_{3}$ using $\zeta_{h}$. We make the following standard assumption on the subspace $X$ [35]

$$
\begin{align*}
& \inf _{\psi \in X}\left\{\mathrm{k} v-\psi \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\frac{h}{p} \mathrm{k} \nabla v-\nabla \psi \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\left(\frac{h}{p}\right)^{\frac{1}{2}} \mathrm{k} v-\psi \mathrm{k}_{L_{\alpha \#}^{2}\left(\Gamma_{ \pm}\right)}\right. \\
&\left.+\frac{h}{p} \mathrm{k} v-\psi \mathbf{k}_{H_{\alpha \#}^{\frac{1}{\alpha}}\left(\Gamma_{ \pm}\right)}\right\} \leq C\left(\frac{h}{p}\right)^{l} \mathrm{k}^{l} \mathrm{k}_{H_{\alpha \#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} . \tag{4.74}
\end{align*}
$$

Similarly, let $X^{\alpha}$ be a finite element subspace of order $p$ of $H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$. The discretised problem corresponding to equation (4.20) is given below. Find $U_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{h}, \phi\right)=(f, \phi)_{\Gamma_{+}} \tag{4.75}
\end{equation*}
$$

with

$$
\begin{align*}
a\left(U_{h}, \phi\right) & =\left(\nabla U_{h}, \nabla \phi\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}-\left(k^{2} U_{h}, \phi\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}-\left(T_{ \pm} U_{h}, \phi\right)_{\Gamma_{ \pm}}  \tag{4.76}\\
(f, \phi)_{\Gamma_{+}} & =-2 i \beta_{1}^{0} \int_{\Gamma_{+}} e^{i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{\phi} \tag{4.77}
\end{align*}
$$

for all $\phi \in X$ such that $U_{h}=0$ on $\partial \Omega_{3}$ and $T_{ \pm}$are given by Definition 3.
Lemma 57. Let $k_{\text {ref }}$ be a positive scalar such that $\mathrm{k} k \mathrm{k}_{\infty} \geq k_{\text {ref }}$ and let $U \in$ $H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, then for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ we have

$$
|a(U, v)| \leq C_{c}{\mathrm{k} U \mathrm{k}_{\mathcal{H}} \mathrm{k} \mathrm{k}_{\mathcal{H}} .}
$$

such that $C_{c}=C d+1$ depends only on the period of the diffraction grating d. Proof. We note by using Cauchy- Schwarz inequality [22, p. 50] that

$$
\begin{align*}
\left|(\nabla U, \nabla v)_{\Omega \backslash \text { int } \Omega_{3}}\right| & =\left|\int_{\Omega \backslash \text { int } \Omega_{3}} \nabla U . \nabla \bar{v} d x d y\right| \\
& \leq \int_{\Omega \backslash \text { int } \Omega_{3}}|\nabla U . \nabla \bar{v}| d x d y \\
& \leq \mathrm{k} \nabla U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} \nabla v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \tag{4.78}
\end{align*}
$$

and also that

$$
\begin{align*}
\left|\left(k^{2} U, v\right)_{\Omega \backslash \text { int } \Omega_{3}}\right| & =\left|\int_{\Omega \backslash \operatorname{int} \Omega_{3}} k^{2} U \bar{v} d x d y\right| \\
& \leq \mathrm{k} k^{2} \mathbf{k}_{\infty} \int_{\Omega \backslash \operatorname{int} \Omega_{3}}|U \bar{v}| d x d y \\
& \leq \mathrm{k} k^{2} \mathbf{k}_{\infty} \mathrm{k} U \mathrm{k}_{L_{\alpha \# 1}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathbf{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \tag{4.79}
\end{align*}
$$

We note that

$$
\begin{align*}
\left|\int_{\Gamma_{ \pm}} T_{ \pm} U \bar{v} d x\right|= & d\left|\sum_{n \in \mathbb{Z}} i \beta_{j}^{n} U^{\left(n_{\alpha}\right)}( \pm B) \overline{v^{\left(n_{\alpha}\right)}( \pm B)}\right|  \tag{4.80}\\
= & d\left|\sum_{n \in \mathbb{Z}} i \beta_{j}^{n}\left(1+n_{\alpha}^{2}\right)^{-1 / 4}\left(1+n_{\alpha}^{2}\right)^{1 / 4} U^{\left(n_{\alpha}\right)}( \pm B) \overline{v^{\left(n_{\alpha}\right)}( \pm B)}\right| \\
\leq & d\left(\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|^{2}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}\right)^{1 / 2} \\
& \times\left(\sum_{n \in \mathbb{Z}}\left(1+n_{\alpha}^{2}\right)^{1 / 2}\left|v^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}\right)^{1 / 2} \tag{4.81}
\end{align*}
$$

We use equations (2.68) and (2.69)

$$
\begin{aligned}
\sum_{n \in \mathbb{Z}}\left|\beta_{j}^{n}\right|^{2}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \leq & c_{2} \sum_{n \in \mathbb{Z}: n_{\alpha}^{2}<\left|k_{j}^{2}\right|}\left|k_{j}^{2}\right|\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& +c_{1} \sum_{n \in \mathbb{Z}: n_{\alpha}^{2}>\left|k_{j}^{2}\right|} n_{\alpha}^{2}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2}, \\
\leq & c_{2}\left|k_{j}^{2}\right| \sum_{n \in \mathbb{Z}}\left(1+n_{\alpha}^{2}\right)^{-1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
& +c_{1} \sum_{n \in \mathbb{Z}}\left(1+n_{\alpha}^{2}\right)^{1 / 2}\left|U^{\left(n_{\alpha}\right)}( \pm B)\right|^{2} \\
\leq & c_{2}\left|k_{j}^{2}\right| \mathbf{k} U \mathrm{k}_{H_{\alpha}^{-\frac{1}{\#}}\left(\Gamma_{ \pm}\right)}^{2}+c_{1} \mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{2}}^{2}\left(\Gamma_{ \pm}\right)
\end{aligned}
$$

from Definition A-17. We can then write

$$
\left|\int_{\Gamma_{ \pm}} T_{ \pm} U \bar{v} d x\right|^{2} \leq C d\left(\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{H_{\alpha}^{-\frac{1}{\#}}\left(\Gamma_{ \pm}\right)}^{2}+\mathrm{k} U \mathrm{k}_{H_{\alpha \# \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2}\right) \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}^{2},
$$

where $C=\max \left\{c_{1}, c_{2}\right\}$ and so we have

$$
\begin{equation*}
\left|\int_{\Gamma_{ \pm}} T_{ \pm} U \bar{v} d x\right|^{2} \leq C d\left(\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}^{2}+\mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}^{2}\right) \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}^{2} \tag{4.82}
\end{equation*}
$$

from the trace theorem given in Theorem A-13 since and so

$$
\begin{equation*}
\left|\int_{\Gamma_{ \pm}} T_{ \pm} U \bar{v} d x\right| \leq C d \mathbf{k} U \mathbf{k}_{\mathcal{H}} \mathbf{k} v \mathbf{k}_{\mathcal{H}} \tag{4.83}
\end{equation*}
$$

from Definition 49. Since $\left({\left.\mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}\right)^{2} \geq 0 \text { then }}\right.$

$$
\begin{aligned}
& \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
\geq & 2 \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)} .
\end{aligned}
$$

Hence

$$
\begin{aligned}
& \left(|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}\right)^{2} \\
& =|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}|v|_{H_{\alpha \# 1}^{1}\left(\Omega \backslash i n t \Omega_{3}\right)}^{2}+\mathrm{k} k \mathrm{k}_{\infty}^{4} \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
& +2 \mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} U \mathrm{k}_{L_{\alpha \# \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \\
& \leq|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+\mathrm{k} k \mathrm{k}_{\infty}^{4} \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \mathrm{k} v \mathrm{k}_{L_{\alpha \# \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
& +\mathrm{k} k \mathrm{k}_{\infty}^{2}\left(\mathrm{k} U \mathrm{k}_{L_{\alpha \# 1}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2} \mathrm{k} v \mathrm{k}_{L_{\alpha \neq 1}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2}\right) \\
& =\left(|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}+\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}^{2}\right)\left(|v|_{H_{\alpha \#}^{1}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}^{2}+\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}^{2}\right) \\
& =\mathrm{k} U \mathrm{k}_{\mathcal{H}}^{2} \mathrm{k} v \mathrm{k}_{\mathcal{H}}^{2} \text {. }
\end{aligned}
$$

So


Proof. Using the duality argument [22, p. 137], we will show below that $C_{1}=$ $C C_{c} C_{\text {reg }} \mathrm{k} k \mathrm{k}_{\infty} \frac{h}{p}$. Let $w$ be the dual solution of

$$
\begin{array}{rlr}
\Delta w+k^{2} w & =\phi & (x, y) \in \Omega \backslash \operatorname{int} \Omega_{3}  \tag{4.86}\\
\left(T_{ \pm}^{*}-\partial_{n}\right) w & =0 & \text { on } \Gamma_{ \pm},
\end{array}
$$

for all $\phi, w \in H_{\alpha \# \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ where $T_{ \pm}^{*}$ are the dual operators of $T_{ \pm}[61$, p. 476]. Then [22, p. 146],

$$
\mathrm{ke}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}=\sup _{\phi \in C_{\infty}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \frac{\left|\left(e_{h}, \phi\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}\right|}{\mathrm{k} \phi \mathrm{k}_{L_{\alpha \#}^{2}}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}
$$

Multiplying both sides of equation (4.86) by $e_{h}$ and integrating this becomes

$$
\begin{align*}
\mathrm{ke}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} & =\sup _{\phi \in C_{\infty}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \frac{\left|\left(e_{h}, \Delta w+k^{2} w\right)_{\Omega \backslash \text { int } \Omega_{3}}\right|}{\mathrm{k} \phi \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}} \\
& =\sup _{\phi \in C_{\infty}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \frac{\left|a\left(e_{h}, w\right)\right|}{\mathbf{k} \phi \mathbf{k}_{L_{\alpha \# 1}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}}, \tag{4.87}
\end{align*}
$$

by integrating by parts. We use Galerkin orthogonality [22, Prop. 2.5.9], which is

$$
\begin{equation*}
a\left(e_{h}, \psi\right)=0 \tag{4.88}
\end{equation*}
$$

for all $\psi \in X$, and so equation (4.87) becomes

$$
\begin{equation*}
\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}=\sup _{\phi \in C_{\infty}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \frac{\left|a\left(e_{h}, w-\psi\right)\right|}{\mathbf{k} \phi \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}} \tag{4.89}
\end{equation*}
$$

So, from equation (4.76)

$$
\begin{aligned}
& \left|a\left(e_{h}, w-\psi\right)\right|=\left|\left(\nabla e_{h}, \nabla(w-\psi)\right)_{\Omega \backslash \text { int } \Omega_{3}}-\left(k^{2} e_{h}, w-\psi\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}-\left(T_{ \pm} e_{h}, w-\psi\right)_{\Gamma_{ \pm}}\right| \\
& \left.\leq\left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash i n t \Omega_{3}\right)}|w-\psi|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}+\mathbf{k} k \mathbf{k}_{\infty}^{2} \mathbf{k}_{h} \mathbf{k}_{L_{\alpha \#}^{2}(\Omega \backslash i n t} \Omega_{3}\right) \mathbf{k} w-\psi \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \\
& \left.+C d \mathbf{k} e_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w-\psi \mathbf{k}_{H_{\alpha \#}^{1}(\Omega \backslash i n t} \Omega_{3}\right)
\end{aligned}
$$

from equation (4.82) and using Cauchy-Schwarz inequality [22, p. 50]. Hence,

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq(C d+1) \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w-\psi \mathbf{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}
$$

using equation (4.84). We use the standard approximation estimate in finite element space given by equation (4.74) with $w-\phi$ to get

$$
\begin{equation*}
\left|a\left(e_{h}, w-\psi\right)\right| \leq C(C d+1) \frac{h}{p} \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w \mathbf{k}_{H_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \tag{4.90}
\end{equation*}
$$

We use the result from Theorem 27 and we have the regularity estimate

$$
\mathrm{k} w \mathrm{k}_{H_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq\left(C_{r e g} \mathrm{k} k \mathrm{k}_{\infty}\right) \mathbf{k} \phi \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}
$$

Using this in equation (4.90) we see

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq C C_{r e g} \mathbf{k} k \mathbf{k}_{\infty}(C d+1) \frac{h}{p} \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} \phi \mathbf{k}_{L_{\alpha}^{2} \neq\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)},
$$

and equation (4.89) gives

$$
\mathrm{k} e_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \leq C C_{r e g} \mathrm{k} k \mathrm{k}_{\infty}(C d+1) \frac{h}{p} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}
$$

and we finish the proof by letting $C_{1}=C C_{r e g} \mathrm{k} k \mathbf{k}_{\infty}(C d+1) \frac{h}{p}$.
The previous three lemmas now allow us to derive the following a priori error estimate for the periodic solution $U_{\alpha, 0}$.

Theorem 60. Let the wavenumber $|k| \geq k_{\text {ref }}>0$, let the maximum mesh size $h \in\left[0, h_{0}\right]$, and let the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$, and $C_{3}=$ $1-\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}>0$ with $C_{1}$ as given in Lemma 59. Let $U_{\alpha, 0}$ be the continuous solution of equation (4.47) then $U_{\alpha, 0_{h}} \in X^{\alpha}$ the corresponding discretised solution is unique. In addition, if $e_{\alpha, 0_{h}}=U_{\alpha, 0}-U_{\alpha, 0_{h}}$ then there exists a constant $C_{c}$ which only depends on the period of the grating $d$ and $k_{\text {ref }}$ such that

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{C_{c}}{C_{3}} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}},
$$

and

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq 2 C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}},
$$

for all test functions $\psi_{\alpha, 0} \in X^{\alpha}$, where $C_{c}$ is given in Lemma 57 and $C_{1}$ is given in Lemma 59.

Proof. Let us denote $e_{h}=U-U_{h}$, and let $\psi=e^{i \alpha x} \psi_{\alpha, 0}$, then by using Lemma 58, we get

$$
\left(\left|e_{h}\right|_{H_{\alpha \#( }^{1}\left(\Omega \backslash i n t \Omega_{3}\right)}^{2}-\mathfrak{R}(k) \mathrm{k}_{e_{h}} \mathrm{k}_{L_{\alpha \# \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}\right) \leq\left|a\left(e_{h}, e_{h}\right)\right| .
$$

Since we have the Galerkin orthogonality, as in equation (4.88), the right hand side can be written

$$
\left(\left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash i n t \Omega_{3}\right)}^{2}-\Re(k) \mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}\right) \leq\left|a\left(e_{h}, U-\psi\right)\right| .
$$

We use Lemma 57 to get

$$
\begin{gathered}
\left(\left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-\Re(k) \mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}\right) \\
\leq C_{c} \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}} \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}} .
\end{gathered}
$$

Since $\left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash i n t \Omega_{3}\right)} \leq \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}}$, then dividing both sides by $\mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}$ gives

$$
\left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}-\Re(k) \operatorname{ke}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq C_{c} \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}}
$$

and so

$$
\begin{aligned}
& \left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash i n t \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty} \mathrm{Ke}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \\
& -\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) \mathrm{k}_{h} \mathrm{k}_{L_{\alpha \# \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq C_{c} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} .
\end{aligned}
$$

From Lemma 59 and using Definition 49, we have

$$
\begin{equation*}
\mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}}-\left(\Re(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}} \leq C_{c} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} . \tag{4.91}
\end{equation*}
$$

Since $C_{3}=1-\left(\mathfrak{R}(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}>0$ then

$$
\begin{equation*}
\mathrm{k}_{h} \mathrm{k}_{\mathcal{H}} \leq \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} . \tag{4.92}
\end{equation*}
$$

So we have from Definition 49

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} & \leq\left|e_{\alpha, 0_{h}}\right|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathbf{k} k \mathrm{k}_{\infty} \mathrm{K} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}, \\
& \leq 2\left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty} \mathrm{K}_{h} \mathrm{k}_{L_{\alpha \# \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)},
\end{aligned}
$$

from equation (4.41) and so

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 2\left(\left|e_{h}\right|_{H_{\alpha \neq 1}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty} \mathrm{k}_{\mathrm{h}_{h}} \mathrm{k}_{L_{\alpha \neq 1}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}\right) \leq 2 \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}} . \tag{4.93}
\end{equation*}
$$

From equation (4.92)

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 2 \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} . \tag{4.94}
\end{equation*}
$$

We also note

$$
\begin{aligned}
\mathrm{k}_{h} \mathrm{k}_{\mathcal{H}} & =\mathrm{k} e^{i \alpha x} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}}, \\
& =\left|e^{i \alpha x} e_{\alpha, 0_{h}}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty} \mathrm{k} e^{i \alpha x} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\alpha \neq}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}, \\
& =\mathrm{k} i \alpha e^{i \alpha x} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\left|e_{\alpha, 0_{h}}\right|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty} \mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}, \\
& =\mathrm{k} i \alpha e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\left|e_{\alpha, 0_{h}}\right|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty} \mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)},
\end{aligned}
$$

using Definition A-10. Then we get

$$
\mathbf{k} e_{h} \mathbf{k}_{\mathcal{H}} \leq 2 \mathbf{k} k \mathbf{k}_{\infty} \mathbf{k} e_{\alpha, 0_{h}} \mathbf{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}+\left|e_{\alpha, 0_{h}}\right|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}
$$

since $|\alpha| \leq \mathrm{k} k \mathrm{k}_{\infty}$. Hence,

$$
\begin{equation*}
\mathrm{ke}_{h} \mathrm{k}_{\mathcal{H}} \leq 2 \mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \tag{4.95}
\end{equation*}
$$

from Definition 49. If we replace $e_{h}$ with $U-\psi$ and use equation (4.94) then

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{C_{c}}{C_{3}} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} .
$$

Once more, from Lemma 59 and equation (4.92), we get

$$
\begin{equation*}
\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \leq C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} . \tag{4.96}
\end{equation*}
$$

From Theorem 50, we have

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}=\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \neq}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)},
$$

Using this in equation (4.96), we get

$$
\begin{align*}
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\# \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} & \leq C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}}, \\
& \leq 2 C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \tag{4.97}
\end{align*}
$$

from equation (4.95). Now, let us show that $U_{\alpha, 0_{h}} \in X^{\alpha}$ exists and is unique. Since $X^{\alpha}$ has a finite dimension then a solution exists if only if it is unique. Let us suppose that we have two solutions $U_{\alpha, 0_{h 1}}$ and $U_{\alpha, 0_{h} 2}$. Then, when $\frac{h}{p}$ goes to zero

$$
\begin{aligned}
\mathrm{k} U_{\alpha, 0_{h 1}}-U_{\alpha, 0_{h 2}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} & \leq \mathrm{k} U_{\alpha, 0_{h 1}}-U_{\alpha, 0} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathrm{k} U_{\alpha, 0_{h}}-U_{\alpha, 0_{h} 2} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash i \mathrm{int} \Omega_{3}\right)} \\
& \leq 0
\end{aligned}
$$

since $C_{1}$ also goes to zero in equation (4.97) which finishes our proof.

### 4.3.2 A priori error estimation of the continuous problem by truncating the $\operatorname{DtN}$ operator

As discussed in Section 2.2, we have introduced the parameter $b$ for three reasons. Firstly, for computational efficiency, so that when we are far from the scatterer and $|y|>b$, then we can use a coarse mesh. Secondly, to derive an a priori error estimate, the Rayleigh expansion is used in the region $b \leq|y| \leq B$. Finally, it also allows us to cope with more general problems involving layered geometry such as cladding or a substrate.
Let $U_{\alpha, 0}^{M}$ be the approximated solution of the continuous problem of equation (4.48) when we truncate the DtN map, by approximating $T_{ \pm}^{\alpha, 0}$ with $T_{ \pm}^{\alpha, 0^{M}}$ for $M \in \mathbb{N}$ where Definition 4 becomes

$$
\begin{equation*}
T_{ \pm}^{\alpha, 0^{M}}(v)=\sum_{n=-M}^{M} i \beta_{j}^{n} v^{\left(n_{\alpha}\right)}( \pm B) e^{i \frac{2 \pi n}{d} x} \tag{4.98}
\end{equation*}
$$

Then, the error estimate by truncating $T_{ \pm}^{\alpha, 0}$ is given in the following theorem.

Theorem 61. Let us choose $M \in \mathbb{N}$ such that $M>M_{0}=|k|+|\alpha|$ and let us denote by

$$
e_{\alpha, 0}^{M}=U_{\alpha, 0}-U_{\alpha, 0}^{M} .
$$

If $\left(\mathfrak{R}(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}<1$, with $C_{1}$ as given in Lemma 59, so that $C_{3}=1-\Re(k)+$ $\mathrm{k} k \mathrm{k}_{\infty} C_{1}>0$ then $U_{\alpha, 0}^{M}$ is unique and we have

$$
\begin{align*}
\mathrm{ke}_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}} & \leq 4 \frac{d}{C_{3}} e^{-(B-b) c_{\min }} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}\left(\Gamma_{1, \pm}\right)}},  \tag{4.99}\\
{\mathrm{k} e_{\alpha, 0}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}} \leq 2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{\min }} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} & \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)}, \tag{4.100}
\end{align*}
$$

with $\Gamma_{1, \pm}=\{(x, \pm b) \in \Omega\}$ where $b$ is as shown in Figure 2.3. The parameter $z_{n}$ is given by equation (2.44) and $c_{\text {min }}=\left.\inf \right|_{n \left\lvert\,>\frac{M d}{2 \pi}\right.} \sin \left(z_{n} / 2\right)$.

Proof. Let $U \in H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ satisfy equation (4.20). Since we truncate the DtN map, we are approximating this problem by finding $U^{M} \in H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$, such that $a^{M}\left(U^{M}, v\right)=(f, v)_{\Gamma_{+}}$with equations (4.18) and (4.19) becoming

$$
\begin{align*}
a^{M}\left(U^{M}, v\right) & =\left(\nabla U^{M}, \nabla v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} U^{M}, v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}^{M} U^{M}, v\right)_{\Gamma_{ \pm}}  \tag{4.101}\\
(f, v)_{\Gamma_{+}} & =\left(-2 i \beta_{1}^{0} e^{i\left(\alpha x-\beta_{1}^{0} B\right)}, v\right)_{\Gamma_{+}} \tag{4.102}
\end{align*}
$$

for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ such that

$$
\begin{equation*}
T_{ \pm}^{M}(v)=\sum_{n=-M}^{M} i \beta_{j}^{n} v^{\left(n_{\alpha}\right)}( \pm B) e^{i n_{\alpha} x} \tag{4.103}
\end{equation*}
$$

Then we have

$$
\begin{aligned}
a(U, v)-a^{M}\left(U^{M}, v\right)= & (\nabla U, \nabla v)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} U, v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm} U, v\right)_{\Gamma_{ \pm}} \\
& -\left(\nabla U^{M}, \nabla v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}+\left(k^{2} U^{M}, v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}+\left(T_{ \pm}{ }^{M} U, v\right)_{\Gamma_{ \pm}} \\
= & 0 .
\end{aligned}
$$

Let us denote $e^{M}=U-U^{M}$, since $T_{ \pm}=T_{ \pm}^{M}+\left(T_{ \pm}-T_{ \pm}^{M}\right)$ then

$$
\begin{equation*}
\left(\nabla e^{M}, \nabla v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} e^{M}, v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}^{M} e^{M}, v\right)_{\Gamma_{ \pm}}=\left(\left(T_{ \pm}-T_{ \pm}^{M}\right) U, v\right)_{\Gamma_{ \pm}} . \tag{4.104}
\end{equation*}
$$

We first note that

$$
\begin{equation*}
\left(\left(T_{ \pm}-T_{ \pm}^{M}\right) U, v\right)_{\Gamma_{ \pm}}=d \sum_{|n| \geq \frac{M d}{2 \pi}} i \beta_{j}^{n} U^{\left(n_{\alpha}\right)}( \pm B) \overline{v^{\left(n_{\alpha}\right)}( \pm B)} . \tag{4.105}
\end{equation*}
$$

Since $i \beta_{j}^{n}=\left(i \cos \left(z_{n} / 2\right)-\sin \left(z_{n} / 2\right)\right) \sqrt{n_{\alpha}^{2}-k^{2}}$ when $n_{\alpha}^{2}>k^{2}$ from equation (2.43) and since $b<B$ as shown in Figure 2.3, therefore we can use equation (2.60) to write

$$
\begin{equation*}
g^{\left(n_{\alpha}\right)}( \pm B)=g^{\left(n_{\alpha}\right)}( \pm b) e^{-(B-b) \sin \left(z_{n} / 2\right)} \sqrt{ } \overline{n_{\alpha}^{2}-k_{j}^{2}}, \quad \text { for } j=1,2 \tag{4.106}
\end{equation*}
$$

Let $M_{0}=|k|+|\alpha|$ then for any $M>M_{0}$, we have $n_{\alpha}^{2}>k^{2}$ for $\left|\frac{2 \pi n}{d}\right|>M$. Hence from equation (4.105)

$$
\begin{aligned}
\left(\left(T_{ \pm}-T_{ \pm}^{M}\right) U, v\right)_{\Gamma_{ \pm}}= & -d \sum_{|n| \geq \frac{M d}{2 \pi}} \sqrt{n_{\alpha}^{2}-k^{2}} e^{\left(i \cos \left(z_{n} / 2\right)-\sin \left(z_{n} / 2\right)\right)(B-b)} \sqrt{ } \overline{n_{\alpha}^{2}-k_{j}^{2}} \\
& \times U^{\left(n_{\alpha}\right)}( \pm b) \overline{v^{\left(n_{\alpha}\right)}}( \pm B)
\end{aligned}
$$

Note that if $n \geq 0$ then $\frac{2 \pi n}{d}-|\alpha| \leq n_{\alpha} \leq \frac{2 \pi n}{d}+|\alpha|$. Therefore, $n_{\alpha} \geq M-|\alpha|$. In a similar way, we note that if $n \leq 0, n_{\alpha}^{2} \geq(M+|\alpha|)^{2}$. Hence,

$$
\begin{aligned}
& \left|\left(\left(T_{ \pm}-T_{ \pm}^{M}\right) U, v\right)_{\Gamma_{ \pm}}\right| \\
& \quad \leq d \sum_{|n| \geq \frac{M d}{2 \pi}} e^{-(B-b) \sin \left(z_{n} / 2\right)} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \sqrt{n_{\alpha}^{2}-k^{2}} U^{\left(n_{\alpha}\right)}( \pm b) \overline{v^{\left(n_{\alpha}\right)}}( \pm B)
\end{aligned}
$$

We note that

$$
\begin{aligned}
n_{\alpha}^{2}-k^{2} & \leq n_{\alpha}^{2}-k_{r e f}^{2} \\
& \leq n_{\alpha}^{2}\left(1-\frac{k_{r e f}^{2}}{n_{\alpha}^{2}}\right) \\
& \leq n_{\alpha}^{2}\left(1-\frac{k_{r e f}^{2}}{(M-|\alpha|)^{2}}\right) \\
& \leq n_{\alpha}^{2}+1
\end{aligned}
$$

since $M \geq|k|+|\alpha|$ then $M-|\alpha|>|k|>k_{\text {ref }}$. Hence

$$
\left.\begin{array}{r}
\left|\left(\left(T_{ \pm}-T_{ \pm}^{M}\right) U, v\right)_{\Gamma_{ \pm}}\right| \leq \quad d \sum_{|n| \geq \frac{M d}{2 \pi}} e^{-(B-b) \sin \left(z_{n} / 2\right)} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \sqrt{n_{\alpha}^{2}+1} \\
\leq \quad \\
\times U^{\left(n_{\alpha}\right)}( \pm b) \overline{v^{\left(n_{\alpha}\right)}}( \pm B) \\
\leq-(B-b) c_{m i n} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}
\end{array} \sum_{n \in \mathbb{Z}} \sqrt{n_{\alpha}^{2}+1}\right) \quad \times U^{\left(n_{\alpha}\right)}( \pm b) \overline{v^{\left(n_{\alpha}\right)}}( \pm B) .
$$

with $c_{\text {min }}=\left.\inf \right|_{|n|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$ and so
using Cauchy's inequality [22, p. 50] with Definition A-17 and such that $\Gamma_{1, \pm}=$ $\{(x, \pm b) \in \Omega\}$. We note that the left hand side of equation (4.104)

$$
\begin{aligned}
\mid\left(\nabla e^{M}, \nabla e^{M}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} & -\left(k^{2} e^{M}, e^{M}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}^{M} e^{M}, e^{M}\right)_{\Gamma_{ \pm}} \mid \\
& =\left|a^{M}\left(e^{M}, e^{M}\right)\right|,
\end{aligned}
$$

from equation (4.101). Taking the real part of the left hand side of equation (4.104), we have

$$
\begin{equation*}
\left|a^{M}\left(e^{M}, e^{M}\right)\right| \geq \mathfrak{R}\left(\left|e^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-k^{2} \mathbf{k} e^{M} \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}\right), \tag{4.108}
\end{equation*}
$$

since $-\mathfrak{R}\left(T_{ \pm}{ }^{M} e^{M}, e^{M}\right)_{\Gamma_{ \pm}} \geq 0$ from equation (2.72). From equation (4.104), we get

$$
\begin{align*}
\left|e^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-\Re\left(k^{2}\right) \mathbf{k e}^{M} \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2} \leq & \left|\left(\left(T_{ \pm}-T_{ \pm}^{M}\right) U, e^{M}\right)_{\Gamma_{ \pm}}\right| \\
\leq & d e^{-(B-b) c_{m i n} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \\
& \times \mathbf{k} U \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)} \mathrm{ke}^{M} \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \tag{4.109}
\end{align*}
$$

from equation (4.107). We use once again the duality argument [22, p. 137] to approximate $\mathrm{k} . \mathrm{K}_{L_{\alpha \#}^{2}(\Omega \backslash i n t} \Omega_{3}$, with the dual problem given by equation (4.86). Similar to the derivation of equation (4.91) in the proof of Theorem 60, we can use Lemma 59 with Theorem A-13 so that we can divide by $\mathrm{ke}^{M} \mathrm{k}_{\mathcal{H}}$ to get

$$
\mathrm{k} e^{M} \mathrm{k}_{\mathcal{H}}-\left(\mathfrak{R}(k)+\mathrm{k} k \mathbf{k}_{\infty}\right) C_{1} \mathrm{k}^{M} \mathrm{k}_{\mathcal{H}} \leq d e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)} . . . . ~}
$$

Since we have $C_{3}=1-\Re(k)+\mathrm{k} k \mathrm{k}_{\infty} C_{1}>0$ then we get

$$
\begin{equation*}
\mathrm{K}^{M} \mathrm{k}_{\mathcal{H}} \leq \frac{d}{C_{3}} e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}}{\mathrm{~K} U \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)}} \tag{4.110}
\end{equation*}
$$

and we can use equations (4.93) and (4.95) together with Theorem 50 to get

$$
\begin{equation*}
\mathrm{K} e_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{d}{C_{3}} e^{-(B-b) c_{\text {min }}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{\#}}\left(\Gamma_{1, \pm}\right)} . \tag{4.111}
\end{equation*}
$$

From Lemma 59 and equation (4.110), we have

$$
\begin{equation*}
\mathbf{k} e^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k}^{\left(\mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)} .\right.} \tag{4.112}
\end{equation*}
$$

From equations (4.93) and (4.95) together with Theorem 50, we conclude that

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0}^{M} \mathrm{k}_{L_{\#}^{2}(\Omega)} \leq 2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{m i n} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}\left(\Gamma_{1, \pm}\right)}} \tag{4.113}
\end{equation*}
$$

To show that $U^{M}$ is unique, we note that $\left|\int_{\Gamma_{ \pm}} T_{ \pm}^{M} U \bar{v} d x\right| \leq\left|\int_{\Gamma_{ \pm}} T_{ \pm}^{M} U \bar{v} d x\right|$. Hence, we can use equation (4.33) and we derive similar to equation (4.34) that

$$
\left|a^{M}\left(U^{M}, v\right)\right| \leq C_{0} \sup \left(1, \mathrm{k} k \mathrm{k}_{\infty}^{2}\right) \mathbf{k} U^{M} \mathbf{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}{\mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} . . . ~}
$$

Hence, $a^{M}\left(U^{M}, U^{M}\right)$ is continuous using Definition A-6. In a similar way when we derive equation (2.72), we have $\mathfrak{R}\left(T_{ \pm}^{M} U^{M}, U^{M}\right)_{\Gamma_{ \pm}} \leq 0$ and so similar to the derivation of equation (4.36) we have

$$
\left|a^{M}\left(U^{M}, U^{M}\right)+\Re\left(k^{2}\right) \mathbf{k} U^{M} \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}\right| \geq M_{1} \mathbf{k} U^{M} \mathbf{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} .
$$

Then, $a^{M}\left(U^{M}, U^{M}\right)$ is $H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$-coercive from Definition B-4. We can then use Lemma B-5 to show the existence of a solution from its uniqueness. Let us suppose that we have two solutions $U_{1}^{M}$ and $U_{2}^{M}$, then we have

$$
\mathrm{k} U_{1}^{M}-U_{2}^{M} \mathrm{k}_{L_{\alpha \neq}^{2}(\Omega)} \leq \mathrm{k} U-U_{1}^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}+\mathrm{k} U-U_{2}^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}
$$

and from equation (4.113), we see that when $M$ tends to $\infty, \mathrm{k} U_{1}^{M}-U_{2}^{M} \mathrm{k}_{L_{\alpha \neq 1}^{2}(\Omega)}=0$. Since $U^{M}$ is unique then $U_{\alpha, 0}^{M}$ is also unique.

### 4.3.3 Total error made by solving numerically the problem

The error that we make by solving numerically the Helmholtz equation for a periodic grating arises from two sources

- discretising using finite elements and
- from truncating the DtN operator when we use the transparent boundary conditions.

If we denote the total error by $e_{\alpha, 0}=U_{\alpha, 0}-U_{\alpha, 0_{h}}^{M}$ then it can be estimated as follows.

Theorem 62. Let $|k| \geq k_{r e f}>0$, the maximum mesh size $h \in\left[0, h_{0}\right]$, the degree of the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$ with $\left(\mathfrak{R}(k)+\mathbf{k} k \mathrm{k}_{\infty}\right) C_{1}<1$ where $C_{1}$ is defined in Lemma 59 so that $C_{3}=1-\left(\mathfrak{R}(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}>0$. Let $M \in \mathbb{N}$ such that $M \geq M_{0}$ and let $U_{\alpha, 0}$ be the continuous solution of equation (4.47), $U_{\alpha, 0_{h}}^{M}$ be the corresponding discretised solution with the truncated DtN operator and the total error be $e_{\alpha, 0}=U_{\alpha, 0}-U_{\alpha, 0_{h}}^{M}$. Then we have

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq & 4 \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \\
& +4 \frac{d}{C_{3}} e^{-(B-b) c_{m i n} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathrm{k} e_{\alpha, 0} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \leq 2 \frac{C_{c}}{C_{3}} C_{1} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \\
&+2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
& \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)^{\prime}},
\end{aligned}
$$

for all test functions $\psi_{\alpha, 0} \in X^{\alpha}$ where $C_{c}$ is given in Lemma 57 and $c_{\text {min }}=\inf _{|n|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$ with $z_{n}$ as defined in equation (2.44). Note that

$$
\begin{equation*}
\inf \left(1, \mathrm{k} k \mathbf{k}_{\infty}\right) \mathrm{k} e_{\alpha, 0} \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq \mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq \sup \left(1, \mathrm{k} k \mathbf{k}_{\infty}\right) \mathrm{k}_{\alpha, 0} \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \tag{4.114}
\end{equation*}
$$

Then we can use the standard finite element estimate equation (4.74) and we can write

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq & 4 \sup \left(1, \mathrm{k} k \mathbf{k}_{\infty}\right)\left(\frac{h}{p}\right)^{l-1} \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{H_{\#}^{l}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \\
& +4 \frac{d}{C_{3}} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{z}}\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

and using the definition of $C_{1}$ from Lemma 59

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \leq & 2 \mathbf{k} k \mathrm{k}_{\infty}\left(\frac{h}{p}\right)^{l} \frac{C_{c}}{C_{3}}(C d+1) C_{r e g} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{H_{\#}^{l}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \\
& +2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{m i n} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathbf{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2( }\left(\Gamma_{1, \pm}\right)}},
\end{aligned}
$$

for any integer $l \geq 2$.

Proof. We have

$$
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq \mathrm{k} U_{\alpha, 0}-U_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}}+\mathrm{k} U_{\alpha, 0}^{M}-U_{\alpha, 0_{h}}^{M} \mathrm{k}_{\mathcal{H}}
$$

where an a priori error estimate of $\mathbf{k} U_{\alpha, 0}-U_{\alpha, 0}^{M} \mathbf{k}_{\mathcal{H}}$ has already been shown in Theorem 61. An a priori error estimate for the second term can be derived in a similar way to that performed in Theorem 61; the only thing that changes is that $T_{ \pm}^{\alpha, 0}$ is approximated by $T_{ \pm}^{\alpha, 0^{M}}$. Let us denote $e_{h}^{M}=U^{M}-U_{h}^{M}$ where $U^{M}=e^{i \alpha x} U_{\alpha, 0}^{M}$ and $U_{h}^{M}=e^{i \alpha x} U_{\alpha, 0_{h}}^{M}$. By a similar argument to that used in Theorem 61 to derive equation (4.108) we can derive the following equations

$$
\begin{equation*}
\mathfrak{R}\left(a^{M}\left(e_{h}^{M}, e_{h}^{M}\right)\right) \geq \mathfrak{R}\left(\left|e_{h}^{M}\right|_{H_{\alpha \# \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-k^{2} \mathbf{k} e_{h}^{M} \mathbf{k}_{L_{\alpha \# 1}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}\right) \tag{4.115}
\end{equation*}
$$

since $-\mathfrak{R}\left(T_{ \pm}^{M} e_{h}^{M}, e_{h}^{M}\right)_{\Gamma_{ \pm}}>0$ using Lemma 8. So,

$$
\left|e_{h}^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}^{2}-\mathfrak{R}\left(k^{2}\right) \mathbf{k} e_{h}^{M} \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathbf{k} e_{h}^{M} \mathbf{k}_{\mathcal{H}} \leq\left|a^{M}\left(e_{h}^{M}, e_{h}^{M}\right)\right|
$$

since $\mathscr{R}(k) \mathrm{K}_{h}^{M} \mathrm{k}_{L_{\alpha \neq \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq \mathrm{K}_{h}^{M} \mathrm{k}_{\mathcal{H}}$. So

$$
\begin{align*}
\left|e_{h}^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2}-\mathcal{R}(k) \mathbf{k} e_{h}^{M} \mathbf{k}_{L_{\alpha \# 1}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \mathrm{k}_{h}^{M} \mathrm{k}_{\mathcal{H}} & \leq\left|a^{M}\left(e_{h}^{M}, U^{M}-\psi\right)\right| \\
& \leq C_{c} \mathrm{k}_{h}^{M} \mathrm{k}_{\mathcal{H}} \mathbf{k} U^{M}-\psi \mathbf{k}_{\mathcal{H}} . \tag{4.116}
\end{align*}
$$

using Galerkin orthogonality and Lemma 58. Similar to Lemma 59, we can show that

$$
\begin{equation*}
\mathbf{k} e_{h}^{M} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq C_{1} \mathrm{ke}_{h}^{M} \mathrm{k}_{\mathcal{H}} \tag{4.117}
\end{equation*}
$$

By using equations (4.116) with (4.117), dividing by $\mathrm{k} e_{h}^{M} \mathrm{~K}_{\mathcal{H}}$ and following a similar argument to the proof of Theorem 60 we get

$$
\mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}}-\left(\Re(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1} \mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}} \leq C_{c} \mathrm{k} U^{M}-\psi \mathbf{k}_{\mathcal{H}} \leq C_{c} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}}
$$

since for $M>M_{0}, U^{M}$ tends to $U$. If we suppose that $\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}<1$ then $C_{3}=1-\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}>0$ and

$$
\begin{equation*}
\mathrm{k}_{h}^{M} \mathrm{k}_{\mathcal{H}} \leq \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} . \tag{4.118}
\end{equation*}
$$

From equations (4.93) and (4.95), we get

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0_{h}}^{M} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} . \tag{4.119}
\end{equation*}
$$

From Lemma 59 and Theorem 50, we get

$$
\begin{equation*}
\mathrm{ke}_{\alpha, 0_{h}}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq 2 \frac{C_{1}}{C_{3}} C_{c} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \tag{4.120}
\end{equation*}
$$

We use the result given by Theorem 61 with equations (4.119) and(4.120) to finish the proof of the total error of discretising and truncating the DtN operator.

### 4.4 Summary

In this chapter, we have transformed the diffraction problem from the $\alpha$-quasi periodic space to a periodic space. To emphasise the essential points in our analysis, we focused on Case 1A and we relegated Case 1B, 2A and 2B to Appendix D. We investigated the a priori error estimates that arise through discretisation and through truncating the DtN map. We started by studying the continuous problem where we gave a variational formulation of the periodic problem. To derive an a priori error estimate the problem must be shown to be well-posed. In order for a problem to be well posed, the solution needs to exist, be unique and to depend continuously on the data. In Lemma 48, we showed that the solution of the variational problem exists and that it is unique. The continuous dependence on the data, was then shown in Theorem 51 when we studied the regularity of the solution. We also showed that there is an equivalence between the variational formulation corresponding to the scattering problem in periodic space with that in $\alpha$-quasi periodic space. The variational formulation in the latter case is algebraically simpler and we also have the regularity results from Chapter 3. Hence, we derived a priori error estimates in the $\alpha$-quasi periodic setting first and then used these to derive the a priori error estimate in the periodic case. If we compare the a priori error estimate in the $\alpha$-quasi periodic space given by combining equations (4.110) and (4.118) with the a priori error estimate in the periodic space given by Theorem 62, they just differ by a constant factor. This shows that there is no significant difference in studying either the $\alpha$-quasi periodic solution or the periodic solution. Having dealt with the continuous problem, we then considered the discrete problem that arises when we approximate the continuous problem with a finite element solution. Since applying the $\alpha$-quasi periodic constraints in the finite element method is more expensive than with periodic constraints, we solve numerically the periodic problem rather than the $\alpha$-quasi periodic one. In Theorem 62, we derived an a priori error estimate that arises due to the discretisation and the truncation of the DtN map. For the discretisation, we derived an explicit dependency of this error on the maximum mesh size $h$, the degree of the polynomial basis $p$ and the wavenumber $k$. It transpired that the form of the dependency of this error on these parameters is $\frac{h}{p} \mathrm{k} k \mathrm{k}_{\infty}$. This shows that for a large wavenumber, we need a finer mesh and a higher order for the polynomial basis. It also transpires that the a priori error estimate for the truncation of the DtN map had an exponential convergence rate with the number of Fourier terms used. This indicates that the number of Fourier terms used when we solve the problem numerically plays a minor role compared to our choice of the mesh size $h$ and the order of the polynomial basis $p$. Having derived these error estimates we then showed that these discretised problems also had unique solutions. We will show in Chapter 7 that this error estimate will allow us to use a uniform mesh to solve the Helmholtz problem and show that the corresponding solution is bounded.

In the following chapter, we will extend our study to an a posteriori error
estimate of the above transformation which will allow us to implement an adaptive algorithm to solve our Helmholtz problem.

## Chapter 5

## An a posteriori error estimate using the dual weighted residual method and the $\alpha, 0$-quasi periodic transformation

In the previous chapter we studied a priori error estimation. This consists of finding an upper bound on the error between the exact and the approximate solutions in terms of the exact solution and the solution to the dual problem. This upper bound also depends on some stability constants whose dependency on the system parameters is unknown although they are independent on the wavenumber $k$, the mesh size $h$ and the degree of the polynomial basis $p$. The goal in studying the a priori error estimate is to guarantee that the discrete solution converges to the exact solution provided that we keep the interpolation error small with respect to the wavenumber $k$. In contrast to the a priori error estimate, the a posteriori error estimate will provide a computable upper bound. This error estimate is later used to compute efficiently the numerical solution using an adaptive method.

In this chapter, we will derive an a posteriori error estimate that arises when we discretise the Helmholtz problem for a periodic grating. Given a grating profile, one of the main concerns is to know how much of the incident wave will be reflected and how much will be transmitted. This is achieved by computing the efficiency of the grating. Hence, rather than estimating the energy norm of the error in the solution to the Helmholtz problem, we want to estimate a particular linear functional of this error. This linear functional denoted by $Q$, is chosen so that we minimise the error made by computing the efficiency of the grating. To begin with, we will introduce some basic concepts concerning a posteriori error estimates, and recall the direct problem corresponding to Cases 1 and 2. We will then introduce the dual problem which will then allow us to estimate this linear functional of the error.

### 5.1 Introduction

When we solve a boundary value problem numerically, it produces an approximation of the solution rather than the exact solution. In order to achieve a reliable and efficient numerical method, we need to know how the system parameters affect the error between the exact and approximate solution. According to [43], these errors can be classified into three different sources which are the data, the modelling and the computation errors. Since we use the finite element method we will focus on the error arising from the discretisation (data and modeling).

Since we have already studied the a priori error estimate of our scattering problem in Chapter 4, we will focus in this chapter on a posteriori error estimation. The approach in this latter case is different from the treatment of the a priori error in that the a posteriori error estimate will be used to refine locally the mesh (hversion) or raise the degree of the polynomial basis (p-version). By doing so, we reduce and control both the numerical error and the computational cost. This process is known as the adaptive computation method [8, 43, 51]

Let us recall that our boundary value problem can be represented by

$$
\begin{equation*}
A U_{\alpha, 0}=f \tag{5.1}
\end{equation*}
$$

where

$$
\begin{equation*}
A=\Delta+\left(k^{2}-\alpha^{2}\right)+2 i \alpha \partial_{x} \tag{5.2}
\end{equation*}
$$

for Cases $1 \mathrm{~A}, 1 \mathrm{~B}$ and 2 A and

$$
\begin{equation*}
A=\nabla_{\alpha} \cdot\left(\frac{1}{k^{2}} \nabla_{\alpha}\right)+I_{d} \tag{5.3}
\end{equation*}
$$

for Case 2B with $\nabla_{\alpha}$ as given in Lemma D-3 and $I_{d}$ is the identity operator, $A: V \rightarrow V$ is a linear operator on $V$ (a Sobolev space with inner product $(., .)_{V}$ and norm $\mathbf{k} . \mathbf{k}$ ), and $f$ is some given data. For all cases, $U_{\alpha, 0}$ also satisfies the following boundary conditions

$$
\begin{array}{ll}
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}, & \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=0, & \text { on } \Gamma_{-} . \tag{5.5}
\end{array}
$$

and in addition for Case 1

$$
\begin{align*}
&\left.U_{\alpha, 0}\right|_{\partial \Omega_{3}}=0, \\
&\left.\partial_{n} U_{\alpha, 0}\right|_{\partial \Omega_{3}}=0,  \tag{5.6}\\
& \text { for Cor Case } 1 \mathrm{~A}, \\
& 1 \mathrm{~B}
\end{align*}
$$

$\Gamma_{+}$is defined in equation (2.37), $\Gamma_{-}$in equation (2.38), and $T_{ \pm}^{\alpha, 0}$ in Definition 4. We denote by $U_{\alpha, 0}$ the exact solution of the boundary value problem and by $U_{\alpha, 0_{h}} \in$ $X^{\alpha}$, such that $X^{\alpha} \subset V$, the corresponding numerical solution. There are two
different approaches to estimate the approximation error $U_{\alpha, 0}-U_{\alpha, 0_{h}}$. The first approach consists on looking for upper and lower bounds on $U_{\alpha, 0}-U_{\alpha, 0_{h}}$ (global error estimate) and the second one is to use a quantitative estimate of some local feature of $U_{\alpha, 0}$ called the quantity of interest (strain, displacement etc) [43], and to look for an estimate of the error in this quantity of interest (goal oriented error estimate).

### 5.1.1 Global error estimate

In this approach, we are looking for the error $e_{h}=U_{\alpha, 0}-U_{\alpha, 0_{h}}$ in a global norm such as the $L^{2}$ norm. Below is the general framework that one follows in order to derive a global error estimate.

To derive a global error estimate for equation (B.2), we first consider the dual problem

$$
\begin{equation*}
A^{*} z=e_{h} \tag{5.7}
\end{equation*}
$$

where $A^{*}$ denotes the adjoint operator of $A$. We then can represent the error $e_{h}=U_{\alpha, 0}-U_{\alpha, 0_{h}}$ in terms of the residual of the finite element solution and the solution $z$ of the continuous dual problem equation (5.7). We have

$$
\begin{align*}
\mathrm{k} e_{h} \mathrm{k}^{2} & =\left(e_{h}, e_{h}\right)_{V}, \\
& =\left(e_{h}, A^{*} z\right)_{V}, \\
& =\left(A e_{h}, z\right)_{V}, \\
& =\left(f-A U_{\alpha, 0_{h}}, z\right)_{V}=-\left(R\left(U_{\alpha, 0_{h}}\right), z\right)_{V} \tag{5.8}
\end{align*}
$$

with the residual $R\left(U_{\alpha, 0_{h}}\right)$ defined by

$$
\begin{equation*}
R\left(U_{\alpha, 0_{h}}\right)=A U_{\alpha, 0_{h}}-f . \tag{5.9}
\end{equation*}
$$

Since $\left(A e_{h}, v\right)_{V}=\left(-R\left(U_{\alpha, 0_{h}}\right), v\right)_{V}=0$ by using Galerkin orthogonality similar to equation (4.88) for all $v \in X^{\alpha}$ then we can write

$$
\begin{equation*}
\mathbf{k} e_{h} \mathbf{k}^{2}=-\left(R\left(U_{\alpha, 0_{h}}\right), z-z_{h}\right)_{V} \tag{5.10}
\end{equation*}
$$

such that $z_{h} \in X^{\alpha}$ is an interpolation of $z$.
We now use a local interpolation estimate which follows from classical interpolation theory [43] of the form

$$
\begin{equation*}
\mathrm{k}(h / p)^{-l}\left(z-z_{h}\right) \mathrm{k} \leq C_{i} \mathrm{k} D^{l} z \mathrm{k} \tag{5.11}
\end{equation*}
$$

where $C_{i}$ is an interpolation constant, $h$ the size of the mesh and $p$ the polynomial basis degree and $D^{l} z$ as given in Definition A-10.

Next we need to prove the regularity for the dual continuous problem so that we can write [43]

$$
\begin{equation*}
\mathrm{k} D^{l} z \mathrm{k} \leq C_{r} \mathrm{k} e_{h} \mathrm{k} . \tag{5.12}
\end{equation*}
$$

By combining equations (5.8), (5.10), (5.11) and (5.12) we have

$$
\begin{aligned}
\mathbf{k} e_{h} \mathbf{k}^{2} & =\left(R\left(U_{\alpha, 0_{h}}\right), z_{h}-z\right)_{V}, \\
& \leq C_{r} C_{i} \mathrm{k}(h / p)^{l} R\left(U_{\alpha, 0_{h}}\right) \mathrm{kk} e_{h} \mathrm{k} .
\end{aligned}
$$

Hence we have

$$
\begin{equation*}
\mathrm{k} U_{\alpha, 0}-U_{\alpha, 0_{h}} \mathrm{k} \leq C_{r} C_{i} \mathrm{k}(h / p)^{l} R\left(U_{\alpha, 0_{h}}\right) \mathrm{k} . \tag{5.13}
\end{equation*}
$$

This global error estimate now rests on finding the global stability constants $C_{r}$ and $C_{i}$ or some estimate if they cannot be found analytically. The determination of $C_{r}$ depends on the solution being approximated since it is linked to the continuous dual problem. If an analytical upper bound for $C_{r}$ cannot be found then we look for an approximation by solving the continuous dual problem numerically. The constant of interpolation $C_{i}$ depends on the shape of the elements, the local order of the polynomial and the choice of the norms in the finite element implementation. It can be determined using interpolation theory or through calibration by numerically solving problems with known exact solutions. Once we get equation (5.13), we choose a given tolerance (denoted by $T O L$ ) and demand that the approximate solution $u_{h}$ satisfies

$$
\mathrm{k} U_{\alpha, 0}-U_{\alpha, 0_{h}} \mathrm{k} \leq T O L
$$

by utilising an adaptive algorithm where each element is iteratively changed until it satisfies the stopping criterion given by

$$
C_{r} C_{i} \mathrm{k}(h / p)^{l} R\left(U_{\alpha, 0_{h}}\right) \mathrm{k} \leq T O L .
$$

We have kept the framework for deriving a global error estimate very general here and for more details see [92, 39, 43].

### 5.1.2 Goal oriented error estimate

In many applications, we are interested in the error that arises in some specific, real valued physical quantity of interest $Q$ that depends on the solution $U_{\alpha, 0}$. The global error does not, however, provide useful bounds for this error in the quantity of interest. Also the sensitivity of the global error to local error sources is not properly represented when we use the global stability constants [43, 51]. This issue is addressed by using a goal oriented error estimate so that the error in the quantity of physical interest can be controlled and at the same time we can optimise the efficiency of computing this quantity. In order to do so, we combine the dual problem with the direct problem to derive an estimate of the error in the target quantity $Q\left(U_{\alpha, 0}\right)-Q\left(U_{\alpha, 0_{h}}\right)$ from each local residual denoted $\rho_{K}\left(U_{\alpha, 0_{h}}\right)$ on each of the mesh cells $K$. In this way, we can control locally the error in computing our target quantity.

In order to derive a goal oriented error estimate for equation (B.2) in the quantity of interest $Q$, we consider the following dual problem rather than using
the dual problem given by equation (5.7)

$$
\begin{equation*}
\left(\phi, A^{*} z\right)_{V}=Q(\phi) \tag{5.14}
\end{equation*}
$$

for all $\phi \in V$. Similar to the global error estimate, we represent the error, which is now in the target quantity $Q\left(e_{h}\right)=Q\left(U_{\alpha, 0}\right)-Q\left(U_{\alpha, 0_{h}}\right)$, in terms of the residual of the finite element solution and the solution $z$ of the continuous dual problem given by equation (5.14). We have

$$
\begin{align*}
Q\left(e_{h}\right) & =\left(e_{h}, A^{*} z\right)_{V} \\
& =-\left(R\left(U_{\alpha, 0_{h}}\right), z\right)_{V} \tag{5.15}
\end{align*}
$$

by equation (5.8).
Once again by using Galerkin orthogonality, we can write

$$
Q\left(e_{h}\right)=-\left(R\left(U_{\alpha, 0_{h}}\right), z-z_{h}\right)_{V}
$$

such that $z_{h} \in X^{\alpha}$ is an interpolation of $z$ and which can be developed using cell-wise integration as

$$
\begin{equation*}
\mathrm{k} Q\left(e_{h}\right) \mathrm{k} \leq\left.\sum_{K} \mathrm{k}\left(R\left(U_{\alpha, 0_{h}}\right), z-z_{h}\right)_{V}\right|_{K} \mathrm{k} \tag{5.16}
\end{equation*}
$$

where $K$ represents each mesh cell. There are different methods for evaluating, or for deriving an approximation to the right hand side of equation (5.16). Examples include the energy norm based estimate, the influence function estimate and the Dual Weighted Residual (DWR) method [43, 51]. In our case, we are going to use the latter method. It is called the dual weighted residual method since the error in equation (5.16) comes from the cell residual $R\left(U_{\alpha, 0_{h}}\right)$ which we denote by $\rho_{K}$ and from the weighted dual solution $z$ that we denote $w_{K}$. We can therefore write equation (5.16) as

$$
\mathrm{k} Q\left(e_{h}\right) \mathrm{k} \leq \sum_{K} \rho_{K} w_{K}
$$

Similar to the global error estimate implementation, we demand that in our adaptive computation method the error in the targeted quantity does not exceed some chosen tolerance. Since we are interested in the grating efficiencies then the quantity of interest $Q$ is directly linked to the computation of these grating efficiencies. We start by generalising the continuous Helmholtz grating problem in such a way that Case 1 and Case 2 are recovered by suitable parameter choices. We will study the dual problem in which we define the quantity of interest $Q$, present the continuous problem and then the variational formulation. Finally, we combine the direct and the dual problem to derive an a posteriori error estimate using the dual weighted residual method.

### 5.2 Direct problem

In this section, we are going to reformulate the continuous and discrete formulation of the direct problem for the four cases.

### 5.2.1 Continuous problem

The continuous variational formulations given by equations (4.11), (D.14), (D.18) and (D.21) can be rewritten as a single problem as follows. Find $U_{\alpha, 0} \in H_{\#}^{1}(F)$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v\right)=(f, v)_{\Gamma_{+}} \tag{5.17}
\end{equation*}
$$

for all $v \in H_{\#}^{1}(F)$ with
$a\left(U_{\alpha, 0}, v\right)$
$= \begin{cases}\left(\frac{1}{q} \nabla U_{\alpha, 0}, \nabla v\right)_{F}-\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}, v\right)_{F}-2 i \alpha\left(\frac{1}{q} \partial_{x} U_{\alpha, 0}, v\right)_{F} & \\ -\left(\frac{1}{q} T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, v\right)_{\Gamma_{ \pm}}+i\left(\alpha n_{x} U_{\alpha, 0}, v\right)_{\partial F}, & \text { for Case 1B } \\ \left(\frac{1}{q} \nabla U_{\alpha, 0}, \nabla v\right)_{F}-\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}, v\right)_{F}-2 i \alpha\left(\frac{1}{q} \partial_{x} U_{\alpha, 0}, v\right)_{F} & \\ -\left(\frac{1}{q} T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, v\right)_{\Gamma_{ \pm}}, & \text {otherwise }\end{cases}$
and

$$
\begin{equation*}
(f, v)_{\Gamma_{+}}=-2 i \beta_{1}^{0} \int_{\Gamma_{+}} \frac{1}{q} e^{-i \beta_{1}^{0} B} \bar{v} \tag{5.18}
\end{equation*}
$$

where

$$
q= \begin{cases}k^{2} & \text { for Case 2B }  \tag{5.19}\\ 1 & \text { for Case 1A , Case 1B and Case 2A }\end{cases}
$$

such that $k$ is given by equation (2.33) and

$$
F= \begin{cases}\Omega \backslash \operatorname{int} \Omega_{3} & \text { for Case 1 }  \tag{5.20}\\ \Omega & \text { for Case 2 }\end{cases}
$$

and $\Gamma_{+}$is defined in equation (2.37), $\Gamma_{-}$in equation (2.38), $T_{ \pm}^{\alpha, 0}$ in Definition 4 and $n_{x}$ is the outward unit normal with respect to the $x$ axis. For Case 2B, we have used the property

$$
\begin{aligned}
\left(\partial_{x} U_{\alpha, 0}, v\right)_{F} & =\int\left(\int \partial_{x} U_{\alpha, 0} \bar{v} d x\right) d y \\
& =\int\left(\left[U_{\alpha, 0} \bar{v}\right]_{x=0}^{x=d}-\int U_{\alpha, 0} \partial_{x} \bar{v} d x\right) d y
\end{aligned}
$$

from integration by parts. Since $U_{\alpha, 0}$ and $v$ are both periodic, then

$$
\begin{equation*}
\left(\partial_{x} U_{\alpha, 0}, v\right)_{F}=-\left(U_{\alpha, 0}, \partial_{x} v\right)_{F} . \tag{5.21}
\end{equation*}
$$

We can also reformulate the corresponding discrete form.

### 5.2.2 Discretised problem

Let $X^{\alpha}$ be a finite dimensional subspace of $H_{\#}^{1}(F)$ (see Appendix B.3) then we want to find $U_{\alpha, 0_{h}} \in X^{\alpha}$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0_{h}}, v_{h}\right)=\left(f, v_{h}\right)_{\Gamma_{+}} \tag{5.22}
\end{equation*}
$$

for all $v_{h} \in X^{\alpha}$ with
$a\left(U_{\alpha, 0}, v\right)= \begin{cases}\left(\frac{1}{q} \nabla U_{\alpha, 0_{h}}, \nabla v_{h}\right)_{F}-\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0_{h}}, v_{h}\right)_{F}-2 i \alpha\left(\frac{1}{q} \partial_{x} U_{\alpha, 0_{h}}, v_{h}\right)_{F} \\ -\left(\frac{1}{q} T_{ \pm}^{\alpha, 0} U_{\alpha, 0_{h}}, v_{h}\right)_{\Gamma_{ \pm}}+i\left(\alpha n_{x} U_{\alpha, 0_{h}}, v_{h}\right)_{\partial F}, & \text { for Case 1B } \\ \left(\frac{1}{q} \nabla U_{\alpha, 0_{h}}, \nabla v_{h}\right)_{F}-\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0_{h}}, v_{h}\right)_{F}-2 i \alpha\left(\frac{1}{q} \partial_{x} U_{\alpha, 0_{h}}, v_{h}\right)_{F} \\ -\left(\frac{1}{q} T_{ \pm}^{\alpha, 0} U_{\alpha, 0_{h}}, v_{h}\right)_{\Gamma_{ \pm}}, & \text {otherwise }\end{cases}$
and

$$
\left(f, v_{h}\right)_{\Gamma_{+}}=-2 i \beta_{1}^{0} \int_{\Gamma_{+}} e^{-i \beta_{1}^{0} B} \overline{v_{h}},
$$

where $q$ and $F$ are given respectively by equations (5.19) and (5.20). We denote the discretisation error as

$$
\begin{equation*}
e_{h}=U_{\alpha, 0}-U_{\alpha, 0_{h}} \tag{5.24}
\end{equation*}
$$

Now that we have presented the direct problem, we need to study the dual problem in order to establish an a posteriori error estimate.

### 5.3 Dual problem

Since we want to focus on the error that arises in computing the grating efficiency, we have chosen a goal oriented error estimate which is the DWR method. Hence, we first need to define our quantity of interest. We will show that this quantity of interest is a continuous linear functional, which is necessary for the dual problem to be well posed. We then formulate the dual problem and use it to study the error in the quantity of interest.

### 5.3.1 Quantity of interest: Q

Let $f \in H_{\#}^{\frac{1}{2}}([0, d])$, we define the map $Q$ where $Q(f) \in \mathbb{C}$ and

$$
\begin{equation*}
Q(f)=\frac{1}{q} \sum_{\left|n_{\alpha}\right|<|k|} c_{n} f^{(n)}( \pm B), \tag{5.25}
\end{equation*}
$$

such that $c_{n}=0$ or $c_{n}=1, d$ corresponds to the period of the grating as shown in Figure 2.3 and $\Gamma_{ \pm}$are defined in equations (2.37) and (2.38). The Fourier coefficient $f^{(n)}( \pm B)$ of $f(x, \pm B)$ is given by

$$
f^{(n)}( \pm B)=\frac{1}{d} \int_{0}^{d} f(x, \pm B) e^{-i \frac{2 \pi n}{d} x} d x
$$

From equations (4.49), (4.50) and Lemma 55, we have

$$
\begin{aligned}
\sum_{\left|n_{\alpha}\right|<|k|}\left|r_{1}^{n}\right|+\left|t_{2}^{n}\right|= & \left|U_{\alpha, 0}^{(0)}(B)-e^{-i \beta_{1}^{0} B}\right| \\
& +\sum_{\left|n_{\alpha}\right|<\left|k_{1}\right|, n \neq 0}\left|U_{\alpha, 0}^{(n)}(B)\right|+\sum_{\left|n_{\alpha}\right|<\left|k_{2}\right|}\left|U_{\alpha, 0}^{(n)}(-B)\right|, \\
\leq & \sup _{n} e^{\Im\left(\beta \beta_{2}^{n}\right)} \sum_{\left|n_{\alpha}\right|<|k|}\left|U_{\alpha, 0}^{(n)}( \pm B)\right|+1, \\
\leq & 1+q e^{\Im\left(k_{2}\right)} Q\left(U_{\alpha, 0}( \pm B)\right)
\end{aligned}
$$

since $\mathfrak{J}\left(\beta_{2}^{n}\right) \leq \mathfrak{J}\left(k_{2}\right)$, when $\left|k_{2}\right|>\left|n_{\alpha}\right|$. Hence, the quantity of interest is chosen so that it is related to the computation of the efficiencies corresponding to the propagating waves, and for that reason we have the condition $\left|n_{\alpha}\right|<|k|$. In addition, the constants $c_{n}$ have been introduced so that we can choose which particular efficiency order $m$ to focus on. Hence we choose

$$
c_{n}=\delta_{m n}
$$

with $\delta_{m n}$ the Kronecker delta (given by equation (2.56)). We will show in the following result that $Q$ is a continuous linear functional.

Lemma 63. Let $f \in H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right), k_{\text {ref }}>0$ such that $|k|>k_{\text {ref }}$ and let $Q$ as given by equation (5.25) then $Q$ is a linear continuous functional and we have

$$
|Q(f)| \leq \frac{1}{k_{r e f}^{2}} \ln \left(|k|-\alpha+\sqrt{1+(|k|-\alpha)^{2}}\right)^{1 / 2} \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}
$$

Proof. The linearity of $Q$ follows from its definition. We show the continuity by using Schwarz's inequality. We have

$$
\begin{aligned}
|Q(f)|= & \left|\sum_{\left|n_{\alpha}\right|<|k|} \frac{1}{q} f^{(n)}( \pm B)\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{1 / 4}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{-1 / 4}\right| \\
& \leq \frac{1}{k_{r e f}^{2}}\left(\sum_{\left|n_{\alpha}\right|<|k|}\left|f^{(n)}( \pm B)\right|^{2}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{1 / 2}\right)^{1 / 2}\left(\sum_{\left|n_{\alpha}\right|<|k|} \frac{1}{\sqrt{1+\left(\frac{2 \pi n}{d}\right)^{2}}}\right)^{1 / 2}
\end{aligned}
$$

By noting that

$$
\sum_{\left|n_{\alpha}\right|<|k|} \frac{1}{\sqrt{1+\left(\frac{2 \pi n}{d}\right)^{2}}} \leq \int_{\alpha}^{|k|} \frac{1}{\sqrt{1+\left(n_{\alpha}-\alpha\right)^{2}}} d n_{\alpha}
$$

We can make the following change of variable $x=n_{\alpha}-\alpha$ and $d x=d n_{\alpha}$, and so

$$
\sum_{\left|n_{\alpha}\right|<|k|} \frac{1}{\sqrt{1+\left(\frac{2 \pi n}{d}\right)^{2}}} \leq \int_{0}^{|k|-\alpha} \not \frac{1}{1+x^{2}} d x .
$$

We can write

$$
\begin{aligned}
\sum_{\left|n_{\alpha}\right|<|k|} \frac{1}{\sqrt{1+\left(\frac{2 \pi n}{d}\right)^{2}}} & =\left[\ln \left(x+\sqrt{ } \frac{\left.\overline{1+x^{2}}\right)}{}\right)\right]_{0}^{|k|-\alpha} \\
& =\ln \left(|k|-\alpha+\sqrt{1+(|k|-\alpha)^{2}}\right)
\end{aligned}
$$

We then have

$$
\begin{aligned}
|Q(f)| \leq & \frac{1}{k_{\text {ref }}^{2}} \ln \left(|k|-\alpha+\sqrt{1+(|k|-\alpha)^{2}}\right)^{1 / 2} \\
& \times\left(\sum_{n_{\alpha} \in \mathbb{Z}}\left|f^{(n)}( \pm B)\right|^{2}\left(1+\left(\frac{2 \pi n}{d}\right)^{2}\right)^{1 / 2}\right)^{1 / 2}, \\
\leq & \frac{1}{k_{\text {ref }}^{2}} \ln \left(|k|-\alpha+\sqrt{1+(|k|-\alpha)^{2}}\right)^{1 / 2} \mathrm{k} f \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}
\end{aligned}
$$

from Definition A-14 which finishes the proof.
We now can proceed to the study of the dual problem.

### 5.3.2 Strong form of the dual problem

In order to estimate $Q\left(U_{\alpha, 0}\right)-Q\left(U_{\alpha, 0_{h}}\right)=Q\left(e_{h}\right)$, we introduce a dual function $z \in H_{\#}^{2}(F)$ which satisfies

$$
\begin{equation*}
(A v, z)=a(v, z)=Q(v) \quad \forall v \in H_{\#}^{1}(F) \tag{5.26}
\end{equation*}
$$

where $a$ is given by equation (5.18) and $A$ by equation (5.2) or (5.3). Hence, the strong form of the dual problem given by equation (5.26) is given by the following lemma.

Lemma 64. Let $z \in H_{\#}^{2}(F)$ be the dual solution corresponding to the dual problem of (5.18) then $z$ satisfies

$$
\begin{cases}-q \nabla \cdot\left(\frac{1}{q} \nabla\right) z+2 i \alpha \partial_{x} z-\left(k^{2}-\alpha^{2}\right) z+\partial_{n} z-T_{ \pm}^{* \alpha, 0} z+i \alpha n_{x} z & \text { for Case 1B }  \tag{5.27}\\ -q \nabla \cdot\left(\frac{1}{q} \nabla\right) z+2 i \alpha \partial_{x} z-\left(k^{2}-\alpha^{2}\right) z+\partial_{n} z-T_{ \pm}^{* \alpha, 0} z & \text { otherwise }\end{cases}
$$

The functional $J_{ \pm}$is defined as follows

$$
\begin{equation*}
J_{ \pm}=\sum_{\left|n_{\alpha}\right|<\left|k_{j}\right|} \frac{c_{n}}{d} e^{-i \frac{2 \pi n}{d} x}, \tag{5.28}
\end{equation*}
$$

with $j=1,2$ respectively for $J_{+}\left(\right.$on $\left.\Gamma_{+}\right)$and $J_{-}\left(\right.$on $\left.\Gamma_{-}\right)$, and $T_{ \pm}^{* \alpha, 0}$ is the dual of $T_{ \pm}^{\alpha, 0}[61, p .476]$.
Proof. By using the divergence theorem, we can integrate by parts $a\left(U_{\alpha, 0}, v\right)$ given by equation (5.18) and we get

$$
\left(\frac{1}{q} \nabla v, \nabla z\right)_{F}=\left(\nabla v, \frac{1}{q} \nabla z\right)_{F}=\left(v,-\nabla \cdot\left(\frac{1}{q} \nabla\right) z\right)_{F}+\left(v, \frac{1}{q} \partial_{n} z\right)_{\partial F},
$$

and

$$
\int_{F}\left(\partial_{x_{i}} v\right) z=-\int_{F} v\left(\partial_{x_{i}} z\right)+\int_{\partial F}(v z) n_{i} d s,
$$

such that $n_{i}$ is the outward unit normal in the direction of $x_{i}$. This leads us to

$$
\int_{F}\left(\partial_{x} v\right) \bar{z}=-\int_{F} v \partial_{x} \bar{z},
$$

from equation (5.21). We then have for Case 1A and Case 2

$$
\begin{align*}
a(v, z)= & \left(v,-\nabla \cdot\left(\frac{1}{q} \nabla z\right)\right)_{F}+2 i \alpha\left(\frac{1}{q} v, \partial_{x} z\right)_{F}+\left(\frac{1}{q} v, \partial_{n} z\right)_{\partial F} \\
& -\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) v, z\right)_{F}-\left(\frac{1}{q} v, T^{* \alpha, 0} z\right)_{\Gamma_{ \pm}}=Q(v), \tag{5.29}
\end{align*}
$$

for all $v \in H_{\#}^{1} F$. Let $\phi \in L_{\#}^{2}(\Gamma)$, we also have

$$
\begin{align*}
T_{ \pm}^{* \alpha, 0} \phi(x) & =\sum_{n \in \mathbb{Z}} i \beta_{j}^{n} \bar{\phi}^{(n)}( \pm B) e^{-i \frac{2 \pi n}{d} x},  \tag{5.30}\\
\bar{\phi}^{(n)}(y) & =\frac{1}{d} \int_{0}^{d} \overline{\phi(x, y)} e^{i \frac{2 \pi n}{d} x} d x . \tag{5.31}
\end{align*}
$$

We note that

$$
\begin{equation*}
Q(v)=\frac{1}{q}\left(J_{ \pm}, v\right), \tag{5.32}
\end{equation*}
$$

such that $J_{ \pm}$is given by (5.28). For Case 1B, we note that $\left(v n_{x}, z\right)_{\partial F}=\left(v, n_{x} z\right)_{\partial F}$ which finishes the proof.

### 5.3.3 Variational form

Since we do not know the exact dual solution $z$, we will need this weak formulation later in Chapter 7 to approximate $z$ by either using a finer mesh or by raising the degree of the polynomial basis in our finite element implementation. The variational formulation of Lemma 64 is given below.
Lemma 65. Let $z \in H_{\#}^{1}(F)$ (weak solution) then, for any $\psi \in H_{\#}^{1}(F)$, we have for Case 1A, Case 2A and Case 2B

$$
\begin{aligned}
a^{*}(\psi, z)= & \left(\frac{1}{q} \nabla \psi, \nabla z\right)_{F}+2 i \alpha\left(\frac{1}{q} \psi, \partial_{x} z\right)_{F}-\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) \psi, z\right)_{F} \\
& -\left(\frac{1}{q} \psi, T_{ \pm}^{* \alpha, 0} z\right)_{\Gamma_{ \pm}}=\left(\frac{1}{q} J_{ \pm}, \psi\right)_{F}
\end{aligned}
$$

and for Case 1B

$$
\begin{aligned}
a^{*}(\psi, z)= & \left(\frac{1}{q} \nabla \psi, \nabla z\right)_{F}+2 i \alpha\left(\frac{1}{q} \psi, \partial_{x} z\right)_{F}-\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) \psi, z\right)_{F} \\
& +\frac{i \alpha}{q}\left(\psi, z n_{x}\right)_{\partial F}-\left(\frac{1}{q} \psi, T_{ \pm}^{* \alpha, 0} z\right)_{\Gamma_{ \pm}}=\left(\frac{1}{q} J_{ \pm}, \psi\right)_{F}
\end{aligned}
$$

Proof. Let $\psi \in H_{\#}^{1}(F)$, then we have from equation (5.27) for Case 1A , Case 2A and Case 2B

$$
\begin{gathered}
\left(\psi,-\nabla \cdot\left(\frac{1}{q} \nabla z\right)\right)_{F}+2 i \alpha\left(\frac{1}{q} \psi, \partial_{x} z\right)_{F}-\left(\frac{1}{q}\left(k^{2}-\alpha^{2}\right) \psi, z\right)_{F} \\
+\left(\frac{1}{q} \psi, \partial_{n} z\right)_{F}-\left(\frac{1}{q} \psi, T_{ \pm}^{* \alpha, 0} z\right)_{\Gamma_{ \pm}}=\left(\frac{1}{q} \psi, J_{ \pm}\right)_{F} .
\end{gathered}
$$

We note that

$$
\left(\psi,-\nabla \cdot\left(\frac{1}{q} \nabla\right) z+\partial_{n} z\right)_{F}=\left(\frac{1}{q} \nabla \psi, \nabla z\right)_{F} .
$$

In a similar way, we can prove the weak form for Case 1B .
Now that we have formulated the direct and the dual problem, we now use them to establish a goal oriented error estimate, using the DWR method, where the target is related to the grating efficiency.

### 5.4 A posteriori error estimation

### 5.4.1 Continuous problem

Let $z$ be the dual solution associated with the dual problem given by Lemma 64 . We then have for any $\phi \in H_{\#}^{1}(F)$ that

$$
\begin{equation*}
a(\phi, z)=\left(\phi, A^{*} z\right)_{F}=\left(\frac{1}{q} J_{ \pm}, z\right)_{F} \tag{5.33}
\end{equation*}
$$

such that

$$
A^{*}= \begin{cases}-\nabla \cdot\left(\frac{1}{q} \nabla\right)+\frac{2 i \alpha}{q} \partial_{x}-\frac{1}{q}\left(k^{2}-\alpha^{2}\right)+\frac{1}{q} \partial_{n}-\frac{1}{q} T_{ \pm}^{* \alpha, 0}+i \alpha n_{x} & \text { for Case 1B }  \tag{5.34}\\ -\nabla \cdot\left(\frac{1}{q} \nabla\right)+\frac{2 i \alpha}{q} \partial_{x}-\frac{1}{q}\left(k^{2}-\alpha^{2}\right)+\frac{1}{q} \partial_{n}-\frac{1}{q} T_{ \pm}^{* \alpha, 0} & \text { otherwise }\end{cases}
$$

where $A^{*}$ is the dual operator of $A$ defined in equation (5.2) or (5.3), [61, p. 476]. Using equation (5.32) leads us to the following problem: Find $z \in H_{\#}^{1}(F)$ such that

$$
\begin{equation*}
a^{*}(\phi, z)=\left(\phi, A^{*} z\right)_{F}=Q(\phi), \tag{5.35}
\end{equation*}
$$

for any $\phi \in H_{\#}^{1}(F)$. We now use the finite element method to discretise the problem.

### 5.4.2 Discretised problem

Let $X^{\alpha}$ be a finite element subspace of order $p$ of $H_{\#}^{1}(F)$ as described in Section B. 3 and let $\zeta_{h}$ be any regular partition of $X^{\alpha}$. We denote by $h$ the maximum mesh size of the triangular elements in this partition. The finite element approximation of the dual problem given by equation (5.35) is to find $z_{h} \in X^{\alpha}$ such that

$$
\begin{equation*}
a\left(\phi_{h}, z_{h}\right)=\left(\phi_{h}, A^{*} z_{h}\right)_{F}=Q\left(\phi_{h}\right), \tag{5.36}
\end{equation*}
$$

for any $\phi_{h} \in X^{\alpha}$. We are now going to look for an upper bound on $Q\left(e_{h}\right)$.

### 5.4.3 Error estimation

The estimate of the linear functional of the error $Q\left(e_{h}\right)$ is given by the following theorem.

Theorem 66. Let $U_{\alpha, 0_{h}}$ be the solution to equation (5.22), $e_{h}$ be given by equation (5.24), $\zeta_{h}=\{K\}$ be a partition of $X^{\alpha}$, and $p_{K}$ and $h_{K}$ be the polynomial order and the mesh size associated with the element $K$. Let us also denote the field equation residual $R_{h}(K)$ x, we then have

$$
R_{h}(K):= \begin{cases}\nabla_{\alpha}\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0_{h}}\right)+U_{\alpha, 0_{h}} & \text { for Case 2B }  \tag{5.37}\\ \nabla \cdot\left(\frac{1}{q} \nabla U_{\alpha, 0_{h}}\right)-\frac{2 i \alpha}{q} \partial_{x} U_{\alpha, 0_{h}}-\frac{1}{q}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0_{h}}, & \text { otherwise }\end{cases}
$$

and the flux residual $r_{h}(E)$ for Case 2 is given by

$$
r_{h}(E):=\left\{\begin{array}{lll}
-\frac{1}{2 q}\left[\partial_{n} U_{\alpha, 0_{h}}\right] & \text { if } & E \subset \partial K \backslash \Gamma_{ \pm},  \tag{5.38}\\
\frac{1}{q} T_{+}^{\alpha, 0} U_{\alpha, 0_{h}}-\frac{1}{q} 2 i e^{-i \beta_{1}^{0} B} & \text { if } & E \subset \Gamma_{+}, \\
\frac{1}{q} T_{-}^{\alpha, 0} U_{\alpha, 0_{h}} & \text { if } & E \subset \Gamma_{-},
\end{array}\right.
$$

and for Case 1

$$
r_{h}(E):=\left\{\begin{array}{lll}
-\frac{1}{2 q}\left[\partial_{n} U_{\alpha, 0_{h}}\right] & \text { if } & E \subset \partial K \backslash\left(\Gamma_{ \pm} \cup \partial \Omega_{3}\right),  \tag{5.39}\\
\frac{1}{q} T_{+}^{\alpha, 0} U_{\alpha, 0_{h}}-\frac{1}{q} 2 i e^{-i \beta_{1}^{0} B} & \text { if } & E \subset \Gamma_{+}, \\
\frac{1}{q} T_{-}^{\alpha, 0} U_{\alpha, 0_{h}} & \text { if } & E \subset \Gamma_{-},
\end{array}\right.
$$

with $\left[\partial_{n} U_{\alpha, 0_{h}}\right]$ denoting the jump of the normal derivatives of $U_{\alpha, 0_{h}}, E$ an edge of the element $K$ and $q$ as defined by equation (5.19). We then have

$$
\begin{equation*}
\left|Q\left(e_{h}\right)\right| \leq \sum_{K \in \zeta_{h}} \rho_{K} w_{k} \tag{5.40}
\end{equation*}
$$

such that the cell residuals $\rho_{k}$ and weights $w_{K}$ are given by

$$
\begin{align*}
\rho_{K} & :=\mathrm{k} R_{h}(K) \mathrm{k}_{L_{\alpha \#}^{2}(K)}+\left(h_{K}\right)^{-1 / 2} \mathrm{k} r_{h}(E) \mathrm{k}_{L_{\alpha \#}^{2}(E)}^{2},  \tag{5.41}\\
w_{K} & :=\mathrm{k} z-z_{h} \mathrm{k}_{L_{\alpha \#}^{2}(K)}+\left(h_{K}\right)^{1 / 2} \mathrm{k} z-z_{h} \mathrm{k}_{L_{\alpha \#}^{2}(E)}^{2} . \tag{5.42}
\end{align*}
$$

Proof. From equations (5.26) and (5.22), we have

$$
\begin{equation*}
Q\left(U_{\alpha, 0}\right)-Q\left(U_{\alpha, 0_{h}}\right)=a\left(U_{\alpha, 0}-U_{\alpha, 0_{h}}, z\right), \tag{5.43}
\end{equation*}
$$

and

$$
Q\left(e_{h}\right)=(f, z)_{\Gamma_{+}}-a\left(U_{\alpha, 0_{h}}, z\right)=(f, z)_{\Gamma_{+}}-\left(A U_{\alpha, 0_{h}}, z\right)
$$

using equations (5.17) and (5.26). Let $\zeta$ be a partition of the domain $F$ into mesh cells $K$, from cell wise integration by parts [39, p. 28], [117, p. 12] and by using Galerkin orthogonality similar to equation (4.88), we have from equations (5.44), (5.2) for Case 1A , Case 1B and Case 2A

$$
\begin{array}{r}
Q\left(e_{h}\right)=\quad \sum_{K}\left(\Delta U_{\alpha, 0_{h}}+2 i \alpha \partial_{x} U_{\alpha, 0_{h}}\right. \\
\left.+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0_{h}}, z-\phi\right)_{K}  \tag{5.44}\\
\\
-\left(\nabla U_{\alpha, 0_{h}} \cdot n+f, z-\phi\right)_{\partial K}
\end{array}
$$

for all $\phi \in X^{\alpha}$ and $\bigcup_{K \in \zeta}=F$. We can use the similar argument for Case 2B to get

$$
\begin{array}{r}
Q\left(e_{h}\right)=\quad \sum_{K}\left(\nabla_{\alpha}\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0_{h}}\right)+U_{\alpha, 0_{h}}, z-\phi\right)_{K} \\
-\left(\nabla U_{\alpha, 0_{h}} \cdot n+f, z-\phi\right)_{\partial K} \tag{5.45}
\end{array}
$$

Hence, using equations (5.4), (5.5) and in addition equation (5.6) for Case 1 we have equations (5.37) and (5.38) or (5.39), then

$$
\left|Q\left(e_{h}\right)\right| \leq \sum_{K}\left|\left(R_{h}(K), z-\phi\right)_{K}+\left(r_{h}(E), z-\phi\right)_{\partial K}\right| .
$$

From the Cauchy-Schwarz inequality [22, p. 50]

$$
\begin{aligned}
& \left|Q\left(e_{h}\right)\right| \\
& \leq \sum_{K} \mathrm{k} R_{h}(K) \mathrm{k}_{L_{\alpha \#}^{2}(K)} \mathrm{k} z-\phi \mathrm{k}_{L_{\alpha \#}^{2}(K)}+\mathrm{k} r_{h}(E) \mathrm{k}_{L_{\alpha \# \#}^{2}(E)} \mathrm{k} z-\phi \mathrm{k}_{L_{\alpha \#}^{2}(E)} \\
& \leq \sum_{K}\left(\mathrm{k} R_{h}(K) \mathrm{k}_{L_{\alpha \#}^{2}(K)}+h_{K}^{-1 / 2} \mathrm{k} r_{h}(E) \mathrm{k}_{L_{\alpha \#}^{2}(E)}\right) \\
& \quad \times\left(\mathrm{k} z-\phi \mathrm{k}_{L_{\alpha \# \#}^{2}(K)}+h_{K}^{1 / 2} \mathrm{k} z-\phi \mathrm{k}_{L_{\alpha \#}^{2}(E)}\right) \\
& =\sum_{K} \rho_{K} w_{K}
\end{aligned}
$$

we finish the proof of the error estimation by choosing $\phi=z_{h}$ and by using equations (5.41) and (5.42).

### 5.5 Summary and conclusion

### 5.5.1 Summary

In this chapter, we have introduced a basic framework for studying a posteriori error estimates. We showed that there are two ways of establishing an a posteriori error estimate. There is the global error estimate and the goal oriented error estimate. Since our interest is driven by the diffraction efficiency, we chose the latter and in particular we used the Dual Weighted Residual (DWR) method. In order to do so, we generalised the continuous problem to cover all four cases. We introduced our quantity of interest, $Q$ and showed that it is a linear continuous functional. This allowed us to formulate the dual problem. We then combined the dual and the direct problems to establish an upper bound for the error estimate in Theorem 66. The evaluation of the error in the functional $Q$ represents the primary output that we desire from our model, namely the diffraction efficiency.

### 5.5.2 Conclusion

In Chapter 4, the a priori error estimate used the duality argument to represent the error between the exact and approximate solution in terms of the exact solution. By combining the Galerkin orthogonality with the regularity estimate of the exact solution, we found an upper bound for this error (see Theorems 62, D-36, D-37 and D-38). Importantly, these upper bounds depend on the exact solution, and as this is not normally known, we cannot compute them. The goal in deriving a priori error estimates is therefore to guarantee that the discrete solution will converge to the exact solution provided that we keep the interpolation error small with respect to the wavenumber $k$. In contrast, for the a posteriori error estimate, the upper bounds given by equation (5.40) can be evaluated. We find the discrete dual
problem $z_{h}$ in the same way that we find the discrete direct solution $U_{\alpha, 0_{h}}$ using Lemma 65 . For the exact solution of the dual problem $z$, since we do not know it analytically, we can approximate it either by solving the dual problem numerically in a very fine mesh or by increasing the polynomial order using the same mesh as $z_{h}$. Hence we can estimate the error in the targeted quantity $Q\left(e_{h}\right)$ from the local contribution of each error indicator $\rho_{K}$ and $w_{K}$ as defined in equations (5.41) and (5.42). In fact, these error indicators are the cell residuals $\rho_{K}$ multiplied by the weights $w_{K}$ taken from the computed solution. The $\rho_{K}$ in turn consist of the field equation residual $\left(R_{h}(K)\right)$ and the flux residual $\left(r_{h}(E)\right)$ which indicate the smoothness of the discrete solution. The weights $w_{K}$ capture the influence of the cell residuals on the targeted error $Q\left(e_{h}\right)$ since if we differentiate $Q\left(e_{h}\right)$ with respect to $\rho_{K}$, we are left with $w_{K}$.

We will show in Chapter 7 that this error estimate will allow us to perform an automatic mesh adaptation based on the local error indicators $\rho_{K}$ and $w_{K}$. The dual weighted residual method uses these error indicators to maximize the accuracy of the computed diffraction efficiency by choosing a tolerance ( $T O L$ ) and demanding that $\sum_{K} \rho_{K} w_{K}<T O L$ as we solve our problem. It will allow us to optimise the computational efficiency. Rather than having a uniform mesh, where we refine all elements of the mesh at each step, we will just refine where the error indicators are large and keep the coarse mesh where the error indicators are small.

We investigated both the a priori error estimate and the a posteriori error estimate using the $\alpha-0$ quasi periodic method in Chapters 4 and 5 . In the following chapter, we will extend the $\alpha, 0$-quasi periodic transformation and introduce a function $w(y)$ which depends on the wavenumber $k$ and use the change of variable $U=e^{i \alpha x} w(y) U_{\alpha, \beta}$. Hence, $U_{\alpha, \beta}$ is still periodic and we will investigate the a priori error estimate for $U_{\alpha, \beta}$.

## Chapter 6

# New formulation of the grating problem and a priori error estimates using the $\alpha, \beta$-quasi periodic transformation 

In this chapter, we will introduce another approach to solve the diffraction problems described in Chapters 2 and 3 which is an extension of the $\alpha, 0$-quasi periodic transformation in Chapter 4. When a series of homogeneous layers are constructed (that is, $\Omega_{3}$ as shown in Figure 2.2 is not present), the analytic solution of the associated diffraction problem can be written in the form $e^{i \alpha x} w(y)$ where Snell' s law is used to calculate some $w(y)$. This suggests that writing $U=e^{i \alpha x} w(y) U_{\alpha, \beta}(x, y)$ and solving our diffraction problem for $U_{\alpha, \beta}$ might improve the a priori error estimate. Indeed, when we solve the Helmholtz problem for homogeneous layers, and we use this transformation, $U_{\alpha, \beta}$ is just a constant. The idea is to remove from $U$ the oscillations $w(y) e^{i \alpha x}$ and then investigate if this new transformation will improve the dependence on the wavenumber $k$ in the a priori error estimate. Since $U$ is $\alpha$-quasi periodic then $U_{\alpha, \beta}$ is periodic with respect to $x$ and we will see that $w$ depends on $\beta$. For these reasons we are going to call this new transformation the $\alpha, \beta$-quasi periodic transformation. Hence, the numerical implementation of the finite element method should be computationally less expensive and less complicated if we base it on $U_{\alpha, \beta}$.

In keeping the thesis to a manageable size and since we have studied in detail the four cases in Chapter 4 and Appendix D, we are going to focus, in this chapter, on Case 1A (perfectly conducting grating with a homogeneous region outside the scatterer (that is $\left.k_{0}=k_{1}=k_{2}\right)$ ). The other cases can be derived in a similar fashion. We start by deriving the differential equation satisfied by $U_{\alpha, \beta}$. We then examine the variational formulation corresponding to the continuous problem. Next, we show that this problem is well posed and derive a formula to compute the efficiency of the diffraction grating from $U_{\alpha, \beta}$. We then use the finite element
method to discretise the problem, and provide a rigorous study of the a priori error estimate. In order to do so, we first derive regularity results for the scattering problem in the periodic space $H_{\#}^{l}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ for $l \geq 1$. This then provides error estimates that give an explicit dependence on the wavenumber. This error estimate provides an insight into the convergence of the solution and will help in our numerical implementation to balance the accuracy against the computational cost. The Chapter concludes with a comparison between this transformation and that examined in Chapter 4.

### 6.1 Restatement of the boundary value problem for the periodic solution

Similar to the analysis in Section 4.1, we seek a periodic function $U_{\alpha, \beta}$ such that

$$
\begin{equation*}
U(x, y)=e^{i \alpha x} w(y) U_{\alpha, \beta}(x, y) \tag{6.1}
\end{equation*}
$$

where $w(y)$ is the analytical solution of equation (2.95) given by Snell's law when $\Omega_{0}$ is homogeneous (that is, $\Omega_{3}$ is not present) in Figure 2.2 and $k_{0}=k_{1}=k_{2}$. This can be written as

$$
\begin{equation*}
w(y)=e^{-i \beta_{1}^{0} y}, \quad-B \leq y \leq B, \tag{6.2}
\end{equation*}
$$

with $\beta_{1}^{0}=k_{1} \cos \theta$ where $k_{1} \in \mathbb{R}$ is the wavenumber, and $\theta$ is the angle of incidence. Therefore the propagating equation is changed, and we have the following lemma.

Lemma 67. Suppose $U_{\alpha, \beta}$ satisfies

$$
-2 i \beta_{1}^{0} \partial_{y} U_{\alpha, \beta}+2 i \alpha \partial
$$

Hence, we can rewrite equation (6.9) to get
$e^{i \alpha x}\left(-i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} U_{\alpha, \beta}(B)+\left.e^{-i \beta_{1}^{0} B} \partial_{n} U_{\alpha, \beta}\right|_{\Gamma_{+}}\right)=\left(e^{-i \beta_{1}^{0} B} T_{+}^{\alpha, 0} U_{\alpha, \beta}-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}\right) e^{i \alpha x}$.
That is

$$
\begin{equation*}
\left.\partial_{n} U_{\alpha, \beta}\right|_{\Gamma_{+}}=T_{+}^{\alpha, 0} U_{\alpha, \beta}-\left(-i \beta_{1}^{0}\right) U_{\alpha, \beta}(x, B)-2 i \beta_{1}^{0} . \tag{6.10}
\end{equation*}
$$

Since $U_{\alpha, \beta}$ is periodic with respect to $x$ then we can write

$$
U_{\alpha, \beta}(x, B)=\sum_{n \in \mathbb{Z}} U_{\alpha, \beta}^{(n)}(B) e^{i \frac{2 \pi n}{d} x} .
$$

Using Definition 4 then equation (6.10) becomes (since this boundary is in $\Omega_{1}$ )

$$
\left.\partial_{n} U_{\alpha, \beta}\right|_{y=B}=\sum_{n \in \mathbb{Z}}\left(i \beta_{1}^{n}+i \beta_{1}^{0}\right) U_{\alpha, \beta}^{(n)}(B) e^{i \frac{2 \pi n}{d} x}-2 i \beta_{1}^{0} .
$$

Using Definition 5 finishes the proof for the boundary condition on $\Gamma_{+}$. On the bottom boundary, when $y=-B$, then a similar argument shows

$$
\left.\partial_{n} U_{\alpha, \beta}\right|_{y=-B}=\sum_{n \in \mathbb{Z}}\left(i \beta_{2}^{n}-i \beta_{1}^{0}\right) U_{\alpha, \beta}^{(n)}(-B) e^{i \frac{2 \pi n}{d} x} .
$$

### 6.1.1 Variational formulation

Let $v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ then we can multiply equation (6.3) by $\bar{v}$ and integrate to get

$$
\begin{gather*}
\int_{\Omega \backslash \text { int } \Omega_{3}} \Delta U_{\alpha, \beta} \bar{v}+2 i \alpha \int_{\Omega \backslash \text { int } \Omega_{3}}\left(\partial_{x} U_{\alpha, \beta}\right) \bar{v} \\
+2 \int_{\Omega \backslash \text { int } \Omega_{3}}\left(-i \beta_{1}^{0} \partial_{y} U_{\alpha, \beta}\right) \bar{v}=0 . \tag{6.11}
\end{gather*}
$$

We can then integrate by parts where we note that
since the normal derivative of $\Gamma$

### 6.2 Discrete problem

In order to solve numerically the scattering problem, we need to discretise the variational formulation corresponding to the continuous problem.

### 6.2.1 Variational formulation

We want to approximate the continuous problem associated with equation (6.13) which is given by equation (6.14). Let $X^{\alpha, \beta}$ be a finite-dimensional subspace of $H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, with $\operatorname{dim}\left(X^{\alpha, \beta}\right)=N<\infty$ and let $\psi_{i}$ for $i=1, . ., N$, be a basis of $X^{\alpha, \beta}$. We discretise the variational form given by the equation (6.14) and this leads us to find $U_{\alpha, \beta_{h}} \in X^{\alpha, \beta}$ for all $v_{h} \in X^{\alpha, \beta}$ such that

$$
\begin{equation*}
a\left(U_{\alpha, \beta_{h}}, v_{h}\right)=\left(f_{\alpha, \beta}, v_{h \alpha, \beta}\right), \tag{6.34}
\end{equation*}
$$

and subject to the constraint

$$
U_{\alpha, \beta_{h}}=0,(x, y) \in \partial \Omega_{3}
$$

with

$$
\begin{align*}
a\left(s_{h}, v_{h}\right)= & \left(\nabla s_{h}, \nabla v_{h}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-2 i \alpha\left(\partial_{x} s_{h}, v_{h}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} \\
& +\left(2 i \beta_{1}^{0} \partial_{y} s_{h}, v_{h}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}^{\alpha, \beta} s_{h}, v_{h}\right)_{\Gamma_{ \pm}},  \tag{6.35}\\
\left(f_{\alpha, \beta}, v_{h}\right)_{\Gamma_{+}}= & \left(-2 i \beta_{1}^{0}, v_{h}\right)_{\Gamma_{+}}, \tag{6.36}
\end{align*}
$$

for all $s_{h} \in X^{\alpha, \beta}$ such that $T_{ \pm}^{\alpha, \beta}$ is as defined in Definition 5. Since $U_{\alpha, \beta_{h}} \in X^{\alpha, \beta}$, there exists $U_{j}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, \beta_{h}}=\sum_{j=1}^{N} U_{j} \psi_{j}(x, y)$. Hence, the discrete problem given by equation (6.34) is equivalent to the following linear algebraic system

$$
\begin{equation*}
A U=L \tag{6.37}
\end{equation*}
$$

with $U=U_{j}$ for $j=, \cdots, N$,

$$
A=a\left(\psi_{i}, \psi_{j}\right)
$$

and

$$
L=\left(f_{\alpha, \beta}, \psi_{j}\right)_{\Gamma_{+}}
$$

for $\{i, j\} \in\{1, . ., N\}$.

### 6.2.2 Truncation of the DtN map

Similar to Section 4.2.2, for computational purposes, we need to truncate the infinite sum inside the DtN map that we use as transparent boundary conditions. Let $M \in \mathbb{N}$ and $M<\infty$. From equation (2.66), $T_{ \pm}^{\alpha, \beta}$ is approximated by $T_{ \pm}^{\alpha, \beta^{M}}$ where

$$
\begin{equation*}
T_{ \pm}^{\alpha, \beta^{M}}=T_{ \pm}^{\alpha, 0^{M}} \pm i \beta_{1}^{0} I_{d} \tag{6.38}
\end{equation*}
$$

with $T_{ \pm}^{\alpha, 0^{M}}$ as given by equation (4.61). Similar to the derivation of equation (4.62), $U_{\alpha, \beta}$ is approximated with a truncated Fourier series

$$
\begin{equation*}
U_{\alpha, \beta_{h}}(x, y)=\sum_{n=-M}^{M} U_{\alpha, \beta_{h}}^{(n)}(y) e^{i \frac{2 \pi n}{d} x}, \tag{6.39}
\end{equation*}
$$

where

$$
U_{\alpha, \beta_{h}}^{(n)}(y)=\frac{1}{d} \int_{0}^{d} U_{\alpha, \beta_{h}}(x, y) e^{-i \frac{2 \pi n}{d} x} d x .
$$

By truncating the fundamental solution in equation (2.55), and using equation (6.1), we can derive a similar equation to equation (4.63)

$$
\begin{equation*}
e^{-i \beta_{1}^{0} y} U_{\alpha, \beta_{h}}=\sum_{n=-M}^{M} r_{j}^{n, M} e^{i \beta_{j}^{n} y+i \frac{2 \pi n}{d} x}+t_{j}^{n, M} e^{-i \beta_{j}^{n} y+i \frac{2 \pi n}{d} x}, j=1,2, \tag{6.40}
\end{equation*}
$$

where the unknowns $r_{j}^{n, M}$ and $t_{j}^{n, M}$ are complex scalars. $U_{\alpha, 0_{h}}, U_{\alpha, \beta_{h}}$ contain the incident wave and must satisfy the UPRC condition (see equation (2.26)) in $\Omega_{1}$ (see Figure 2.3). It follows that $t_{1}^{n, M}=0$, for $n \sigma 0$ and $t_{1}^{0, M}=1$ in $\Omega_{1}$ and in $\Omega_{2}$, all the $r_{2}^{n, M}=0$. We proceed exactly as we have done for the continuous problem but the solution is now truncated as given by equations (6.39) and (6.40). The coefficients $\beta_{j}^{n}$ are given by equation (2.43). From Lemma 71 and equation (4.64), we can derive

$$
\begin{array}{ll}
e^{-i \beta_{1}^{0} y} U_{\alpha, \beta h}^{(n)}(y)=r_{1}^{n, M} e^{i \beta_{1}^{n} y}+\delta_{n 0} e^{-i \beta_{1}^{n} y}, & \text { in } \Omega_{1},  \tag{6.41}\\
e^{-i \beta_{1}^{0} y} U_{\alpha, \beta_{h}}^{(n)}(y)=t_{2}^{n, M} e^{-i \beta_{2}^{n} y}, & \text { in } \Omega_{2} .
\end{array}
$$

At $y= \pm B$, equations (2.58) and (2.59) become

$$
\begin{align*}
r_{1}^{n, M} & =e^{-i \beta_{1}^{0} B} U_{\alpha, \beta_{h}}^{(n)}(B) e^{-i \beta_{1}^{n} B}-\delta_{n 0} e^{-2 i \beta_{1}^{0} B},  \tag{6.42}\\
t_{2}^{n, M} & =e^{i \beta_{1}^{0} B} U_{\alpha, \beta_{h}}^{(n)}(-B) e^{-i \beta_{2}^{n} B} .
\end{align*}
$$

Similarly to the continuous problem, but now we use equations (6.38) and (6.41) so that the boundary conditions for the discrete problem are given by

$$
\begin{array}{ll}
\left(T_{+}^{\alpha, \beta^{M}}-\frac{\partial}{\partial \eta}\right) U_{\alpha, \beta_{h}}=2 i \beta_{1}^{0}, & \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, \beta^{M}}-\frac{\partial}{\partial \eta}\right) U_{\alpha, \beta_{h}}=0, & \text { on } \Gamma_{-} . \tag{6.44}
\end{array}
$$

Therefore, instead of solving directly equation (6.34) we approximate $U_{\alpha, \beta}$ by $U_{\alpha, \beta_{h}}^{M}$ and we solve numerically the following problem. We want to find $U_{\alpha, \beta_{h}}^{M} \in X^{\alpha, \beta}$ such that for all $v_{h} \in X^{\alpha, \beta}$ we have

$$
\begin{equation*}
a^{M}\left(U_{\alpha, \beta_{h}}^{M}, v_{h}\right)=\left(f_{\alpha, \beta}, v_{h}\right), \tag{6.45}
\end{equation*}
$$

subject to the constraint

$$
U_{\alpha, \beta_{h}}^{M}=0,(x, y) \in \partial \Omega_{3}
$$

where

$$
\begin{align*}
a^{M}\left(s_{h}, v_{h}\right) & =\left(\nabla s_{h}, \nabla v_{h}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-2 i \alpha\left(\partial_{x} s_{h}, v_{h}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} \\
& +2 i \beta_{1}^{0}\left(\partial_{y} s_{h}, v_{h}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}^{\alpha, 0^{M}} s_{h}, v_{h}\right)_{\Gamma_{ \pm}} \mp i \beta_{1}^{0}\left(s_{h}, v_{h}\right)_{\Gamma_{ \pm}},  \tag{6.46}\\
\left(f_{\alpha, \beta}, v_{h}\right)_{\Gamma_{ \pm}} & =\left(-2 i \beta_{1}^{0}, v_{h}\right)_{\Gamma_{+}},
\end{align*}
$$

for all $s_{h} \in X^{\alpha, \beta}$. This leads to a linear algebraic system

$$
\begin{equation*}
A^{M} U^{M}=L \tag{6.47}
\end{equation*}
$$

with $U^{M}=U_{j}^{M}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, \beta_{h}}^{M}=\sum_{j=1}^{N} U_{j}^{M} \psi_{j}(x, y)$.

$$
A^{M}=a^{M}\left(\psi_{i}, \psi_{j}\right),
$$

and

$$
L=\left(f_{\alpha, \beta}, \psi_{j}\right)_{\Gamma_{+}}
$$

for $\{i, j\} \in\{1, . ., N\}$.

### 6.2.3 Efficiency of the grating

The efficiency of the grating with respect to each diffraction order $n$ can be computed by using equations (2.55), (4.49) and (6.40), to give

$$
\begin{align*}
& R_{n}^{M}=\frac{\beta_{1}^{n}}{\beta_{1}^{0}}\left|r_{1}^{n, M}\right|^{2}  \tag{6.48}\\
& T_{n}^{M}=\frac{\beta_{2}^{n}}{\beta_{1}^{0}}\left|t_{2}^{n, M}\right|^{2}
\end{align*}
$$

such that $r_{1}^{n, M}$ and $t_{2}^{n, M}$ are given by equation (6.42). We use the same notation as in Section 4.2.3, where we replaced $E_{t}, E_{r}$ and $E_{a b s}$ with $E_{t}^{M}, E_{r}^{M}$ and $E_{a b s}^{M}$ in Definition 54. In a similar way to Section 4.2.4, we can check the energy balance using the truncated DtN map using equation (4.72) such that $r_{1}^{n, M}$ and $t_{2}^{n, M}$ are now given by equation (6.42).

### 6.3 A priori error estimates for the exact solution $U_{\alpha, \beta}$

The well-posedness of our problem that we derived in Section 6.1 .2 will now allow us to derive an a priori error estimate for $U_{\alpha, \beta}$. We will first study the error estimation by discretising the continuous Helmholtz problem given by equation (6.14).

Since we truncate our domain, we apply absorbing boundary conditions through $T_{ \pm}^{\alpha, \beta}$. As these involve an infinite sum, we then need to truncate the $T_{ \pm}^{\alpha, \beta}$ in order to perform numerical computation. Hence, we need to study a second a priori error estimate that arises from this truncation. We combine these two error estimates in order to provide the total a priori error estimate. We will also show that the discrete solution is unique. Before we derive an a priori error estimate, we first need to establish the dual problem corresponding to equation (6.14) of the form

$$
\begin{equation*}
a^{*}(\psi, v)=(\psi, \phi) \tag{6.49}
\end{equation*}
$$

for all $\psi \in L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right), v \in H_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and for a given $\phi \in L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$. This dual problem is needed to derive an upper bound of the error estimate in the $L_{\#}^{2}$-norm in terms of the error estimate in the H -norm when we discretise our problem. The formulation of the dual problem is given below.

### 6.3.1 Dual problem

Lemma 73. Let $v \in H_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$. Then the strong form of the dual problem corresponding to equation (6.14) is given by

$$
\begin{aligned}
-\Delta v+2 i \alpha \partial_{x} v-2 i \beta_{1}^{0} \partial_{y} v & =\phi, \\
T_{ \pm}^{\alpha, 0^{*}} v-\partial_{n} v & =0,
\end{aligned}
$$

for a given $\phi \in L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and where $T_{ \pm}^{\alpha, 0^{*}}$ is the dual operator of $T_{ \pm}^{\alpha, 0}[61$, p. 476].

Proof. From the Green identities [22, p. 130], and by integrating by parts first with respect to $x$ and then integrate with respect to $y$ we have

$$
\begin{align*}
\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\partial_{x} \psi\right) \bar{v} & =-\int_{\Omega \backslash \mathrm{int} \Omega_{3}} \psi \partial_{x} \bar{v}+\int_{\partial \Omega \backslash \mathrm{int} \Omega_{3}} \psi \bar{v} n_{1} d s,  \tag{6.50}\\
& =-\int_{\Omega \backslash \mathrm{int} \Omega_{3}} \psi \partial_{x} \bar{v}+\int_{\Gamma_{L} \cup \Gamma_{R}} \psi \bar{v} n_{1} d s .
\end{align*}
$$

This leads us to

$$
\begin{align*}
\int_{\Omega \backslash i \operatorname{int} \Omega_{3}}\left(\partial_{x} \psi\right) \bar{v} & =-\int_{\Omega \backslash \operatorname{int} \Omega_{3}} \psi \partial_{x} \bar{v}+\int_{\Gamma_{L}} \psi \bar{v}\left(n_{1 L}-n_{1 R}\right) d s \\
& =-\int_{\Omega \backslash \operatorname{int} \Omega_{3}} \psi \partial_{x} \bar{v}, \tag{6.51}
\end{align*}
$$

since $\psi$ and $v$ are periodic with respect to $x$ and the unit normal outward vector corresponding to $\Gamma_{L}$ and $\Gamma_{R}$, denoted respectively by $n_{1 L}$ and $n_{1 R}$, satisfy

$$
n_{1 R}=-n_{1 L} .
$$

Once again we use the Green identities [22, p. 130], and proceed in a similar way to the proof of equation (6.50) but exchange the order of integration in equation (6.50). We then have

$$
\begin{align*}
\int_{\Omega \backslash \mathrm{int} \Omega_{3}} i \beta_{1}^{0}\left(\partial_{y} \psi\right) \bar{v} & =-i \beta_{1}^{0} \int_{\Omega \backslash \operatorname{int} \Omega_{3}} \psi \partial_{y} \bar{v}+\int_{\partial \Omega \backslash \operatorname{int} \Omega_{3}} i \beta_{1}^{0} \psi \bar{v} n_{2} d s, \\
& =-\int_{\Omega \backslash \mathrm{int} \Omega_{3}} i \beta_{1}^{0} \psi \partial_{y} \bar{v} \pm i \beta_{1}^{0} \int_{\Gamma_{ \pm}} \psi \bar{v} d x . \tag{6.52}
\end{align*}
$$

From the Green identities [22, p. 130] with equations (6.51), (6.52), (2.66), (6.13) and (6.2) we get

$$
\begin{aligned}
a(\psi, v)= & (\psi,-\Delta v)_{\Omega \backslash \operatorname{int} \Omega_{3}}+\left(\psi, \partial_{n} v\right)_{\partial \Omega \backslash \operatorname{int} \Omega_{3}}+2 i \alpha\left(\psi, \partial_{x} v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} \\
& -2 i \beta_{1}^{0}\left(\psi, \partial_{y} v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} \pm \int_{\Gamma_{ \pm}} i \beta_{1}^{0} \psi \bar{v} d x \\
& \mp \int_{\Gamma_{ \pm}} i \beta_{1}^{0} \psi \bar{v} d x-\left(\psi, T_{ \pm}^{\alpha, 0^{*}} v\right) .
\end{aligned}
$$

By definition of the dual, we have

$$
\begin{align*}
a^{*}(\psi, v)= & (\psi,-\Delta v)_{\Omega \backslash \operatorname{int} \Omega_{3}}+\left(\psi, \partial_{n} v\right)_{\partial \Omega \backslash \text { int } \Omega_{3}}+2 i \alpha\left(\psi, \partial_{x} v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} \\
& -2 i \beta_{1}^{0}\left(\psi, \partial_{y} v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} \pm \int_{\Gamma_{ \pm}} i \beta_{1}^{0} \psi \bar{v} d x \\
& \mp \int_{\Gamma_{ \pm}} i \beta_{1}^{0} \psi \bar{v} d x-\left(\psi, T_{ \pm}^{\alpha, 0^{*}} v\right) . \tag{6.53}
\end{align*}
$$

Hence, we can put together the integral inside $\Omega \backslash \operatorname{int} \Omega_{3}$ separately from the integrals on the boundaries of $\Omega \backslash$ int $\Omega_{3}$ and we can use equation (6.49) and so

$$
\begin{aligned}
-\Delta v+2 i \alpha \partial_{x} v-2 i \beta_{1}^{0} \partial_{y} v & =\phi, \\
\partial_{n} v-T_{ \pm}^{\alpha, 0^{*}} v & =0
\end{aligned}
$$

which represents the strong form of the dual problem which finishes the proof.

### 6.3.2 An a priori error estimation of the discretised problem

Let $U_{\alpha, \beta}$ be the corresponding solution of the continuous problem given by equation (6.14), and let $U_{\alpha, \beta_{h}}$ be the solution of the discretised problem given by equation (6.34). We are going to study the relation between the norm of the error in approximating the problem using the H -norm and the $L_{\#}^{2}$-norm in the following theorem.

Theorem 74. If we denote $e_{\alpha, \beta_{h}}=U_{\alpha, \beta}-U_{\alpha, \beta_{h}}$ and if $k_{1} h_{0} / p_{0}<1$ then for all $h \leq h_{0}$ and for all $p \geq p_{0}$, the solution to equation (6.34) satisfies

$$
\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq \tilde{C}_{1} \frac{h}{p} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}
$$

where $\tilde{C}_{1}=2 k_{1} C C_{c} C_{\text {reg }}$ with $C_{c}=C d+1$ and $C_{\text {reg }}$ as defined in Theorem 27.
Proof. We apply the duality argument [22, p. 137] similar to the proof of Lemma 59. From [22, p. 146], we have

$$
\begin{equation*}
\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}=\sup _{\phi \in C_{\infty}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \frac{\left|\left(e_{\alpha, \beta_{h}}, \phi\right)_{\Omega \backslash \text { int } \Omega_{3}}\right|}{\mathrm{k} \phi \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}} . \tag{6.54}
\end{equation*}
$$

Let $v$ be the solution to the adjoint problem given by Lemma 73. We have for a given $\phi \in L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ that

$$
a^{*}(\psi, v)=(\psi, \phi)_{\Omega \backslash \operatorname{int} \Omega_{3}}, \forall \psi \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)
$$

We then have for $\psi=e_{\alpha, \beta_{h}}$ by using equations (6.49) and (6.53)

$$
\begin{aligned}
\left(e_{\alpha, \beta_{h}}, \phi\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} & =a\left(e_{\alpha, \beta_{h}}, v\right), \\
& =a\left(e_{\alpha, \beta_{h}}, v-\psi\right)
\end{aligned}
$$

for any $\psi \in X^{\alpha, \beta}$ by using Galerkin orthogonality similar to equation (4.88).
From equation (6.15), we have

$$
\begin{equation*}
|a(s, v)|=\left|(\nabla s, \nabla v)_{\Omega \backslash \operatorname{int} \Omega_{3}}\right|+2 k_{1}\left|(\nabla s, v)_{\Omega \backslash i n t} \Omega_{3}\right|+\left|\left(T_{ \pm}^{\alpha, \beta} s, v\right)_{\Gamma_{ \pm}}\right| . \tag{6.55}
\end{equation*}
$$

Using Lemma 9 and since $e^{-i \beta y}$ is independent on $x$, we get

$$
\begin{align*}
\int_{\Gamma_{ \pm}} T_{ \pm}^{\alpha, 0} s \bar{v} & =\int_{\Gamma_{ \pm}} e^{i \alpha x-i \beta y}\left(T_{ \pm}^{\alpha, 0} s\right) \overline{e^{i \alpha x-i \beta y}}, \\
& =\int_{\Gamma_{ \pm}} T_{ \pm}\left(e^{i \alpha x-i \beta y} s\right) \overline{e^{i \alpha x-i \beta y}} . \tag{6.56}
\end{align*}
$$

We first look for an upper bound for $T_{ \pm}^{\alpha, \beta}$. From equations (4.80), (2.64) and (6.16) , we have

$$
\begin{align*}
\left|\left(T_{ \pm}^{\alpha, \beta} s, v\right)_{\Gamma_{ \pm}}\right| & =d\left|\sum_{n \in \mathbb{Z}} i\left(\beta_{1}^{n} \pm \beta_{1}^{0}\right) s^{(n)}( \pm B) \overline{v^{(n)}( \pm B)}\right| \\
& \leq d\left|\sum_{n \in \mathbb{Z}}\right| \beta_{1}^{n}+\beta_{1}^{0}\left|s^{(n)}( \pm B) \overline{v^{(n)}( \pm B)}\right| \tag{6.57}
\end{align*}
$$

We note that when $k^{2}>\frac{2 \pi n}{d}$, then

$$
\begin{equation*}
\left|\beta_{1}^{n}+\beta_{1}^{0}\right| \leq 2 \beta_{1}^{0} \leq 2 k_{1} . \tag{6.58}
\end{equation*}
$$

For $k_{1}<\left(\frac{2 \pi n}{d}\right)$, we have

$$
\begin{aligned}
\beta_{1}^{n}+\beta_{1}^{0} & =\sqrt{\left(\frac{2 \pi n}{d}\right)^{2}+\frac{4 \pi n \alpha}{d}+\alpha^{2}-k^{2}}+\sqrt{ } \overline{k^{2}-\alpha^{2}} \\
& =\sqrt{\frac{2 \pi n^{2}}{d}\left(1+\frac{d \alpha}{2 \pi n}+d^{2} \frac{\alpha^{2}-k^{2}}{4 \pi^{2} n^{2}}\right)}+\sqrt{ } \frac{k^{2}-\alpha^{2}}{} .
\end{aligned}
$$

Let us denote $X^{2}=d^{2} \frac{\alpha^{2}-k^{2}}{4 \pi^{2} n^{2}}$, since $\frac{2 \pi n}{d}>k_{1}$ then $\frac{2 \pi n}{d}{ }^{2} \gg k_{1}^{2}$ then $X^{2}$ tends to 0 . We note by using Taylor series that $\left(1-X^{2}\right)^{1 / 2}$ when $X^{2}$ tends to zero, can be approximated with $1-X$ and so

$$
\begin{equation*}
\left|\beta_{1}^{n}+\beta_{1}^{0}\right| \leq C\left|\frac{2 \pi n}{d}\right| \tag{6.59}
\end{equation*}
$$

where $1<C^{2}=\sup \left(c_{1}, c_{2}\right)<5$, the constants $c_{1}$ and $c_{2}$ are given by equations (2.68) and (2.69). We then combine equations (6.56), (6.58), (6.59) and use a similar argument to that use to prove equation (4.82) with $T_{ \pm}^{\alpha, \beta}$ to get

$$
\begin{aligned}
\left|\left(T_{ \pm}^{\alpha, \beta} s, v\right)_{\Gamma_{ \pm}}\right| & \leq C d \mathbf{k} s \mathbf{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\left(2 k_{1} \mathbf{k} v \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathbf{k} v \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\right) \\
& \leq C d(h / p)^{-1} \mathbf{k} s \mathbf{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\left(2 k_{1} h / p \mathbf{k} v \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+h / p \mathbf{k} v \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash i \operatorname{int} \Omega_{3}\right)}\right) .
\end{aligned}
$$

By supposing $2 k_{1} h / p<1$, we have

$$
\begin{aligned}
& \left|\left(T_{ \pm}^{\alpha, \beta} s, v\right)_{\Gamma_{ \pm}}\right| \\
& \leq C d(h / p)^{-1} \mathrm{k} s \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\left(\mathrm{k} v \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+h / p \mathrm{k} v \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\right) \\
& \leq C d(h / p)^{-1} \mathrm{k} s \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\left(\mathrm{k} e^{i \alpha x-i \beta_{1}^{0} y} v \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+2 h / p \mathrm{k} e^{i \alpha x-i \beta_{1}^{0} y} v \mathrm{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\right),
\end{aligned}
$$

from Theorem 69. Since $e^{i \alpha x-i \beta y}(v)$ satisfies equation (4.39), then we have

$$
\begin{equation*}
\left|\left(T_{ \pm}^{\alpha, \beta} s, v\right)_{\Gamma_{ \pm}}\right| \leq 2 C C d(h / p)^{-1} \mathbf{k} s \mathbf{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}(h / p)^{2} \mathbf{k} e^{i \alpha x-i \beta_{1}^{0} y} v \mathrm{k}_{H_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} . \tag{6.60}
\end{equation*}
$$

We now find an upper bound of the other term left in equation (6.55). We note that

$$
\begin{aligned}
& |s|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}|v|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+2 k_{1}|s|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} s \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \\
& =(h / p)^{-1}|s|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\left(h / p|v|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+2 k_{1} h / p|v|_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\right) .
\end{aligned}
$$

Since $2 k_{1} h / p<1$ and similar to the proof of equation (6.60) we get

$$
\begin{gather*}
|s|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}|v|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+2 k_{1}|s|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} s \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \\
=2 C(h / p)^{-1}|s|_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}(h / p)^{2} \mathrm{k} e^{i \alpha x-i \beta_{1}^{0} y} v \mathrm{k}_{H_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} . \tag{6.61}
\end{gather*}
$$

Now we can use equations (6.60), (6.61) and (6.55) to get

$$
\begin{align*}
&\left|a\left(e_{\alpha, \beta_{h}}, v-\psi\right)\right|\left.\leq 2 h / p(C d+1) \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{H_{\#}^{1}(\Omega \backslash i n t} \Omega_{3}\right) \\
&\left.\mathrm{k} e^{i \alpha x-i \beta_{1}^{0} y}(v-\psi) \mathrm{k}_{H_{\alpha \#}^{2}(\Omega \backslash i n t} \Omega_{3}\right) \\
&\left.2 h / p C C_{r e g} k_{1}(C d+1) \mathrm{k} s \mathrm{k}_{H_{\#}^{1}(\Omega \backslash i n t} \Omega_{3}\right)  \tag{6.62}\\
&\left.\mathrm{k} e^{i \alpha x-i \beta_{1}^{0} y} \phi \mathrm{k}_{H_{\alpha \#}^{2}(\Omega \backslash i n t} \Omega_{3}\right), \\
&\left.2 h / p C C_{r e g} k_{1}(C d+1) \mathrm{k} s \mathrm{k}_{H_{\#}^{1}(\Omega \backslash i n t} \Omega_{3}\right) \\
&\left.\mathrm{k} e^{i \alpha x-i \beta_{1}^{0} y} \phi \mathrm{k}_{H_{\alpha \#}^{2}(\Omega \backslash i n t} \Omega_{3}\right),
\end{align*}
$$

from Theorem 27 and equation (6.26).
We finish the proof by taking the supremum over $\phi$ using equation (6.54) and by denoting $\tilde{C}_{1}=2 C C_{c} C_{r e g} k_{1}$, where $C_{c}=C d+1$.

Now that we have the relation between the norm of the error estimate by discretising in H and in $L_{\#}^{2}$, let us establish that the Galerkin method satisfies quasi-optimal convergence [114, 109], which is an upper bound of the error $e_{\alpha, \beta_{h}}$ in terms of $U_{\alpha, \beta}-\psi$ for all $\psi \in X^{\alpha, \beta}$.

Lemma 75. Let $h_{0}$ and $p_{0}$ satisfy $k_{1} h_{0} / p_{0} \ll 1$, then for any maximum mesh size $h \in\left[0, h_{0}\right]$ and any degree of the polynomial basis $p \in\left[p_{0}, \infty\right]$

$$
\begin{equation*}
\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \leq C_{q} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}} \tag{6.63}
\end{equation*}
$$

for all $\psi \in X^{\alpha, \beta}$, where $C_{q}=\tilde{C}_{c} /\left(M_{G}-\tilde{C}_{1} k_{1}\left(\xi_{1} \xi+M_{G}\right) h / p\right)$. The constants $M_{G}, \xi, \xi_{1}$ are given by equation (6.21), $\tilde{C}_{c}$ by equation (6.20) and $\tilde{C}_{1}$ is given in Theorem 74.

Proof. From equation (6.22) and using the equivalence of the $\mathbf{H}$-norm and the $L_{\#}^{2}$-norm in Definition 49, we have

$$
\begin{align*}
\left|a\left(e_{\alpha, \beta_{h}}, e_{\alpha, \beta_{h}}\right)\right| & \geq M_{G} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}^{2}-k_{1}\left(\xi_{1} \xi+M_{G}\right) \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}, \\
& \geq\left(M_{G}-\tilde{C}_{1} k_{1}\left(\xi_{1} \xi+M_{G}\right) h / p\right) \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}^{2}, \tag{6.64}
\end{align*}
$$

using Theorem 74. Since $a$ is continuous from equation (6.19) and by using Galerkin orthogonality similar to equation (4.88), we can divide by $\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}$ equation (6.64) and we get

$$
\tilde{C}_{c} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}} \geq\left(M_{G}-\tilde{C}_{1} k_{1}\left(\xi_{1} \xi+M_{G}\right) h / p\right) \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}},
$$

for all $\psi \in X^{\alpha, \beta}$. Hence, for $h<h_{0}$ and $p \geq p_{0}, M_{G}-\tilde{C}_{1} k_{1}\left(\xi_{1} \xi+M_{G}\right) h / p>0$ which finishes our proof.

We can now use the equivalence of the H -norm and the $L_{\#}^{2}$-norm of $e_{\alpha, \beta_{h}}$ and the quasi-optimal convergence result to establish the a priori error estimate.

Theorem 76. Let $h_{0}$ and $p_{0}$ satisfy $k_{1} h_{0} / p_{0} \leq 1$, then for any maximum mesh size $h \in\left[0, h_{0}\right]$ and any degree of the polynomial basis $p \in\left[p_{0}, \infty\right]$ we have the following error estimates

$$
\begin{align*}
& \mathrm{k} U_{\alpha, \beta}-U_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \leq \frac{\tilde{C}_{c}+C_{q}}{C_{d}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}  \tag{6.65}\\
& \mathrm{k} U_{\alpha, \beta}-U_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq \tilde{C}_{1} h \frac{\tilde{C}_{c}+C_{q}}{p C_{d}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}
\end{align*}
$$

for all $\psi \in X^{\alpha, \beta}$, with $C_{d}=1-\tilde{C}_{1} k_{1} \xi \xi_{1} \frac{h}{p} \geq 0$ where $C_{q}$ is given by equation (6.63), $\tilde{C}_{1}$ is given by Theorem 74, $\xi$ and $\xi_{1}$ are given by equation (6.21) and $\tilde{C}_{c}$ is given by equation (6.20). In addition, the discretised solution $U_{\alpha, \beta_{h}}$ is unique.

Proof. For all $v_{\alpha, \beta} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, since $U_{\alpha, \beta_{h}} \in X^{\alpha, \beta} \subset H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, and $e_{\alpha, \beta_{h}}=U_{\alpha, \beta}-U_{\alpha, \beta_{h}}$ belongs to $H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$, it follows from equation (6.22) that

$$
\begin{aligned}
M_{G} \mathrm{k} e_{\alpha, \beta_{h}} & \mathrm{k}_{\mathcal{H}}^{2}-C_{G} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
& \leq\left|a\left(e_{\alpha, \beta_{h}}, e_{\alpha, \beta_{h}}\right)\right| \\
& \leq\left|a\left(e_{\alpha, \beta_{h}}, U_{\alpha, \beta}-U_{\alpha, \beta_{h}}-\psi+\psi\right)\right|, \\
& \leq\left|a\left(e_{\alpha, \beta_{h}}, U_{\alpha, \beta}-\psi\right)\right|,
\end{aligned}
$$

for $\psi \in X^{\alpha, \beta}$, by using Galerkin orthogonality similar to equation (4.88). We then can write

$$
\begin{aligned}
& \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}^{2}-k_{1}^{2} \xi_{1} \xi \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
& \quad \leq \tilde{C}_{c} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}+\sup \left(\xi_{1} / \xi, M_{G}\right) \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}^{2},
\end{aligned}
$$

from equation (6.19) since $a$ is continuous and using equations (6.21) and (6.23).

$$
\begin{align*}
& \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}^{2}-k_{1}^{2} \xi \xi_{1} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
& \quad \leq \tilde{C}_{c} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}+\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}^{2} \tag{6.66}
\end{align*}
$$

since $\sup \left(\xi_{1} / \xi, M_{G}\right)=\left(\xi_{1} / \xi, 1-\xi_{1} / \xi\right) \leq 1$. Using Definition 49, we note that $k_{1} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \leq \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}$. Hence, we can write

$$
\begin{aligned}
& \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}^{2}-k_{1} \xi \xi_{1} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\neq}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \\
& \leq\left(\tilde{C}_{c}+C_{q}\right) \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}
\end{aligned}
$$

from equation (6.63). We then divide by $\mathbf{k} e_{\alpha, \beta_{h}} \mathbf{k}_{\mathcal{H}}$ and so

$$
\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}-k_{1} \xi_{1} \xi \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq\left(\tilde{C}_{c}+C_{q}\right) \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}
$$

Hence,

$$
\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}}-\tilde{C}_{1} k_{1} \xi_{1} \xi \frac{h}{p} \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \leq\left(\tilde{C}_{c}+C_{q}\right) \mathrm{k} U_{\alpha, \beta}-\psi \mathbf{k}_{\mathcal{H}},
$$

from Theorem 74. That is

$$
\left(1-\tilde{C}_{1} k_{1} \xi_{1} \xi \frac{h}{p}\right) \mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \leq\left(\tilde{C}_{c}+C_{q}\right) \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}} .
$$

Let $C_{d}=1-\tilde{C}_{1} k_{1} \xi_{1} \xi \frac{h}{p}$ and let us suppose that $C_{d}>0$ then we have

$$
\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{\mathcal{H}} \leq \frac{\tilde{C}_{c}+C_{q}}{C_{d}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}} .
$$

Once again, we invoke Theorem 74 to prove the error estimate in the $L^{2}$-norm. Hence, we have

$$
\mathrm{k} e_{\alpha, \beta_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \leq\left(\tilde{C}_{c}+C_{q}\right) \tilde{C}_{1} \frac{h}{p C_{d}} \mathrm{k} U_{\alpha, \beta}-\psi \mathbf{k}_{\mathcal{H}} .
$$

To prove the uniqueness of the solution let us suppose that there exists two solutions $U_{\alpha, \beta}^{h, 1}$ and $U_{\alpha, \beta}^{h, 2}$ satisfying equation (6.34). We have

$$
\begin{aligned}
& \mathrm{k} U_{\alpha, \beta}^{h, 1}-U_{\alpha, \beta}^{h, 2} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}=\mathrm{k} U_{\alpha, \beta}^{h, 1}-U_{\alpha, \beta}+U_{\alpha, \beta}-U_{\alpha, \beta}^{h, 2} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}, \\
& \leq \mathrm{k} U_{\alpha, \beta}^{h, 1}-U_{\alpha, \beta} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}+\mathrm{k} U_{\alpha, \beta}^{h, 2}-U_{\alpha, \beta} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}, \\
& \leq 2 \tilde{C}_{1}\left(\tilde{C}_{c}+C_{q}\right) \frac{h}{p C_{d}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} .
\end{aligned}
$$

We can easily see that taking $h / p$ tending to zero, we finish the proof of uniqueness.

Before studying the error made by truncating the operators $T_{ \pm}^{\alpha, 0}$ in equation (6.13), let us establish a relation between the two continuous variational formulations $a(s, v)$ given by equation (6.13) and $a^{M}(s, v)$ (when we truncate the DtN map inside equation (6.13)).

### 6.3.3 Comparison between the continuous variational formulation with a truncated DtN map and that with a full DtN map

Let us denote

$$
\begin{align*}
a^{M}\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right)= & \left(\nabla s_{\alpha, \beta}, \nabla v_{\alpha, \beta}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-2 i \alpha\left(\partial_{x} s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Omega \backslash \text { int } \Omega_{3}} \\
& +2 i \beta_{1}^{0}\left(\partial_{y} s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}^{\alpha, 0^{M}} s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{ \pm}} \mp 2 i \beta_{1}^{0}\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{ \pm}},  \tag{6.67}\\
\left(f_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{+}}= & \left(-2 i \beta_{1}^{0}, v_{\alpha, \beta}\right)_{\Gamma_{+}}, \tag{6.68}
\end{align*}
$$

for all $s_{\alpha, \beta}, v_{\alpha, \beta} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, such that $M \in \mathbb{N}$ and $T_{ \pm}^{\alpha, 0^{M}}$ is given by equation (4.61). Then, we want to find $U_{\alpha, \beta}^{M} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ such that

$$
\begin{equation*}
a^{M}\left(U_{\alpha, \beta}^{M}, v_{\alpha, \beta}\right)=\left(f_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{+}} . \tag{6.69}
\end{equation*}
$$

Let us also denote

$$
\begin{align*}
S a\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right)= & \left(\nabla s_{\alpha, \beta}, \nabla v_{\alpha, \beta}\right)_{\Omega \backslash \text { int } \Omega_{3}}-2 i \alpha\left(\partial_{x} s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Omega \backslash \text { int } \Omega_{3}} \\
& +2 i \beta_{1}^{0}\left(\partial_{y} s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Omega \backslash \text { int } \Omega_{3}} \mp 2 i \beta_{1}^{0}\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{ \pm}} . \tag{6.70}
\end{align*}
$$

It follows that

$$
\begin{align*}
a\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right) & =S a\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right)-\left(T_{ \pm}^{\alpha, 0} s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{ \pm}}  \tag{6.71}\\
a^{M}\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right) & =S a\left(s_{\alpha, \beta}, v_{\alpha, \beta}\right)-\left(T_{ \pm}^{\alpha, 0^{M}} s_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{ \pm}} .
\end{align*}
$$

Let us also note the following relation between $a$ as defined by equation (6.13) and $a^{M}$ as given by equation (6.67). If $U_{\alpha, \beta}$ and $U_{\alpha, \beta}^{M}$ are respectively the solution of equation (6.14) and equation (6.69) then

$$
\begin{equation*}
a\left(U_{\alpha, \beta}, v\right)-a^{M}\left(U_{\alpha, \beta}^{M}, v\right)=0 \tag{6.72}
\end{equation*}
$$

for all $v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$. If we define $T_{ \pm}^{\alpha, 0^{R}}$ via

$$
\begin{equation*}
T_{ \pm}^{\alpha, 0}=T_{ \pm}^{\alpha, 0^{M}}+T_{ \pm}^{\alpha, 0^{R}} . \tag{6.73}
\end{equation*}
$$

then we have the following relations for all $s, v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$

$$
a(s, v)=S a(s, v)-\left(T_{ \pm}^{\alpha, 0} s, v\right)_{\Gamma_{ \pm}}
$$

from equation (6.70). Hence,

$$
\begin{equation*}
a(s, v)=a^{M}(s, v)-\left(T_{ \pm}^{\alpha, 0^{R}} s, v\right)_{\Gamma_{ \pm}} . \tag{6.74}
\end{equation*}
$$

Now that we have established the relation between $a$ and $a^{M}$ in equation (6.74) we can derive an a priori error estimate when we truncate $T_{ \pm}^{\alpha, \beta}$ in $a$.

### 6.3.4 An a priori error estimate from the truncation of the DtN operators

For computational reasons the DtN map must be truncated. It is therefore important to derive an estimate of the error that then arises due to this approximation.

Lemma 77. Let $U_{\alpha, \beta} \in H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ be the solution of equation (6.14), and let $U_{\alpha, \beta}^{M}$ be the solution of equation (6.69). Let us denote the error estimate that arises when the DtN map is truncated by

$$
\begin{equation*}
e_{\alpha, \beta}^{M}=U_{\alpha, \beta}-U_{\alpha, \beta}^{M} . \tag{6.75}
\end{equation*}
$$

Then, we have

$$
\begin{equation*}
a^{M}\left(e_{\alpha, \beta}^{M}, e_{\alpha, \beta}^{M}\right)=\left(T_{ \pm}^{\alpha, 0^{R}} U_{\alpha, \beta}, e_{\alpha, \beta}^{M}\right)_{\Gamma_{ \pm}} . \tag{6.76}
\end{equation*}
$$

Proof. We have from equation (6.75) and from Definition A-6 that

$$
\begin{aligned}
a^{M}\left(e_{\alpha, \beta}^{M}, e_{\alpha, \beta}^{M}\right) & =a^{M}\left(U_{\alpha, \beta}, e_{\alpha, \beta}^{M}\right)-a^{M}\left(U_{\alpha, \beta}^{M}, e_{\alpha, \beta}^{M}\right), \\
& =a\left(U_{\alpha, \beta}, e_{\alpha, \beta}^{M}\right)+\left(T_{ \pm}^{\alpha, 0^{R}} U_{\alpha, \beta}, e_{\alpha, \beta}^{M}\right)_{\Gamma_{ \pm}}-a^{M}\left(U_{\alpha, \beta}^{M}, e_{\alpha, \beta}^{M}\right),
\end{aligned}
$$

from equation (6.74). Then, we can write

$$
a^{M}\left(e_{\alpha, \beta}^{M}, e_{\alpha, \beta}^{M}\right)=\left(T_{ \pm}^{\alpha, 0^{R}} U_{\alpha, \beta}, e_{\alpha, \beta}^{M}\right)_{\Gamma_{ \pm}},
$$

by using equation (6.72).
For all $s, v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and $\Gamma_{1, \pm}=\{(x, \pm b) \in \Omega\}$ as introduced in Section 4.3.2, we can use the region $b \leq|y| \leq B$ to derive an a priori error estimate. In a similar way to the derivation of equation (4.107), we get from equation (6.73)

$$
\begin{aligned}
& \left|\left(T_{ \pm}^{\alpha, 0^{R}} s, v\right)_{\Gamma_{ \pm}}\right|=\left|\left(\left(T_{ \pm}^{\alpha, 0}-T_{ \pm}^{\alpha, 0^{M}}\right) s, v\right)_{\Gamma_{ \pm}}\right|,
\end{aligned}
$$

with $c_{\text {min }}=\left.\inf \right|_{\left.n\right|_{>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right) \text { where } z_{n} \text { is defined by equation (2.44). From }}$ Theorem A-13, we get

$$
\begin{align*}
& \left(T_{ \pm}^{\alpha, 0^{R}} s, v\right)_{\Gamma_{ \pm}} \leq d e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{1}^{2}} \boldsymbol{k s}_{s} \mathbf{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k}^{\left(\mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}\right.} \\
& =C_{T} \mathrm{k} s \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \mathrm{kv} \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \tag{6.77}
\end{align*}
$$

with

$$
\begin{equation*}
C_{T}=d e^{-(B-b) c_{\min }} \sqrt{(M-|\alpha|)^{2}-k_{1}^{2}} . \tag{6.78}
\end{equation*}
$$

It will help our subsequent calculation of the a priori error estimate to establish a relation between the error by truncating the $\operatorname{DtN}$ operator in the $L_{\#}^{2}$-norm with that in the H -norm.

Theorem 78. Let $U_{\alpha, \beta}$ be the solution of equation (6.14) and let $e_{\alpha, \beta}^{M}$ be as given by equation (6.75) then we have the following estimate

$$
\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq \tilde{C}_{1} \frac{h}{p} \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}
$$

where $\tilde{C}_{1}$ is given in Theorem 14.
Proof. As in Lemma 59 we use the duality argument [22, p. 137]. Let $\phi \in$ $H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$, let $v \in H_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be the solution of the dual problem given in Lemma 73. From the definition of the dual norm [22, p. 146], we can write

$$
\begin{aligned}
\mathrm{ke}_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} & =\sup _{\phi \in C_{\infty}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \frac{\left|\left(e_{\alpha, \beta}^{M}, \phi\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}\right|}{\mathbf{k} \phi \mathbf{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}}, \\
& =\sup _{\phi \in C_{\infty}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \frac{\left|a\left(e_{\alpha, \beta}^{M}, v\right)\right|}{\mathbf{k} \phi \mathbf{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}} .
\end{aligned}
$$

In a similar way to the proof of Theorem 74, we can show that

$$
\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \leq \tilde{C}_{1} \frac{h}{p} \mathrm{k}_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}
$$

where $\tilde{C}_{1}$ is defined in Theorem 74 .
Finally then, the following results gives an estimate of the error made by truncating $T_{ \pm}^{\alpha, \beta}$ in the continuous variational formulation given by equation (6.13).

Theorem 79. There exists a positive integer $M_{0}=|k|+|\alpha|>0$ such that for all $M \geq M_{0}$, the problem with the truncated DtN map given by equation (6.69) satisfies the following error estimates.

$$
\begin{aligned}
\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} & \leq\left(C_{T}+C_{q}\right) \tilde{C}_{1} \frac{h}{p C_{d}} \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}}, \\
\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}} & \leq\left(C_{T}+C_{q}\right) / C_{d} \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}}
\end{aligned}
$$

with $C_{d} \geq 0$, is defined in Theorem 76 where $\tilde{C}_{1}$ is given by Theorem 74, $C_{q}$ by equation (6.63), $\xi$ and $\xi_{1}$ are given by equation (6.21) and $C_{T}$ by equation (6.78). Furthermore, the solution $U_{\alpha, \beta}^{M}$ is unique.
Proof. In a similar way to derive equation (6.22), we get

$$
M_{G} \mathbf{k} e_{\alpha, \beta}^{M} \mathbf{k}_{\mathcal{H}}^{2}-C_{G} \mathbf{k} e_{\alpha, \beta}^{M} \mathbf{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \leq\left|a^{M}\left(e_{\alpha, \beta}^{M}, e_{\alpha, \beta}^{M}\right)\right| .
$$

Hence,

$$
M_{G} \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}^{2}-C_{G} \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \leq\left|\left(T_{ \pm}^{\alpha, 0^{R}} U_{\alpha, \beta}, U_{\alpha, \beta}-\psi\right)_{\Gamma_{ \pm}}\right|,
$$

from equation (6.76). In a similar way to derive equation (6.66), we have

$$
\begin{aligned}
& \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}^{2}-k_{1}^{2} \xi_{1} \xi \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
& \quad \leq\left|\left(T_{ \pm}^{\alpha, 0^{R}} U_{\alpha, \beta}, e_{\alpha, \beta}^{M}\right)_{\Gamma_{ \pm}}\right|+\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}^{2} .
\end{aligned}
$$

We then use equations (6.77), (6.22) and Definition 49 to get

$$
\begin{aligned}
& \mathbf{k} e_{\alpha, \beta}^{M} \mathbf{k}_{\mathcal{H}}^{2}-k_{1} \xi_{1} \xi \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}} \\
& \leq C_{T} \mathbf{k} U_{\alpha, \beta} \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}+\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}^{2} .
\end{aligned}
$$

We then divide by $\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}$ and use Definition 49 again with equation (6.63) to give

$$
\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}-k_{1} \xi_{1} \xi \mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \leq\left(C_{T}+C_{q}\right) \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} .
$$

Using Theorem 78 gives

$$
\mathbf{k} e_{\alpha, \beta}^{M} \mathbf{k}_{\mathcal{H}}-k_{1} \tilde{C}_{1} \xi_{1} \xi \frac{h}{p} \mathbf{k} e_{\alpha, \beta}^{M} \mathbf{k}_{\mathcal{H}} \leq\left(C_{T}+C_{q}\right) \mathbf{k} U_{\alpha, \beta} \mathbf{k}_{\mathcal{H}} .
$$

Let us suppose that $C_{d} \geq 0$, as defined in Theorem 76, then we have

$$
\begin{equation*}
\mathrm{k} e_{\alpha, \beta}^{M} \mathbf{k}_{\mathcal{H}} \leq\left(C_{T}+C_{q}\right) / C_{d} \mathbf{k} U_{\alpha, \beta} \mathbf{k}_{\mathcal{H}} . \tag{6.79}
\end{equation*}
$$

If we invoke Theorem 78, with equation (6.79) then we immediately arrive at the error estimate in the $L^{2}$-norm

$$
\begin{equation*}
\mathbf{k} e_{\alpha, \beta}^{M} \mathbf{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq\left(C_{T}+C_{q}\right) \tilde{C}_{1} \frac{h}{p C_{d}} \mathbf{k} U_{\alpha, \beta}^{M} \mathbf{k}_{\mathcal{H}} . \tag{6.80}
\end{equation*}
$$

To prove the uniqueness, let us suppose that we have two solutions $U_{\alpha, \beta_{1}}^{M}$ and $U_{\alpha, \beta_{2}}^{M}$. Then we have
$\mathrm{k} U_{\alpha, \beta_{1}}^{M}-U_{\alpha, \beta_{2}}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \leq \mathrm{k} U_{\alpha, \beta}-U_{\alpha, \beta_{1}}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)}+\mathrm{k} U_{\alpha, \beta}-U_{\alpha, \beta_{2}}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)}$
which goes to zero when $\frac{h}{p}$ goes to zero from equation (6.80). Hence $U_{\alpha, \beta_{1}}^{M}$ is equal to $U_{\alpha, \beta_{2}}^{M}$ and the solution is unique.

### 6.3.5 Estimate of the total error

Similar to Section 4.3.3, the error that we make by solving numerically the Helmholtz equation for a periodic grating arises from two sources

- when we truncate the DtN operators which describe the boundary conditions.
- when we discretise the continuous problem.

Let us denote by $U_{\alpha, \beta_{h}}^{M} \in X^{\alpha, \beta}$ the solution of

$$
\begin{equation*}
a^{M}\left(U_{\alpha, \beta_{h}}^{M}, v_{\alpha, \beta}\right)=\left(f_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{+}}, \tag{6.81}
\end{equation*}
$$

for all $v_{\alpha, \beta} \in X^{\alpha, \beta}$, with $a^{M}$ and $\left(f_{\alpha, \beta}, v_{\alpha, \beta}\right)_{\Gamma_{+}}$given by equations (6.67) and (6.68), $M \in \mathbb{N}$ and $T_{ \pm}^{\alpha, 0^{M}}$ given by equation (4.61). The estimate of the total error, $e_{\alpha, \beta}=U_{\alpha, \beta}-\underset{U_{\alpha, \beta_{h}}^{M}}{\infty}$, that we make by solving numerically the Helmholtz equation is given by the following theorem.

Theorem 80. Let us suppose that $C_{d} \geq 0$ as defined in Theorem 76. Then there exist positive constant $h_{0} \leq 1, p_{0} \leq \infty$ and a positive integer $M_{0} \in \mathbb{N} \backslash\{0\}$ such that for any $h \in\left[0, h_{0}\right] p \in\left[p_{0}, \infty\right]$ and $M \geq M_{0}$, the total error satisfies

$$
\begin{aligned}
& \mathbf{k} e_{\alpha, \beta} \mathbf{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \leq \\
& \left(\tilde{C}_{c}+C_{q}\right) \tilde{C}_{1} \frac{h}{p C_{d}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}+\left(C_{T}+C_{q}\right) \tilde{C}_{1} \frac{h}{p C_{d}} \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}},
\end{aligned}
$$

and

$$
\begin{equation*}
\mathrm{k} e_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} \leq\left(\tilde{C}_{c}+C_{q}\right) / C_{d} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}+\left(C_{T}+C_{q}\right) / C_{d} \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} \tag{6.82}
\end{equation*}
$$

for all $\psi \in X^{\alpha, \beta}$, with $\tilde{C}_{c}$ as given in equation (6.20), $C_{d}$ as defined in Theorem 76, $C_{q}$ as defined in equation (6.63), $\xi$ and $\xi_{1}$ are given by equation (6.21) and $\tilde{C}_{1}$ as defined in Theorem 74. In addition, the problem equation (6.81) has a unique solution $U_{\alpha, \beta_{h}}^{M}$.

Proof. We have

$$
\begin{aligned}
\mathrm{k} e_{\alpha, \beta} \mathrm{k}_{\mathcal{H}}=\mathrm{k} U_{\alpha, \beta}-U_{\alpha, \beta_{h}}^{M} \mathrm{k}_{\mathcal{H}} & =\mathrm{k} U_{\alpha, \beta}-U_{\alpha, \beta}^{M}+U_{\alpha, \beta}^{M}-U_{\alpha, \beta_{h}}^{M} \mathrm{k}_{\mathcal{H}}, \\
& \leq \mathrm{k} U_{\alpha, \beta}-U_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}+\mathrm{k} U_{\alpha, \beta}^{M}-U_{\alpha, \beta_{h}}^{M} \mathrm{k}_{\mathcal{H}}, \\
& =\mathrm{k} e_{\alpha, \beta}^{M} \mathrm{k}_{\mathcal{H}}+\mathrm{k} U_{\alpha, \beta}^{M}-U_{\alpha, \beta_{h}}^{M} \mathrm{k}_{\mathcal{H}} .
\end{aligned}
$$

In a similar way to the proof of Theorem 76 , it can be shown that

$$
\begin{equation*}
\mathrm{k} U_{\alpha, \beta}^{M}-U_{\alpha, \beta_{h}}^{M} \mathrm{k}_{\mathcal{H}} \leq \frac{\tilde{C}_{c}+C_{q}}{C_{d}} \mathrm{k} U_{\alpha, \beta}^{M}-\psi \mathrm{k}_{\mathcal{H}} \tag{6.83}
\end{equation*}
$$

and

$$
\mathrm{k} U_{\alpha, \beta}^{M}-U_{\alpha, \beta_{h}}^{M} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq \tilde{C}_{1} \frac{\tilde{C}_{c}+C_{q}}{C_{d}} \frac{h}{p} \mathrm{k} U_{\alpha, \beta}^{M}-\psi \mathrm{k}_{\mathcal{H}}
$$

by using $U_{\alpha, \beta}$ instead of $U_{\alpha, \beta}^{M}$. For $M>M_{0}, U_{\alpha, \beta}^{M}$ tends to $U_{\alpha, \beta}$ and so we can replace $\mathbf{k} U_{\alpha, \beta}^{M}-\psi \mathbf{k}_{\mathcal{H}}$ with $\mathbf{k} U_{\alpha, \beta}-\psi \mathbf{k}_{\mathcal{H}}$. Then we can use equation (6.83) and Theorem 79 to get the total error estimate in the H -norm. We finish the proof by following a similar argument for the $L^{2}$-norm. In a similar way to the proof of Theorem 76 and Theorem 79, it can be shown that the solution $U_{\alpha, \beta_{h}}^{M}$ is unique.

### 6.4 Summary and conclusion

### 6.4.1 Summary

In this chapter, we transformed the diffraction problem Case 1A, using an extension of the $\alpha, 0$-quasi periodic transformation proposed in Chapter 4 . This was achieved by setting $U=e^{i \alpha x} e^{-i \beta_{1}^{0} y} U_{\alpha, \beta}$, where $e^{i \alpha x} e^{-i \beta_{1}^{0} y}$ is the analytic solution of the Helmholtz problem when we have a series of homogeneous layers (that is $\Omega_{3}$ is not present). We started by deriving the boundary value problem for $U_{\alpha, \beta}$ which was followed by the variational formulation of the continuous problem. Similar to Chapter 4, our aim was to derive an a priori error estimate. Hence the problem had to be shown to be well-posed. Recall that, in order for a problem to be well posed, the solution needs to exist, be unique and to depend continuously on the data. In Lemma 68 we showed that the solution of the variational problem exists and that it is unique. The continuous dependence on the data, was then shown in Theorem 70 when we studied the regularity of the solution. Comparing Theorems 51 and 70 , it is evident that $U_{\alpha, \beta}$ has the same $k_{1}$ dependence as $U_{\alpha, 0}$. Having dealt with the continuous problem, we then considered the discrete problem that arises when we approximate the continuous problem with a finite element solution and when we truncate the DtN operators. This then allowed us to derive an $a$ priori error estimate, due to discretisation and truncation of the DtN map, in Theorem 80. Having derived this error estimate we then showed that the discretised problem also had a unique solution. Again in our error estimate, we showed an explicit dependency on the maximum mesh size $h$, the degree of the polynomial basis $p$ and the wavenumber $k$.

### 6.4.2 Conclusion

We use Theorem A-13 and the note following Definition 49 and we note that

$$
\mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{1 / 2}\left(\Gamma_{1, \pm}\right)} \leq \mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}} .
$$

Using this result in Theorem 62, the upper bound for the a priori estimate of the total error in the $\mathbf{H}$-norm for the $\alpha, 0$-quasi periodic transformation is

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq 4 C_{c} / C_{3} \mathrm{k} U_{\alpha, 0}-\psi \mathrm{k}_{\mathcal{H}}+4 C_{T} / C_{3} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \tag{6.84}
\end{equation*}
$$

for all $\psi \in X^{\alpha}$ and where we use the definition of $C_{T}$ given by equation (6.78). The constant $C_{c}$ as given in Lemma 57 is independent of the wavenumber $k_{1}$, the mesh size $h$, and the degree of the polynomial basis $p, C_{3}=1-\left(\mathfrak{R}(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}>0$ as defined in Theorem 60 and $C_{1}$ is defined in Lemma 59. Hence, replacing $C_{3}$ by its definition and choosing $\psi=0$, equation (6.84) can be rewritten as

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{C_{c}+C_{T}}{1-2 C C_{c} C_{r e g} k_{1}^{2} h / p} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}} . \tag{6.85}
\end{equation*}
$$

In the following, we rewrite the upper bound corresponding to the a priori estimate of the total error in the H -norm for the $\alpha, \beta$-quasi periodic transformation given in Theorem 80 so that we can derive a qualitative comparison between the upper bounds. From Theorems 74 and 76, we have

$$
\begin{equation*}
C_{d}=1-2 C C_{r e g} C_{c} \xi \xi_{1} k_{1}^{2} \frac{h}{p} \tag{6.86}
\end{equation*}
$$

Hence, the upper bound corresponding to the a priori estimate of the total error in the $\mathbf{H}$-norm for the $\alpha, \beta$-quasi periodic transformation given in Theorem 80 can be rewritten as

$$
\begin{align*}
& \mathrm{k} e_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} \leq \\
& \frac{\tilde{C}_{c}+C_{q}}{1-2 C C_{r e g} C_{c} \xi \xi_{1} k_{1}^{2} \frac{h}{p}} \mathrm{k} U_{\alpha, \beta}-\psi \mathrm{k}_{\mathcal{H}}+\frac{C_{T}+C_{q}}{1-2 C C_{r e g} C_{c} \xi_{1} \xi k_{1}^{2} \underline{p}} \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} \tag{6.87}
\end{align*}
$$

for all $\psi \in X^{\alpha, \beta}$. To allow us to compare qualitatively between the upper bounds let us investigate the case where $\xi_{1}$ goes to zero, the denominator in equation (6.87) given by equation (6.86) tends to 1 . Hence, choosing $\psi=0$ equation (6.87) becomes

$$
\begin{equation*}
\mathbf{k} e_{\alpha, \beta} \mathbf{k}_{\mathcal{H}} \leq\left(\tilde{C}_{c}+C_{T}+2 C_{q}\right) \mathbf{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} . \tag{6.88}
\end{equation*}
$$

Since $\xi_{1}$ tends to zero then $C_{q}=\frac{\tilde{C}_{c}}{1-2 k_{1}^{2} C C_{c} C_{\text {reg }} h / p}$ then by denoting $X=2 C C_{c} C_{\text {reg }} k_{1}^{2} h / p$ and by noting that $\tilde{C}_{c}=C_{c}+(d+1)$, we can rewrite equation (6.89) as

$$
\begin{align*}
\mathrm{k} e_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} & \leq\left(\tilde{C}_{c}+C_{T}+2 \frac{C_{c}+d+1}{1-X}\right) \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}}, \\
& \leq \frac{\left(\tilde{C}_{C}+C_{T}\right)(1-X)+2 C_{c}+2 d+2}{1-X} \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} \tag{6.89}
\end{align*}
$$

and we can also rewrite equation (6.84) as follows

$$
\begin{align*}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} & \leq 4 \frac{C_{c}+C_{T}}{1-X} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}}, \\
& \leq \frac{2 C_{c}+2 C d+2+4 C_{T}}{1-X} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}}, \tag{6.90}
\end{align*}
$$

since $C_{c}=C d+1$. From Lemma 6, we have $2 \leq C^{2}<{ }^{\sqrt{\prime}} \overline{5}$ hence $C>1$ and $2 C d>2$. Therefore, when $X$ tends to 1 , the equivalent upper bound for the $\alpha, 0$-quasi periodic transformation is

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq \lim _{X \rightarrow 1} \frac{2 C_{c}+2 C d+2+4 C_{T}}{1-X} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}}, \tag{6.91}
\end{equation*}
$$

whereas the upper bound for the $\alpha, \beta$-quasi periodic transformation is

$$
\begin{equation*}
\mathrm{k} e_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} \leq \lim _{X \rightarrow 1} \frac{2 C_{c}+2 d+2}{1-X} \mathrm{k} U_{\alpha, \beta} \mathrm{k}_{\mathcal{H}} . \tag{6.92}
\end{equation*}
$$

From equations (6.91) and (6.92), we can see that the $\alpha, 0$-quasi periodic transformation blows up faster than the $\alpha, \beta$-quasi periodic transformation and this makes the error bound given in equation (6.91) for $U_{\alpha, 0}$ larger than the error bound given in equation (6.92) for $U_{\alpha, \beta}$. We have $X$ tending to 1 either by having $k^{2}$ small with $h / p$ large or by having $k^{2}$ large with $h / p$ small. We are interested in the latter case, since it is known that there are difficulties in solving the Helmhotz equation when the wavenumber $k$ becomes large. Equations (6.91) and (6.92) indicate that the $\alpha, \beta$-quasi periodic transformation is more efficient than the $\alpha, 0$-quasi periodic transformation proposed in Chapter 6 as we have suggested.

In the next chapter, we will show some numerical evidence and the advantage of this difference on the convergence behaviour between the $\alpha, \beta$-quasi periodic solution and $\alpha, 0$-quasi periodic solution for a different range of wavenumber $k_{1}$.

## Chapter 7

## Numerical methods and numerical results

This chapter is concerned with the numerical methods that we use in our implementations in order to solve the Helmholtz problem. To validate our code, numerical results will be given that compare the results from different numerical methods and experiments in the literature to our implementation. We will then provide numerical results that compare the $\alpha, \beta$-quasi periodic method with the $\alpha, 0$-quasi periodic method across a range of wavenumbers. We then produce numerical results that show the advantage of using the Dual Weighted Residual (DWR) method as compared to the uniform mesh or the global a posteriori error method proposed in [13]. We have chosen the finite element method because of its flexibility to adapt to complex scattering geometry. In our final example, we consider a scattering geometry from a real world application concerning the Morpho butterfly wing [98, 129, 124, 102].

### 7.1 Numerical methods

In this section, we show how the discrete forms given in Sections 4.2, 5.2.2 and 6.2 are implemented to solve numerically our problem. Some of this implementation requires standard techniques that can be found in the literature $[21,64,113,66]$. In these cases we will not go into the details but restrict ourselves to the main steps with appropriate links to the previous chapters. Where the implementation presents some novel aspects we will be more expansive and provide enough details to convey our ideas. We start by describing our implementation when we use a uniform mesh. We then show the implementation for an adaptive mesh.

### 7.1.1 Uniform mesh

Let $X^{\alpha}$ (resp. $X^{\alpha, \beta}$ ) be a finite dimensional subspace of $H_{\#}^{1}(F)$ with $\operatorname{dim}(X)=$ $N<\infty$ and let $\psi_{i}$ for $i=1, . ., N$, be a basis of $X^{\alpha}\left(\right.$ resp. $\left.X^{\alpha, \beta}\right)$ where $F=\Omega \backslash$ int $\Omega_{3}$

|  | $\alpha, 0-\mathrm{qp}$ |  |  |  | $\alpha, \beta$-qp |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Case | 1 A | 1 B | 2 A | 2 B | 1 A |
| F | $\Omega \backslash \mathrm{int} \Omega_{3}$ |  | $\Omega$ |  | $\Omega \backslash \mathrm{int} \Omega_{3}$ |
| $b_{1}$ | 1 | 1 | 1 | $1 / k^{2}$ | 1 |
| $b_{2}$ | $-2 \mathrm{i} \alpha$ | $-2 \mathrm{i} \alpha$ | $-2 \mathrm{i} \alpha$ | $-\mathrm{i} \alpha / k^{2}$ | $-2 \mathrm{i} \alpha$ |
| $b_{3}$ | 0 | 0 | 0 | $\mathrm{i} \alpha / k^{2}$ | 0 |
| $b_{4}$ | 0 | 0 | 0 | 0 | $2 i \beta_{1}^{0}$ |
| $b_{5}$ | $-\left(k^{2}-\alpha^{2}\right)$ | $-\left(k^{2}-\alpha^{2}\right)$ | $-\left(k^{2}-\alpha^{2}\right)$ | $-\left(k^{2}-\alpha^{2}\right) / k^{2}$ | 0 |
| $b_{6}$ | 0 | $i \alpha$ | 0 | 0 | 0 |
| $b_{7}$ | 0 | 0 | 0 | 0 | $i \beta_{1}^{0}$ |
| $b_{8}$ | -1 | -1 | -1 | $-1 / k_{1}^{2}$ | -1 |
| $b_{9}$ | -1 | -1 | -1 | $-1 / k_{2}^{2}$ | -1 |
| $b_{10}$ | $-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}$ | $-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}$ | $-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}$ | $-2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} / k_{1}^{2}$ | $-2 i \beta_{1}^{0}$ |

Table 7.1: The coefficients $b_{j}$ of equations (7.1) and (7.2) corresponding to Cases $1 \mathrm{~A}, 1 \mathrm{~B}, 2 \mathrm{~A}$ and 2 B for the $\alpha, 0$-quasi periodic transformation ( $\alpha, 0-\mathrm{qp}$ ) and Case 1 A for the $\alpha, \beta$-quasi periodic transformation ( $\alpha, \beta,-\mathrm{qp}$ ).
for Case 1 and $F=\Omega$ for Case 2. Then equations (4.58), (D.63), (D.67), (D.71) and (6.35) present this general form

$$
\begin{align*}
a\left(\psi_{i}, \psi_{j}\right)= & b_{1}\left(\nabla \psi_{i}, \nabla \psi_{j}\right)_{F}+b_{2}\left(\partial_{x} \psi_{i}, \psi_{j}\right)_{F}+b_{3}\left(\psi_{i}, \partial_{x} \psi_{j}\right)_{F} \\
& +b_{4}\left(\partial_{y} \psi_{i}, \psi_{j}\right)_{F}+b_{5}\left(\psi_{i}, \psi_{j}\right)_{F}+b_{6}\left(n_{x} \psi_{i}, \psi_{j}\right)_{\partial \Omega_{3}} \\
& \pm b_{7}\left(\psi_{i}, \psi_{j}\right)_{\Gamma_{ \pm}}+b_{8}\left(T_{+}^{\alpha, 0^{M}} \psi_{i}, \psi_{j}\right)_{\Gamma_{+}}+b_{9}\left(T_{-}^{\alpha, 0^{M}} \psi_{i}, \psi_{j}\right)_{\Gamma_{+}} \tag{7.1}
\end{align*}
$$

since we need to truncate $T_{ \pm}^{\alpha, 0}$. The values of $b_{j}$ for $j=1, \cdots, 9$ and $F$ are given in Table 7.1. We also have equations (4.59), (D.64), (D.68), (D.72) and (6.36) into this general form

$$
\begin{equation*}
\left(\hat{f}, \psi_{j}\right)_{\Gamma_{+}}=\left(b_{10}, \psi_{j}\right)_{\Gamma_{+}} \tag{7.2}
\end{equation*}
$$

where the value of $b_{10}$ is also given in Table 7.1. In our finite element code, the operations are done element wise by looping over all elements of a given triangulation. Hence, each element is mapped to a reference element through an affine transformation [21, 64, 113]. We therefore map each basis $\phi_{i}$ into the local basis
$\tilde{\phi}_{i}$ through this transformation and so equations (7.1) and (7.2) become

$$
\begin{align*}
a\left(\tilde{\psi}_{i}, \tilde{\psi}_{j}\right)= & \tilde{b}_{1}\left(\boldsymbol{\nabla} \tilde{\psi}_{i}, \boldsymbol{\nabla} \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{2}\left(\partial_{x} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{3}\left(\tilde{\psi}_{i}, \partial_{x} \tilde{\psi}_{j}\right)_{F} \\
& +\tilde{b}_{4}\left(\partial_{y} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{5}\left(\tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{6}\left(n_{x} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\partial \Omega_{3}} \\
& \pm \tilde{b}_{7}\left(\tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\Gamma_{ \pm}}+\tilde{b}_{8}\left(T_{+}^{\alpha, 0} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\Gamma_{+}}+\tilde{b}_{9}\left(T_{-}^{\alpha, 0} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\Gamma_{+}} \tag{7.3}
\end{align*}
$$

and

$$
\begin{equation*}
\left(\hat{f}, \tilde{\psi}_{j}\right)_{\Gamma_{+}}=\left(\tilde{b}_{10}, \tilde{\psi}_{j}\right)_{\Gamma_{+}} . \tag{7.4}
\end{equation*}
$$

We need the following steps to implement equations (7.3) and (7.4).

- step 1: Assemble the mass matrix

By denoting

$$
\begin{align*}
K_{i j} & =\tilde{b}_{1}\left(\nabla \tilde{\psi}_{i}, \nabla \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{2}\left(\partial_{x} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{3}\left(\tilde{\psi}_{i}, \partial_{x} \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{4}\left(\partial_{y} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{F} \\
& +\tilde{b}_{5}\left(\tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{F}+\tilde{b}_{6}\left(n_{x} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\partial \Omega_{3}} \pm \tilde{b}_{7}\left(\tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\Gamma_{ \pm}} \tag{7.5}
\end{align*}
$$

we can use numerical integration to calculate the surface and element integrals to get $K_{i j}[70,87,38]$.

- step 2: Assemble the load vector

Similar to step 1, numerical integration can be used to compute $\left(\tilde{b}_{10}, \tilde{\psi}_{j}\right)_{\Gamma_{+}}$.

- step 3: Assemble the DtN operators

Let us denote $B_{i j}=b_{8}\left(T_{+}^{\alpha, 0^{M}} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\Gamma_{+}}+b_{9}\left(T_{-}^{\alpha, 0^{M}} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\Gamma_{+}}$, from equations (4.80) and (6.56), we note that

$$
\begin{equation*}
\left(T_{ \pm}^{\alpha, 0^{M}} \tilde{\psi}_{i}, \tilde{\psi}_{j}\right)_{\Gamma_{ \pm}}=d \sum_{m=-M}^{M} i \beta_{l}^{m} \tilde{\psi}_{i}^{(m)}( \pm B) \overline{\tilde{\psi}_{j}^{(m)}( \pm B)} \tag{7.6}
\end{equation*}
$$

We note that $\tilde{\psi}_{j}^{(m)}$ can be computed as follows using equation (2.53)

$$
\begin{align*}
\tilde{\psi}_{j}^{(m)}( \pm B) & =1 / d \int_{0}^{d} \tilde{\psi}_{j}(x, \pm B) e^{-i 2 \pi m / d x} d x \\
& =1 / d \int_{x_{j}}^{x_{j+1}} \tilde{\psi}_{j}(x, \pm B) e^{-i 2 \pi m / d x} d x \tag{7.7}
\end{align*}
$$

where $x_{j}$ and $x_{j \pm 1}$ delimit the element edge where $\tilde{\psi}$ is different from zero and hence $\operatorname{supp} \psi_{j}(x, \pm B)=\left[x_{j}, x_{j+1}\right]$. We can write

$$
\tilde{\psi}_{j}(x, \pm B)=\sum_{l=0}^{N} c_{l} x^{l}
$$

since $\tilde{\psi}_{j}(x, \pm B)$ is a polynomial of order $N$. Hence, equation (7.7) becomes

$$
\begin{align*}
\tilde{\psi}_{j}^{(m)}( \pm B) & =1 / d \int_{x_{j}}^{x_{j+1}} \sum_{l=0}^{N} c_{l} x^{l} e^{-i 2 \pi m / d x} d x \\
& =1 / d \sum_{l=0}^{N} c_{l} G_{l}\left(x_{j}, x_{j+1}, m, d\right) \tag{7.8}
\end{align*}
$$

where

$$
\begin{equation*}
G_{l}\left(x_{j}, x_{j+1}, m, d\right)=\int_{x_{j}}^{x_{j+1}} x^{l} e^{-i 2 \pi m / d x} d x . \tag{7.9}
\end{equation*}
$$

It can be shown by integrating by parts and by induction that

$$
\begin{align*}
& G_{l}\left(x_{j}, x_{j+1}, l, m, d\right)= \\
& \begin{cases}\frac{x^{l+1}}{l+1}, & \text { if } m=0 \\
\sum_{t=0}^{q} \frac{l!}{(l-t)!(2 i \pi m / d)^{t}} & {\left[x^{l-t} \frac{e^{-2 i \pi m / d}}{-2 i \pi m / d}\right]_{x=x_{j}}^{x=x_{j+1}}} \\
+\frac{l!}{l-(q+1)!} \frac{1}{(2 i \pi m / d)^{q+1}} G_{l-(q+1)}, & \text { otherwise }\end{cases} \tag{7.10}
\end{align*}
$$

for $q \leq l-1$. We can then use equations (7.8), (7.10) and (7.6) to compute $B_{i j}$.

- step 4: Apply the periodicity boundary conditions for all cases and the Dirichlet boundary conditions for Case 1A

Since our problem is find $U_{\alpha, 0}$ (resp. $\left.U_{\alpha, \beta}\right)$ such that $a\left(U_{\alpha, 0}, \psi\right)=(\hat{f}, \psi)_{\Gamma_{+}}$ (resp. $a\left(U_{\alpha, \beta}, \psi\right)=(\hat{f}, \psi)_{\Gamma_{+}}$). By denoting $\left(\hat{f}, \tilde{\psi}_{i}\right)_{\Gamma_{+}}=F_{i}$ and by noting that $a\left(\tilde{\psi}_{i}, \tilde{\psi}_{j}\right)=K_{i j}+B_{i j}$ our solution $\sum_{j=1}^{N} U_{j} \phi_{j}(x, y)$ satisfies

$$
\begin{equation*}
\left(K_{i j}+B_{i j}\right) U_{j}=F_{i} \tag{7.11}
\end{equation*}
$$

for $i, j=1, \cdots N$. For the periodicity boundary condition, let us denote $N_{p}$ the set of nodes $i$ such that $\phi_{i}$ belongs to the boundary $x=0$ or $x=d$ and $N_{O}$ the set of nodes $j$ such that $\phi_{j}$ belongs to $\partial \Omega_{3}$ for Case 1 A . We use the techniques described in [2] which apply the periodicity constraints and the Dirichlet constraints to the nodes $i \in N_{p}$ and $j \in N_{O}$.

- step 5: Solve the system

We can solve the system (7.11) and find the coefficients of the approximate solution to our Helmholtz problem.

- step 6: Refine the meshes such that each element is subdivided into four new elements and repeat step 1 to step 5 until the solution converges. In
our numerical implementation, either the solution is the field or the grating efficiency. If we look at the field, we have convergence when the absolute value of the field coincides to a known or given solution and if it is the grating efficiency, we have convergence when we the grating efficiency from successive refinements present the same $n$ first digits.


### 7.1.2 Adaptive mesh

Once more the operations are done element wise by looping over all elements of a given triangulation. Hence, each element is mapped to a reference element through an affine transformation.

- step 1: Choose a tolerance TOL.
- step 2: Repeat step 1 to 5 for the uniform mesh to solve the direct problem Use the same technique as described with the uniform mesh with equations (4.58), (D.63), (D.67), (D.71), (4.59), (D.64), (D.68) and (D.72) to solve the direct problem since they satisfy the general form given in equations (7.3) and (7.4).
- step 3: Repeat step 1 to 5 for the uniform mesh to solve the dual problem to find $z_{h}$ using Lemma 65.
Since the variational form of the dual problem as given in Lemma 65 has the general form given in equations (7.3) and (7.4) we use the same technique described for the uniform mesh to solve the dual problem.
- step 4: Repeat step 1 to 5 for the uniform mesh to solve the dual problem using Lemma 65 with a higher polynomial order or a finer mesh to give approximation.
- step 5: Compute the upper bound given by equation (5.40) and denoted this upper bound by

$$
\begin{equation*}
I_{e r r}=\sum \rho_{K} w_{K} \tag{7.12}
\end{equation*}
$$

We can use the standard techniques (reference element and numerical integration) $[70,38,113,66]$ to compute equation (5.37). For the flux residual given by equations (5.38) and (5.39), we can compute $T_{ \pm}^{\alpha, 0^{M}} \psi_{j}$ using Definition 4 and equation (7.8) which leaves us with the jump derivative $\left[\nabla \psi_{i} . n\right]$. Let us denote by $K$ and $K^{1}$ two triangles which share an edge $E$, and let the affine transformation which maps $K$ to the reference element (and $\psi_{i}$ to $\tilde{\psi}_{i}$ ), be described by $J \underline{x}+b$ where $J$ is a $2 \times 2$ matrix. Let the affine
transformation which maps $K^{1}$ to the reference element (and $\psi_{i}^{1}$ to $\tilde{\psi}_{i}^{1}$ ), be described by $J^{1} \underline{x}+b$ where $J^{1}$ is a $2 \times 2$ matrix. We then have

$$
\left[\nabla \psi_{i} \cdot n\right]=\left(J^{-T} \nabla \tilde{\psi}_{i}+J^{1-T} \nabla \tilde{\psi}_{i}^{1}\right) \cdot n
$$

since $n^{1}=-n$. Hence by using $y=g(x)$ to represent the curve $\partial K \bigcap \partial K^{1}$, we have

$$
\begin{aligned}
& \int_{E}\left|\left[\nabla \psi_{i} \cdot n\right]\right|^{2} \\
& =\int_{x=x_{0}}^{g\left(x_{0}\right)}\left|\sqrt{1+g^{\prime}(x)^{2}} J^{-T} \nabla \tilde{\psi}_{i}(x, g(x))+\sqrt{1+g^{\prime}(x)^{2}} J^{1-T} \nabla \tilde{\psi}_{i}^{1}(x, g(x))\right|^{2} d x
\end{aligned}
$$

where the edge $E$ is delimited by $g\left(x_{0}\right)$ and $x_{0}$.

- step 6: Let $N_{T}$ be the list of elements $K$ where $\rho(K) w_{K}$ is in decreasing order and such that $\sum_{N_{T}} \rho(K) w_{K} \geq 0.7 I_{\text {err }}$. Refine the mesh in $N_{T}$.
- step 7: Check that the mesh is periodic if it is not then make it periodic.
- Repeat step 1 to step 7 until $I_{e r r} \leq T O L$.

Having briefly described the numerical technique used to solve numerically our Helmholtz problem, we present some numerical results in the following section.

### 7.2 Numerical results

In this section, we start by comparing our numerical results using the $\alpha, 0$-quasi periodic method and the $\alpha, \beta$-quasi periodic method with different results in the literature in order to validate our code.

### 7.2.1 Code validation

We will start by comparing our numerical results from the $\alpha, 0$-quasi periodic transformation with the experimental results given in [95]. In this first example, we have a perfectly conducting echelette grating given by Figure 7.1 used in a $-m$ Littrow mount. For this grating, the $m$ th-order diffracted wave propagates backwards in the opposite direction to the incident wave and most of the energy is concentrated into the $m$ th-order diffracted wave provided that

$$
2 \sin \theta=m \frac{\lambda}{d}
$$

where $\theta$ is the angle of incidence, $\lambda$ is the wavelength and $d$ is the period of the grating [95, 79]. The facet angle of the grooves is called the blaze angle and
the wavelength corresponding to the maximum energy is called blaze wavelength $[95,79]$. This type of grating is used in laser system design for wavelength selective reflectors and in spectroscopic devices. In our case, the blaze angle is equal to 5 degrees, and $m=1$.

By doing so, we want to investigate the variation of the reflection efficiency $R_{-1}$ given by equation (4.71), when the ratio $\lambda / d$ varies from 0 to 2 . We solve numerically for the reflection efficiency $R_{-1}$ using the $\alpha, 0$-quasi periodic method described by equations (4.58), (D.63), (4.59) and (D.64) where $B=0.15$, the degree of the polynomial basis is $p=3$ for both TE and TM cases and the degrees of freedom (dof given by equation (B.7)) is 3577 for the TE case and 14065 for the TM case. In Figure 7.2, the experimental data is given by the discrete points and our numerical results are given by the full line (for the TE case) and the dashed line (for the TM case). We can conclude from Figure 7.2 that we have a good agreement between the numerical and experimental results outside the region of resonance ( $\beta_{j}^{n}=0$ in equation (2.43) that is $\lambda=2 d /(2 n+1)$ ).


Figure 7.1: An echelette grating with an apex angle equal to $\pi / 2$, blaze angle $a=5$ degrees and period $d$.

We now consider a second example where we have a perfectly conducting cylinder grating given by Figure 7.3 with a radius $a$ and wavenumber $k_{1}$ outside the scatterer. This semi-analytical approach uses scattering theory, involving multiple scattering between adjacent cylinders, with an eigenfunction expansion in cylindrical coordinates that assumes radial symmetry. This restricts this approach to problems with cylindrical geometry. We will compare a semi-analytic solution given in [81] with the numerical solution obtained by using the $\alpha, 0$-quasi periodic method for the TE case. This type of scattering geometry is useful in helping to develop an understanding of the loads on offshore platforms and wave power devices. We have implemented the series solution in [81], using $N=61$ and $M=6$


Figure 7.2: A perfectly conducting echelette grating (Case 1) as shown in Figure 7.1. Comparison between the reflection efficiency of order -1 from the $\alpha, 0$ quasi periodic method (dashed line for TM and full line for TE) and the experimental results (circle for TM and square for TE) in [95].
terms in each expansion. In our numerical results using the $\alpha, 0$-quasi periodic method described in equations (4.58) and (4.59), we chose the polynomial order $p=3$ and the number of degrees of freedoms (dof) as 14112. In this numerical experiment, we chose a wavenumber $k=\pi$, angle of incidence $\theta=\pi / 4$, radius of the cylinder $a=0.1$, period $d=1$ and upper domain boundary $y=B=1$. In Figures 7.4 and 7.5 we compare the spatial distribution of the imaginary and real parts of the solution for each method and we see a good agreement between both approaches.


Figure 7.3: Cylinder grating.
In this third example, the grating is composed of two dielectric transmitting cylinders as shown in Figure 7.6. We want to find the reflection efficiency of order zero $\left(R_{0}\right)$ as defined in equation (4.71) for the TM case (Case 2B) where we let the ratio $\lambda / d$ vary from 0.7 to 1 . We use the $\alpha, 0$-quasi periodic method to solve the problem numerically from equations (D.71) and (D.72). We then compare this with the result obtained using the lattice sum technique given in [128]. This latter method is found in 2-D photonic bandgap structures in which the cylindrical harmonic expansion is used so that the array of cylinders are viewed as a spatially periodic lattice and this allows for coupling between adjacent cylinders. Therefore, this method is limited to scattering with polygonal geometry. By using a polynomial basis of degree 4 with 21633 degrees of freedom, we have plotted $R_{0}$ as a function of $\lambda / d$ in Figure 7.7. We can conclude from this figure that our numerical results from the $\alpha, 0$-quasi periodic method are in good agreement with


Figure 7.4: A periodic grating consisting of a perfectly conducting cylindrical array of period $d=1$ (see Figure 7.3). The spatial distribution of the imaginary part of the wave amplitude is shown for (a) semi-analytic solution in [81] and (b) the $\alpha, 0$-quasi periodic method. The radius of the cylinder is $a=0.1$, the wavenumber is $k=\pi$, and the angle of incidence is $\theta=\pi / 4$.
(a)



Figure 7.6: Double layered dielectric transmitting cylinders.


Figure 7.7: A two layer, periodic grating consisting of dielectric transmitting cylinders in a TM field (Case 2B) (see Figure 7.6). Comparison between the reflection efficiency of order 0 from the $\alpha, 0$-quasi periodic method (full line) and the lattice sum technique (dashed line) [128]. The reflection efficiency is shown as a function of the ratio of the wavelength of the incident field $(\lambda)$ to the lattice period $(d)$.
those from the lattice sum technique. It is also of interest to examine the effects that the degree of the polynomial basis $p$, the number of Fourier terms $N$ and the number of degrees of freedom dof have on the computational cost in using the $\alpha, 0$ quasi periodic method. In the example below we have set $d=1.085, \lambda=1.55$ and have kept two of the parameters fixed $\{p, N, d o f\}$ and let one of the parameters vary. We plot the computational cost against the parameter which varies. We can conclude from Figures 7.8(a), 7.10(a) and 7.9(a) that the computational cost grows faster with the degrees of freedom and the order of the polynomial basis than the number of Fourier terms. In Figure 7.8(b), we let $\mathbf{N}=\{11,15,21,31\}$ and we plot the variation of $R_{0}$ with respect to $N=\mathbf{N}_{i}$ for $i \in\{1,2,3,4\}$

$$
\begin{equation*}
\mathrm{k} R_{0}\left(\mathbf{N}_{i}\right)-R_{0}\left(\mathbf{N}_{i-1}\right) \mathrm{k} \tag{7.13}
\end{equation*}
$$

against $\mathbf{N}_{i}$ using $p=3$ and dof $=12193$. We do the same in Figure 7.10(b) but we keep fixed $N=15$, dof $=12193$ and vary $p \in \mathrm{P}$ such that $\mathrm{P}=\{3,4,5\}$ and we plot the variation of $R_{0}$ with respect to $\mathrm{P}_{i}$ for $i \in\{1,2,3\}$

$$
\begin{equation*}
\mathrm{k} R_{0}\left(\mathrm{P}_{i}\right)-R_{0}\left(\mathrm{P}_{i-1}\right) \mathrm{k} \tag{7.14}
\end{equation*}
$$

against $\mathbf{P}_{i}$. In Figure 7.9(b) we keep fixed $N=15, p=3$ and vary dof $\in \mathbf{D}$ such that $\mathrm{D}=\{781,3073,12193,48577\}$ and we plot the variation of $R_{0}$ with respect to $\mathrm{D}_{i}$

$$
\begin{equation*}
\mathrm{k} R_{0}\left(\mathrm{D}_{i}\right)-R_{0}\left(\mathrm{D}_{i-1}\right) \mathrm{k} \tag{7.15}
\end{equation*}
$$

against $\mathrm{D}_{i}$. We conclude from Figures 7.8(b), 7.9(b) and 7.10(b) that the accuracy of the diffraction efficiency is less affected by changes in the number of Fourier terms as opposed to changes in the order of the polynomial basis and the number of degrees of freedom.

In the final example in this section, we consider the perfectly conducting echelette grating (Case 1) as shown in Figure 7.11. We want to find the reflection efficiency of order 0 ( $R_{0}$ defined in equation (4.71)) using equations (4.58), (4.59), (D.63) and (D.64) when we let the ratio $l / d$ vary, where $l$ is the grating depth and $d$ is the grating period. We will use both the $\alpha, 0$-quasi periodic method and the $\alpha, \beta$-quasi periodic method to solve the problem numerically. We then compare this with results from method of variation of boundaries given in [25] in Figure 7.12. The method of variation of boundaries is a perturbation technique based on expansions in a small parameter $(l)$ which corresponds to the height of the echelette surface. These echelette designs are used in solar energy absorbers and antireflecting surfaces. We can conclude from Figure 7.12 that our numerical results from both the $\alpha, 0$-quasi periodic method (red diamond for both TE and TM ) and the $\alpha, \beta$-quasi periodic method (blue cross for both TE and TM ) are in good agreement with the numerical result using the method of variation of boundaries (black circle for both TE and TM). The agreement is so good that the points lie precisely on top at each other. In a similar manner to the previous example it is of interest to study the convergence of the error in the calculation of the


Figure 7.8: A double layered, dielectric transmitting, periodic grating consisting of cylinders interacting with a TM field (Case 2B) (see Figure 7.6). (a) Dependence of the computational cost in calculating $R_{0}$ using the $\alpha, 0$-quasi periodic method (on a uniform mesh) on the number of Fourier terms $N$ and (b) accuracy in calculating $R_{0}$ given by equation (7.13) when the number of Fourier terms varies from $\mathbf{N}_{i-1}$ to $\mathrm{N}_{i}$.


Figure 7.9: A double layered, dielectric transmitting, periodic grating consisting of cylinders interacting with a TM field (Case 2B) (see Figure 7.6. (a) Dependence of the computational cost in calculating $R_{0}$ using the $\alpha, 0$-quasi periodic method (on a uniform mesh) on the number of degrees of freedom dof and (b) accuracy in calculating $R_{0}$ given by equation (7.15) when the degrees of freedom varies from $\mathrm{D}_{i-1}$ to $\mathrm{D}_{i}$.


Figure 7.10: A double layered, dielectric transmitting, periodic grating consisting of cylinders interacting with a TM field (Case 2B) (see Figure 7.6. (a) Dependence of the computational cost in calculating $R_{0}$ using the $\alpha, 0$-quasi periodic method (on a uniform mesh) on the polynomial degree $p$ and (b) accuracy in calculating $R_{0}$ given by equation (7.14) when the degree of the polynomial basis varies from $\mathrm{P}_{i-1}$ to $\mathrm{P}_{i}$.


Figure 7.11: A perfectly conducting, echelette grating with depth $l$ and period $d$.


Figure 7.12: A perfectly conducting, echelette grating with depth $l$ and period $d$ (see Figure 7.11) interacting with (a) a TE field (Case 1A), comparison between the reflection efficiency of order 0 from the $\alpha, 0$-quasi periodic method (diamond), the $\alpha, \beta$-quasi periodic method (cross) and the method of variation of boundaries (circle) and (b) a TM field (Case 1B), comparison between the reflection efficiency of order 0 from the $\alpha, 0$-quasi periodic method (diamond), the $\alpha, \beta$-quasi periodic method (cross) and the method of variation of boundaries (circle).
grating efficiencies as the parameters in the algorithm are varied. In the following examples, we fix the pitch of our grating to be $l / d=0.05$, perform uniform mesh refinement and study the convergence of the efficiencies of orders 0,1 and 2 of the reflected waves. The degree of the polynomial basis $(p)$ is varied between 3 and 7 and we calculate the error in calculating the grating efficiency as

$$
\begin{equation*}
\left|R_{j}^{E}-R_{j}(\alpha, \beta)\right| \tag{7.16}
\end{equation*}
$$

where $R_{j}^{E}$ is the numerical result given in [25] (used as the exact values) and $R_{j}(\alpha, \beta)$ is the numerical result obtained from using the $\alpha, \beta$-quasi periodic method. We then plot this error against the degree of the polynomial basis $p$. The results shown in Figure 7.13 show that as the degree of the polynomial basis decreases, the mesh must be refined in order to reproduce the numerical results given in [25] which used the method of variation of boundaries. We note that when the number of vertices is bigger than 150, and the degree of the polynomial basis is greater than 4 , there is no significant difference in reproducing the reference solution in [25]. In the next section, we investigate how efficient the $\alpha, 0$-quasi periodic method is compared with the $\alpha, \beta$-quasi periodic method for different ranges of the wavenumber $k$.

### 7.2.2 Variation with $k$ of the $\alpha, 0$ and the $\alpha, \beta$-quasi periodic methods

By introducing the $\alpha, \beta$-quasi periodic method in Chapter 6 it is envisaged that this will provide a more accurate and stable method for addressing the grating problem. In particular, the $\alpha, \beta$-quasi periodic method should be able to overcome the high wavenumber numerical instabilities that is normally observed when using finite element methods $[59,16]$. A series of numerical examples below investigate this hypothesis. To begin with we will examine the simplest scenario when in fact the domain is scatterer free and hence we know precisely that the transmitted energy $E_{t}$, as given by Definition 54, is equal to one. We will study a TM wave (Case 1B) described by equations (D.63) and (D.64). We define the relative error $\epsilon(\alpha, 0)$ (resp. $\epsilon(\alpha, \beta)$ ) for the $\alpha, 0$-quasi periodic method (resp. $\alpha, \beta$-quasi periodic method) when we compute the transmitted energy $E_{t}$, as $\epsilon(\alpha, 0)=\left|E_{t}(\alpha)-E_{t}\right|$ and $\epsilon(\alpha, \beta)=\left|E_{t}(\alpha, \beta)-E_{t}\right|$. We then compare the logarithm of $\epsilon(\alpha, 0)$ with $\epsilon(\alpha, \beta)$ for a fixed number of nodes and a fixed degree of the polynomial basis. We can see from Figure 7.14 that the $\alpha, \beta$-quasi periodic method is far more accurate than the $\alpha, 0$-quasi periodic method. The $\alpha, \beta$-quasi periodic method has a consistent accuracy of $10^{-14}$ for a wide range of wavenumbers. In fact to achieve a relative error of $10^{-14}$ at $k=20 d$ the $\alpha, \beta$-quasi periodic method requires 625 dof whereas the $\alpha, 0$-quasi periodic method requires 9409 dof. It is known for single scattering that there is an instability in numerical methods when we have high wavenumbers $[59,16]$. Here we observe however that not only is the $\alpha, \beta$-quasi periodic method


Figure 7.13: A perfectly conducting grating (see Figure 7.11) interacting with a TE field (Case 1A). The numerical results given in [25] are used as the exact values. The errors in the computation of efficiencies of the diffracted waves are plotted against the number of vertices $d o f_{h}$, as defined in Section B.3.3, using polynomial bases of degrees $p=3$ (full line) up to $p=7$ (dashed dot line) in (a) for the efficiency of order 0 , (b) for the efficiency of order 1 and (c) for the efficiency of order 2.
reliably accurate, it is also able to cope with high wavenumbers where the standard approach fails.

In this second example of this section, we consider again the perfectly conducting echelette grating as shown in Figure 7.11. As before we use the results given in [25] (denoted $R_{j}^{E}$ ) as the exact values of the diffraction efficiency $R_{j}$ for $j=0,1$ and 2 . We then use a uniform mesh and we calculate the diffraction efficiency obtained with the $\alpha, 0$-quasi periodic method (denoted $R_{j}^{\alpha, 0}$ ), using equations (4.58) and (4.59), and with the $\alpha, \beta$-quasi periodic method (denoted $R_{j}^{\alpha, \beta}$ ) using equations (6.35) and (6.36). We examine the absolute error $|\epsilon(\alpha, 0 / \alpha, \beta)|=\left|R_{j}^{E}-R_{j}^{\alpha, 0 / \alpha, \beta}\right|$ and plot this error as a function of the degree of the polynomial basis $p$ (keeping $h$


Figure 7.14: A scatterer free calculation using a TM wave (Case 1B). Comparison of the logarithm of the relative error in computing the transmitted energy from the $\alpha, 0$-quasi periodic method (full line) and the $\alpha, \beta$-quasi periodic method (dashed line). The degrees of freedom is fixed for both cases and the wavenumber $k$ is varied.
fixed and the same dof in each method). We can see from Figure 7.15 that there is no significant difference in the rate of convergence of the numerical solutions obtained from the two quasi periodic methods with the echelette grating when the wavenumber $k=1 / 0.4368$ as given in [25] is not large. In the following ex-


Figure 7.15: A perfectly conducting echelette grating (see Figure 7.11) interacting with a TE field (Case 1A). Comparison between the absolute error from the $\alpha, 0$ (dashed line) and the $\alpha, \beta$-quasi periodic methods (full line) in calculating (a) the reflection efficiency of order $0\left(R_{0}\right)$, (b) the reflection efficiency of order $1\left(R_{1}\right)$ and (c) the reflection efficiency of order $2\left(R_{2}\right)$.
ample, we consider the perfectly conducting cylinder grating shown in Figure 7.3 interacting with a TM wave (Case 1B). Here the exact solution is not computable, and so we compute the transmitted energy (given by Definition 54) and keep refining uniformly the mesh until this converges (that is, when the first four digits are the same). We denote this converged value by $E_{t}$. We define the relative error $\epsilon(\alpha, 0)$ (resp. $\epsilon(\alpha, \beta))$ for the $\alpha, 0$-quasi periodic method described by equations (D.63) and (D.64) (resp. for the $\alpha, \beta$-quasi periodic method described in a similar way to equations (6.35) and (6.36)) as $\epsilon(\alpha, 0)=\left|E_{t}(\alpha, 0)-E_{t}\right| / E_{t}$ (resp. $\left.\epsilon(\alpha, \beta)=\left|E_{t}(\alpha, \beta)-E_{t}\right| / E_{t}\right)$. We then fix the number of degrees of freedom at the
point where $\epsilon(\alpha, \beta)=0$ in the $\alpha, \beta$-quasi periodic method and examine how each method performs as the wavenumber is varied. It can be seen in Figure 7.16(a) that the $\alpha, \beta$-quasi periodic method is considerably more accurate and requires far fewer computational resources than the $\alpha, 0$-quasi periodic method. To achieve a relative error of $10^{-14}$ at $k=30 d$ the $\alpha, \beta$-quasi periodic method requires 3600 dof whereas the $\alpha, 0$-quasi periodic method requires 55872 dof. We now use the same geometry but this time set the cylinder as a transmitting dielectric grating in a TE field (Case 2A). As above, we compare the logarithm of the relative error in computing the transmitted energy where we fix the number of degrees of freedom at the point where the $\alpha, \beta$-quasi periodic method converges and examine how each method performs as the wavenumber is varied. A similar conclusion is obtained when we examine Figure $7.16(\mathrm{~b})$ where it can clearly be seen that the $\alpha, \beta$-quasi periodic method outperforms the $\alpha, 0$-quasi periodic method by some degree. In fact, to achieve a relative error of $10^{-14}$ the $\alpha, \beta$-quasi periodic method requires 6433 dof whereas the $\alpha, 0$-quasi periodic method requires 101761 dof. In Table 7.2 , we show the effect of each uniform refinement on the number of elements $N_{K}$, the $d o f_{p}$ when $p=4$ is fixed and the total degrees of freedom dof. This shows us that, since each element is divided into four new elements, the dof becomes four times larger. Hence, the computational cost becomes significantly large for the refined mesh compared to the unrefined one. This makes the computational cost from using the $\alpha, \beta$-quasi periodic mesh cheaper since it converges with fewer nodes as opposed to the $\alpha, 0$-quasi periodic. The previous sections showed the

| Case | number of refinement | $N_{K}$ | $d o f_{p}$ | $d o f$ |
| :--- | :---: | :---: | ---: | ---: |
| Case 1B | 3 | 768 | 3168 | 3600 |
|  | 4 | 3072 | 12480 | 14112 |
|  | 5 | 12288 | 49536 | 55872 |
| Case 2A | 3 | 1408 | 5696 | 6433 |
|  | 4 | 5632 | 22656 | 25537 |
|  | 5 | 22528 | 90368 | 101761 |

Table 7.2: The variation of the number of elements $N_{k}$, the degrees of freedom from p-refinement $d o f_{p}$ and the degrees of freedom dof defined in Section B.3.3 when we make uniform refinement with a cylinder grating as shown in Figure 7.3 for Case 1B and Case 2A.
benefits of using the $\alpha, \beta$-quasi periodic method and throughout the computations were conducted on a uniform mesh. We now want to investigate the benefits that can be obtained by using an adaptive grid and in particular where this adaptivity is driven by the Dual Weighted Residual method constructed in Chapter 5.


Figure 7.16: Cylinder grating as shown in Figure 7.3. Comparison of the logarithm of the relative error in computing the transmitted energy from the $\alpha, 0$-quasi periodic method (full line) and $\alpha, \beta$-quasi periodic method (dashed line) (a) for Case 1B and (b) for Case 2A.

### 7.2.3 Comparison with the DWR method

In the following example, we consider the transmitting dielectric lamellar grating as shown in Figure 7.17. This type of grating is used in modeling multiscale phenomena grating problems, and has been studied in [13] using a hybrid approach that combines a perfectly matching layer technique and an adaptive finite element method driven by a global a posteriori error estimate and it can be applied in solving optimal design problems. For our investigation we will fix the wavenumbers as $k_{1}=2 \pi$ and $k_{2}=(0.22+6.71 i) 2 \pi$, the angle of incidence $\theta=\pi / 6$ and the period $d=1$. We will consider a TM field (Case 2B) as described by equations (D.71) and (D.72), and compute the reflection efficiency of order zero $\left(R_{0}\right)$ as given by equation (4.71). Since we will use the DWR method, we employ Lemma 65 and equation (5.40) to find the dual solutions and the error bounds. We follow the algorithm in Section 7.1.2 and we choose a tolerance $T O L=10^{-4}$. To provide a basis for a relative error we use the global method in [13] with 201205 degrees of freedom which gives $R_{0}^{E}=0.8484815$. We then compare the relative error using the global a posteriori error estimate in [13] and the DWR method defined by $\epsilon(D W R /$ Global $)=\left|R_{0}^{D W R / G l o b a l}-R_{0}^{E}\right| / R_{0}^{E}$. We can see from Figure 7.18 that the DWR method converges faster than the global a posteriori error method studied in [13]. We also note that the indicative computed error $I_{\text {err }}$ given in equation (7.12) decreases monotonically which shows the convergence of our DWR method. We choose to stop at a relative error of $10^{-6}$ since the mesh becomes very irregular and this will lead to a numerical instability if we keep refining [3, 20]. In addition, since we focus on the local error, there is a pollution coming from the global error. In practice, we can avoid this problem by refining the mesh in the neighborhood of the high stress gradients; that is, at the singularities in the geometry. We cannot always guarantee that the areas of high stress gradients coincide with the areas of interest and this may lead to pollution error. Hence, we can improve the goaloriented error estimation method by using a proper balance between the local error and the global error so that the mesh is refined to assure a high level of accuracy of the quantity of interest and at the same time we do not underestimate the effect of the global error [51]. We can also compare the DWR method with the use of a uniform mesh. As above, we compare the relative error using the uniform mesh and the DWR method defined by $\epsilon(D W R /$ Uniform $)=\left|R_{0}^{D W R / U n i f o r m}-R_{0}^{E}\right| / R_{0}^{E}$. We can see from Figure 7.19 that the error associated with the DWR method does not decrease monotonically unlike that associated with the uniform mesh. However when the dof $>10^{4}$ the DWR converges faster and requires fewer degrees of freedom than using the uniform mesh when we use with the $\alpha, 0$-quasi periodic method.


Figure 7.17: Transmitting dielectric lamellar grating.


Figure 7.18: The transmitting dielectric lamellar grating shown in Figure 7.17 in a TM field (Case 2B). The relative error in computing the reflection efficiency $R_{0}$, using the global a posteriori error estimate in [13] (dashed line), the DWR method (straight line) and the indicative computed error $I_{\text {err }}$ (dotted line) given by equation (7.12) which shows an upper bound of the discretisation error in solving numerically the Helmholtz problem.


Figure 7.19: The transmitting dielectric lamellar grating shown in Figure 7.17 in a TM field (Case 2B). The relative error in computing the reflection efficiency $R_{0}$, using uniform mesh (full line) and the DWR method (dashed line).

### 7.2.4 Application of the DWR to a scatterer with a complex geometry: the butterfly wings

One of the main reasons for using a finite element approach is of course its ability to tackle any prescribed geometry. The examples we have discussed so far have all dealt with regular geometries and this of course has allowed us to compare our method with exact or semi-analytic methods that capitalise on this regularity. In order to develop more sophisticated grating structures that can help push this technology forward it is necessary to be able to investigate irregular geometries. One recent exploration of what could be achieved by freeing up these geometrical constraints has been inspired by a naturally occurring periodic diffraction grating. When scattered by light, the Morpho butterfly wing produces colour and this allows a dynamical control of light flow and wavelength interaction [125]. The scattering of the Morpho butterfly wing has inspired applications in biomimetics such as gas sensors, electronic display screens and paints for cars [98, 129, 124, 102]. A cross-section of the butterfly wing geometry,which is a transmitting dielectric grating, has been studied experimentally in [125] using a focused laser technique to examine the absolute reflectivity and transmittivity. We have scanned one period of this image and extracted the coordinates of the grating interface which we subsequently smoothed using a Savitzky Golay filter [100, p. 183,644-645] (Figure 7.20). We have set the wavenumber inside the butterfly wing as $k_{2}=(1.57+0.06 i) 2 \pi / 0.455$ and outside the butterfly wing as $k_{1}=2 \pi / 0.455$; which belongs to the range of values given in [125]. We impose a TM field (Case 2B), as described by equations (D.71) and (D.72), and use our DWR method. Therefore, we need Lemma 65 and equation (5.40) to find the dual solutions and the error bounds. The degrees of freedom used is $d o f=11557$, with $N=15$ Fourier terms and the degree of the polynomial basis $p=3$. We choose a tolerance $T O L=0.001$ in the algorithm in Section 7.1.2 to reproduce the imaginary and real parts of the magnetic field. The final indicator computed error $I_{e r r}$ as given in equation (7.12) shows that the discretisation error in solving the Helmholtz problem was kept under $I_{e r r}=3.0312310^{-4}$ so that the level of accuracy in our numerical solutions is of this order. This numerical result is a first step in showing that we can use the DWR method to solve numerically the diffraction phenomenon using a single scale transmitting dielectric grating from the butterfly wing. At this stage it is difficult to compare this result with the experimental data $[125,98,130]$ as there are no plots of the field given. In our study, we produce the solution field and there are no such images in those literature. We could produce the reflection coefficient spectrum but the experimental situation in [125, 98, 130] is far more complex since the butterfly wing is pre-treated in a liquid and therefore we need to develop a more sophisticated model.


Figure 7.20: Reproduction of the image of the Morpho butterfly wing over one period $d$.


Figure 7.21: A transmitting dielectric Morpho butterfly wing grating (Case 2B) as shown in Figure 7.20. The spatial distribution of the magnetic field is shown for (a) the imaginary part and (b) the real part.

### 7.3 Summary and Conclusion

In this chapter, we described the implementation of our finite element methods for solving the periodic grating Helmholtz problem. In the case where the implementation presented some novel aspects, we gave details to expose this novelty. We gave details of both the uniform mesh and the adaptive mesh algorithm as this provides a computational basis for comparing the two approaches. Our first task was to show the validity of the our implementation by comparing our results with those from the literature. Our numerical results presented in Section 7.2.1 showed good agreement with the experimental results in [95], the analytical approach in [81], the lattice sum technique in [128] and the method of variation of boundaries in [25]. We also found that the order of the polynomial basis and the degrees of freedom had a significant impact on the accuracy of the diffraction efficiency and on the computational cost as compared to the number of the Fourier terms. We then compared the $\alpha, 0$-quasi periodic approach to the $\alpha, \beta$-quasi periodic method to investigate any improvements in accuracy, stability and computational cost. In Section 7.2.2, we found that the results from the $\alpha, \beta$-quasi periodic transformation and $\alpha, 0$-quasi periodic transformation converged at the same rate for the echelette grating. This result changed for larger wavenumbers for a perfectly conducting cylindrical grating. It transpired that the $\alpha, \beta$-quasi periodic method converges at a far lower number of degrees than the $\alpha, 0$-quasi periodic method (for example to achieve a relative error of $10^{-14}$ for $k=30 d$ the $\alpha, \beta$-quasi periodic method required 3600 dof whereas the $\alpha, 0$-quasi periodic method required 55872 dof) for Case 1B and we noted the same behaviour for Case 2A (for example to achieve a relative error of $10^{-14}$ for $k=30 d$ the $\alpha, \beta$-quasi periodic method required 6433 dof whereas the $\alpha, 0$-quasi periodic method required 101761 dof). We then investigated the merits of using an adaptive grid driven by a Dual Weighted Residual (DWR) method. In Section 7.2.3, we concluded that the DWR method converged faster and required fewer degrees of freedom than the global a posteriori error estimate proposed in [13]. Finally, we took advantage of the geometrical freedom that the finite element method allows and examined a naturally occurring, periodic diffraction grating in the form of a butterfly wing. Its diffraction properties have previously been experimentally measured but this is the first attempt to mirror those results using a finite element approach. The TM field was successfully calculated and showed the complex interaction between the scatterer and the field. There are still some open questions regarding this approach to numerically solving these diffraction problems. It will be interesting to investigate numerically the sensitivity of the grating efficiency to small changes in the geometry of the grating profile. This would be important when we consider the inverse problem where we have a desired grating efficiency spectrum and we wish to construct the grating profile that would give rise to the spectrum. Also, we still do not have an analytical relation to justify our choice of parameters such as the number of the Fourier terms, the truncation of the domain $B$, the order of the polynomial
basis and the mesh size $h$. It transpires that the accuracy of the numerical results is not significantly affected by the truncation of the $\operatorname{DtN}$ operators. It would be good therefore to have an analytical result to show that there is advantage in using finite element methods because fewer Fourier coefficients are needed.

## Chapter 8

## Conclusion

Diffraction gratings have been used for example in crystalline silicon solar cells [94], gas sensors [98], high intensity colour displays [36], and in medical imaging through x-ray $[62,132]$. The gratings are also used on credit cards or other identification cards as a security measure, providing an image that can be read by an optical scanner [19]. In order to develop these technologies further it would be useful to have fast and reliable mathematical models so that putative designs can be constructed. The appropriate model is given by the Helmholtz equation but this needs to be solved numerically even for fairly simple diffraction gratings.

We started by describing the physical and mathematical aspects of the problem of diffraction when an electromagnetic wave interacts with a periodic grating. From Maxwell's equations, it can be shown that the problem can be decomposed into two elementary mathematical problems which are the transverse magnetic (TM) and the transverse electric (TE) Helmholtz problems. For each problem, the grating can be perfectly conducting or transmitting and so we studied four cases. In order to keep this thesis at a reasonable length, we have just shown the results for Case 1A (TE case for the perfectly conducting grating) when we studied the a priori error estimate using the $\alpha, 0$-quasi periodic method and the results for Case 1A (TE case for the perfectly conducting grating) in the main body of the thesis. We truncated the domain with respect to the $y$ direction and introduced the Dirichlet to Neumann ( DtN ) maps. We then formulated the boundary value problems corresponding to the truncated domain, where the incident wave was included via the boundary conditions. We then considered an equivalent but alternative formulation that incorporated the incident wave via an inhomogeneous forcing term (with compact support). We derived a regularity result for multiple scattering that showed an explicit dependence on the wavenumber $k$ and the forcing term $f$. This regularity result was then used to prove the well-posedness of the variational formulation. It also gave us a hold on the convergence and the stability of the solution when we later solved numerically the scattering problems. In fact, if we let $h$ denote the maximum mesh size of our elements, and $p$ be the highest order of the finite element basis, since we know explicitly the dependence of the
regularity results on the wavenumber $k$, then the a priori error estimate presented a power factor of $k \frac{h}{p}$. Hence, we can choose the mesh size $h$ and the order of the polynomial basis $p$ for a given wavenumber $k$ to balance the computational time and the accuracy of our approximated solution.

We then studied the $\alpha, 0$-quasi periodic method which transformed the diffraction problem from the $\alpha$-quasi periodic space to a periodic space. We gave a variational formulation on an appropriate Sobolev space and demonstrated the well-posedness of this continuous problem. We then investigated the a priori error estimates that arise through discretisation and through truncating the DtN map. In this way, we made sure that the approximate solution is unique. In Theorem 62, we derived an a priori error estimate for Case 1A that arises due to the discretisation and the truncation of the DtN map. For the discretisation, we derived an explicit dependency of this error on $h, p$ and $k$. It transpired from our a priori error analysis that the number of Fourier terms used, when we solve the problem numerically, plays a minor role compared to our choice of the mesh size $h$ and the order of the polynomial basis $p$.

We then derived an a posteriori error estimate that arises when we discretise the Helmholtz problem using the Dual Weighted Residual method which consisted in estimating a particular linear functional of this error. We started by recalling the direct problem and introducing the dual problem. We then estimated this linear functional of the error in equation (5.40) and we showed in Section 5.5.2 that this upper bound can be evaluated. This allowed us to perform an automatic mesh adaptation based on the local error indicators $\rho_{K}$ and $w_{K}$ as defined in equations (5.41) and (5.42) in Chapter 7.

It has been reported for single scattering that when we have a high wavenumber $k$, numerical methods such as the finite element method become unstable. In an attempt to circumvent this problem we transformed the diffraction problem by writing the solution $U$ as a product of the analytical solution of the scatterer free Helmholtz problem with an unknown solution $U_{\alpha, \beta}$. We were able to show in fact that the dependence of this transformed wave equation on the wavenumber $k$ is of order 1 compared to order 2 for the original solution $U_{\alpha, 0}$. In order to test our analytical results a series of numerical investigations was undertaken in Chapter 7 . We started by describing the implementation of our finite element method and gave details of both the uniform mesh and the adaptive mesh algorithm so that we could compare the two approaches. This was followed by the validation of our implementation where we compared our results with those from the literature. In Section 7.2.1, we got good agreement with the experimental results in [95], the analytical approach in [81], the lattice sum technique in [128] and the method of variation of boundaries in [25]. We also demonstrated that the order of the polynomial basis and the degrees of freedom had a significant impact on the accuracy of the diffraction efficiency and on the computational cost as compared to the number of the Fourier terms. In Section 7.2.2, we compared the $\alpha, 0$-quasi periodic approach to the $\alpha, \beta$-quasi periodic method. We found that the results
from the $\alpha, \beta$-quasi periodic transformation and $\alpha, 0$-quasi periodic transformation converged at the same rate for the echelette grating in the low wavenumber regime. This result changed for larger wavenumbers for a perfectly conducting cylindrical grating and a transmitting dielectric grating. In these cases, the $\alpha, \beta$-quasi periodic method converged at a far lower number of degrees of freedom than the $\alpha, 0$-quasi periodic method. We then investigated the advantages of using an adaptive grid driven by a Dual Weighted Residual (DWR) method. In Section 7.2.3, we concluded that the DWR method converged faster and required fewer degrees of freedom than the global a posteriori error estimate proposed in [13]. Finally, we took advantage of the geometrical freedom that the finite element method allows and examined a naturally occurring, periodic diffraction grating in the form of a butterfly wing. Its diffraction properties have previously been experimentally measured but this is the first attempt to mirror those results using a finite element approach. The TM field was successfully calculated and showed the complex interaction between the scatterer and the field.

Modeling of wave interaction with a diffraction grating is a very active research topic, and there are lots of open problems in this field. For example, how the efficiency of the diffraction grating is affected by a small change in the geometry of the scatterer. From a design perspective it is important to investigate the inverse problem of achieving a desired device performance using a model driven approach. In this area, the theory of uniqueness and existence is still not fully understood. Another open question is the dependence of the numerical results on the extent of the truncated domain. We can also think of improving the computational cost of solving the problem by performing parallel computation. It will be also interesting to extend our numerical methods to dynamic gratings so that we can investigate acousto-optic devices for tunable optical filters [108]. Finally, it will be interesting to extend our methods to the three dimensional Helmholtz problem for biperiodic gratings.

## Bibliography

[1] S. Adjerid, J. Flaherty, and I. Babuska. A posteriori error estimation for the finite element method-of-lines solution of parabolic problems. Math. Models Meth. Appl. Sci., 9(2):261-286, 1999.
[2] M. Ainsworth. Essential boundary conditions and multi-point constraints in finite element analysis. Comp. Meth. Appl. Mech. Eng., 190(48):6323-6339, 2001.
[3] B. Aksoylu and M. Holst. Optimality for multilevel preconditioners for local mesh refinement in three dimensions. SIAM J. Numer. Anal., 44(3):10051025, 2006.
[4] T. Arens and T. Hohage. On radiation conditions for rough surface scattering problems. IMA J. App. Maths, 70:839-847, 2005.
[5] T. Arens and A. Kirsch. The factorization method in inverse scattering from periodic structures. Inverse Problems, 19:1195-1211, 2003.
[6] Axelsson and V. A Barker. Finite element of boundary value problems. Academic Press Inc, Orlando, 1984.
[7] V. A. Babenko, L. G. Astafyeva, and V. N. Kuzmin. Electromagnetic scattering in disperse media. Springer-Praxis, Chichester, 2003.
[8] W. Bangerth and R. Rannacher. Adaptive finite elements methods for differential equations. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.
[9] G. Bao. Finite element approximation of time harmonic waves in periodic structures. SIAM J. Numer. Anal., 32(4):1155-1169, 1995.
[10] G. Bao. Numerical analysis of diffraction by periodic structures:TM polarization. Numer. Math, 75(1):1-16, 1996.
[11] G. Bao, Y. Cao, and H. Yang. Numerical solution of diffraction problems by a least squares FEM. Math. Meth. Appl. Sci., 23(12):1073-1092, 2000.
[12] G. Bao and D. C.Dobson. Modeling and optimal design of diffractive optical structures. Surv. Math. Ind, 8:37-62, 1998.
[13] G. Bao, Z. Chen, and H. Wu. Adaptive finite element method for diffraction gratings. J. Opt. Soc. Am. A, 22(6):1106-1114, 2005.
[14] G. Bao, D. Dobson, and K. Ramdani. A constraint on the maximum reflectance of rapidly oscillating dielectric gratings. SIAM J. Control. Opt, 40(6):1858-1866, 2002.
[15] G. Bao and W. W. Symes. A trace theorem for solutions of linear partial differential equations. Math. Meth. Appl. Sci., 14(8):553-562, 1991.
[16] G. Bao, G. Wei, and S. Zhao. Numerical solution of the Helmholtz equation with high wavenumbers. Int. J. Numer. Meth. Eng., 59(3):389-408, 2004.
[17] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. Acta Numerica, pages 1-102, 2001.
[18] P. M. Van Den Berg. Diffraction theory of a reflection grating. Appl. Sci. Res, 24(1):261-293, 1971.
[19] G. Berger, K. Müller, C. Denz, I Földvári, and A. Péter. Digital data storage in a phase-encoded holograhic memory system: data quality and security. Proc. SPIE, 4988:104-111, 2003.
[20] A. Borgioli and T. Cwik. Numerical differentiation for adaptively refined finite element meshes. In Antennas and Propagation Society International Symposium, IEEE, volume 1, pages 256-259, Massachusetts, 1998.
[21] D. Braess. Finite elements. Cambridge University Press, Cambridge, 1997.
[22] S. C. Brenner and L. R. Scott. The mathematical theory of finite element methods, volume 15 of Texts in Applied Mathematics. Springer, New york, 2002.
[23] O. P. Bruno and F. Reitich. Numerical solution of diffraction problems: a method of variation of boundaries. J. Opt. Soc. Am. A, 10(6):1168-1175, 1993.
[24] O. P. Bruno and F. Reitich. Numerical solution of diffraction problems: a method of variation of boundaries. II. Finitely conducting gratings, Padé approximants, and singularities. J. Opt. Soc. Am. A, 10(11):2307-2316, 1993.
[25] O. P. Bruno and F. Reitich. Numerical solution of diffraction problems: a method of variation of boundaries. III Doubly periodic gratings. J. Opt. Soc. Am. A, 10(12):2251-2562, 1993.
[26] A. Buffa. Trace theorems for functional spaces related to Maxwell equations: an overview, in computational electromagnetics. In Lecture Notes in Computational Science and Engineering 28, pages 23-34. Springer-Verlag, Berlin, 2002.
[27] E. Burnab, A. Quarteroni, and B. Stamm. Interior penalty continuous and discontinuous finite element approximations of hyperbolic equations. J. Sci. Comp, 43(3):293-312, 2007.
[28] S. Chandler-Wilde. Boundary value problems for the Helmholtz equation in a half-plane. In Mathematical and numerical aspects of wave propagation (Mandelieu-La Napoule, 1995), pages 188-197. SIAM, Philadelphia, 1995.
[29] S. Chandler-Wilde and P. Monk. Existence, uniqueness and variational methods for scattering by unbounded rough surfaces. SIAM J. Math. Anal., 37(2):598-618, 2005.
[30] S. Chandler-Wilde and B. Zhang. Electromagnetic scattering by an inhomogeneous conducting or dielectric layer on a perfectly conducting plate. Proc. R. Soc. Lond. Ser. A, 454:519-542, 1998.
[31] S. N. Chandler-Wilde and Z. Bo. A uniqueness result for scattering by infinite rough surfaces. SIAM J. Appl. Math., 58(6):1774-1790, 1998.
[32] S. N. Chandler-Wilde and I. G. Graham. Boundary integral methods in highfrequency scattering, in "highly oscillatory problems". Cambridge University Press, pages 154-193, 2009.
[33] S. N. Chandler-Wilde, P. Monk, and M. Thomas. The mathematics of scattering by unbounded, rough, inhomogeneous layers. Journal of Computational and Applied Mathematics, 204:549-559, 2007.
[34] R. Chang. Physical chemistry for the biosciences. University science books, 2005.
[35] P. G. Ciarlet. The finite element method for elliptic equations. North Holland, Amsterdam, 1978.
[36] R. E. Coath. Investigating the use of replica Morpho butterfly scales for colour displays. Society, 5(May):1-9, 2007. (http://printfu.org/blue+ morpho+didius+butterfly).
[37] D. Colton and R. Kress. Inverse acoustic and electromagnetic scattering theory, volume 93 of Applied Mathematical Sciences. Springler-Verlag, Berlin, 1992.
[38] P. Davis and P. Rabinowitz. Method of numerical integration. New York Academic Press, New York, 1984.
[39] R. DeVore, A. Iserles, and E. Süli. Foundations of computational mathematics. Cambridge University Press, Cambridge, 2001.
[40] Z. Ding. A proof of the trace theorem of Sobolev spaces on Lipschitz domains. Proc. Am. Math. Soc., 124(2):591-600, 1996.
[41] J. Elschner and G. Schmidt. Inverse scattering for periodic structures: stability of polygonal interfaces. Inverse Problems, 17:1817-1829, 2001.
[42] C. Erath, S. FerrazLeite, S. Funken, and D. Praetorius. Energy norm based a posteriori error estimation for boundary element methods in two dimensions. Applied Numerical Mathematics, 59:2713-2734, 2009.
[43] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Introduction to adaptive methods for differential equations. In Acta numerica, pages 105-158. Cambridge University Press, Cambridge, 1995.
[44] R. Estrada and R. P. Kanwal. Asymptotic analysis: a distributional approach. Birkhäuser, Boston, 1994.
[45] L. Demkowicz et al. Computing with hp-adaptive finite elements, volume 2. Chapman and Hall, New York, 2008.
[46] S. Fan, J. D. Joannopoulos, Joshua N. Winn, Adrian Devenyi, J. C. Chen, and Robert D. Meade. Guided and defects modes in periodic dielectric waveguides. J. Opt. Soc. Am. B, 12(7):1267-1272, 1995.
[47] S. Gedny, J. Lee, and R. Mittra. A combined FEM/MOM approach to analyze the plane wave diffraction by arbitrary gratings. IEEE Trans. Microwave Theory Tech., 40(2):363-370, 1992.
[48] C. I. Goldstein. A finite element method for solving Helmholtz type equations in waveguides and other unbounded domains. Mathematics of Computation, 39(160):309-324, 1982.
[49] B. Gralak, G. Tayeb, and S. Enoch. Morpho butterfly wings modeled with lamellar grating theory. Opt. Express, 9(11):567-578, 2001.
[50] G. Granet. Reformulation of the lamellar grating problem through the concept of adaptive spatial resolution. J. Opt. Soc. Am. A, 16(10):2510-2516, 1999.
[51] T. Grätsch and K. J. Bathe. A posteriori error estimation techniques in practical finite element analysis. Comput. and Structures, 83(4-5):235-265, 2005.
[52] H. Gross, R. Model, M. Bär, M. Wurm, B. Bodermann, and A. Rathsfeld. Mathematical modelling of indirect measurements in scatterometry. Measurement, 39(9):782-794, 2006.
[53] T. Gudi, N. Nataraj, and A. Pani. An hp-local discontinuous Galerkin method for some quasilinear elliptic boundary value problems of nonmonotone type. Math. Comp, 77(262):731-756, 2008.
[54] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. Princeton University Bulletin, pages 49-52, 1902.
[55] P. S. Heckbert. Introduction to finite element methods. Global Illumination course, 93, 1993.
[56] L. Hormander. The analysis of linear partial differential operators III, volume III. Springer, New York, 1985. A series of comprehensive studies in mathematics.
[57] M. Huber, J. Schoberl, A. Sinwel, and S. Zaglmayr. Simulation of diffraction in periodic media with a coupled finite element and plane wave approach. SIAM J. Sci. Comput, 31(2):1500-1517, 2009.
[58] F. Ihlenburgh. Finite element analysis of acoustic scattering, volume 132. Springer, 1998.
[59] F. Ihlenburgh and I. Babuska. Finite element solution of the Helmholtz equation with high wave number part I: The h-version of the FEM. Comput. Math. Appl., 30(9):9-37, 1995.
[60] U. S. Inan and A. S. Inan. Electromagnetic waves. Prentice Hall, New Jersey, 2000.
[61] K. Ito. Encyclopedic dictionary of mathematics. MIT Press, Cambridge, 2nd edition, 1987.
[62] T. H. Jensen, M. Bech, O. Bunk, T. Donath, C. David, R. Feidenhans, and F. Pfeiffer. Directional x-ray dark-field imaging. Phys. Med. Biol, 55(12):3317-3323, 2010.
[63] H. Jia, K. Yasumoto, and H. Toyama. Reflection and transmission properties of layered arrays of circular cylinders embedded in magnetized ferrite slab. IEEE transactions on antennas and propagation, 53(3):1145-1153, 2005.
[64] J. Jin. The finite element method in electromagnetics. Wiley, New York, 2002.
[65] H. A. Kalhor and M. K. Moaveni. Analysis of diffraction gratings by finitedifference coupling technique. J. Opt. Soc. Am, 63(12):1584-1588, 1973.
[66] P. Kattan. Matlab guide to finite elements. Springer-Verlag, Berlin, 2008.
[67] M. Koshiba and M. Suzuki. Finite element method formulation in terms of electric field vector for electromagnetic waveguide problems. IEEE Trans. Microwave Theory Tech., 33(10):900-905, 1985.
[68] E. Kreyszig. Introductory functional analysis with applications. Wiley, New York, 1989.
[69] G. A. Kriegsmann. Complete transmission through a two-dimensional diffraction grating. SIAM J. Appl. Math., 65(1):24-42, 2004.
[70] A. Krommer and C. Ueberhuber. Numerical integration on advanced computer system. Springer-Verlag, Berlin, 1994.
[71] P. Kurasov. Distribution theory for discontinuous test functions and differential operators with generalized coefficients. J. Math. Anal. Appl., 201(256):297-323, 1996.
[72] G. Larson and T. J. Barth. A posteriori error estimation for adaptive discontinuous Galerkin approximations of hyperbolic systems. In Proc. International Symposium on Discontinuous Galerkin Methods Theory, Computation and Applications (B. Cockburn, G. E. Karniadakis, and C. W. Shu, eds.). Springer, Berlin, 2000.
[73] R. Lech and J. Mazur. Electromagnetic curtain effect and tunneling properties of multilayered periodic structures. IEEE antennas and wireless propagation letters, 7:201-205, 2008.
[74] A. Lechleiter and S. Ritterbusch. A variational method for wave scattering from penetrable rough layers. IMA J. App. Maths, 75:366-391, 2010.
[75] C. M. Linton and D. Evans. The interaction of waves with arrays of vertical circular cylinders. J. Fluid Mech, 215:549-569, 1990.
[76] C. M. Linton and D. Evans. Integral equations for a class of problems concerning obstacles in waveguides. J. Fluid Mech, 245:349-365, 1992.
[77] C. M. Linton and M. McIver. The existence of Rayleigh-Bloch surface waves. J. Fluid Mech., 470:85-90, 2002.
[78] C. M. Linton and I. Thompson. Resonant effects in scattering by periodic arrays. Wave Motion, 44:165-175, 2007.
[79] E. G. Loewen and E. Popov. Diffraction of gratings and applications. Marcel Dekker, INC, New York, 1997.
[80] L. M. Lytvynenko and S. L. Prosvirnin. Wave reflection by a periodic layered metamaterial. Eur. Phys. J. Appl. Phys, 46(3):1-6, 2009.
[81] H. D Maniar and J. N Newman. Wave diffraction by a long array of cylinders. J. Fluid. Mech, 339:309-330, 1997.
[82] D. Maystre. Rigorous vector theories of diffraction gratings. In Progress in optics, E. Wolf, ed., volume 21, pages 1-67. North Holland, Amsterdam, 1984.
[83] D. Maystre and M. Bagieu. Waterman and Rayleigh methods for diffraction grating problems:extension of the convergence domain. J. Opt. Soc. Am. A, 15(6):1566-1576, 1998.
[84] W. McLean. Strongly elliptic systems and boundary integral equations. Cambridge University Press, Cambridge, 2000.
[85] J. M. Melenk and S. Sauter. Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. Math. Comp, 79:1871-1914, 2010.
[86] T. A. Milligan. Modern antenna design. John Wiley and sons, 2005.
[87] W. Milne. Numerical calculus: Approximations, Interpolation, Finite Differences, Numerical Integration and Curve Fitting. Princeton university press, Princeton, 1949.
[88] D. S. Mitrinovic and J. D Keckic. The Cauchy method of residues: theory and applications, volume 2. Springer, 2001.
[89] A. Mohan, R. Raskar, and J. Tumblin. Agile spectrum imaging: programmable wavelength modulation for cameras and projectors. Eurographics, 27(2):1-9, 2008.
[90] P. Monk. Finite element methods for Maxwell's equations. Oxford University Press, Oxford, 2003.
[91] Y. Nakata and M. Koshiba. Boundary-element analysis of plane-wave diffraction from groove-type dielectric and metallic gratings. J. Opt. Soc. Am, 7(8):1494-1502, 1990.
[92] J. T. Oden and M. Ainsworth. A posteriori error estimation in finite element analysis. John Wiley and sons, New York, 2000.
[93] C. Palmer. Diffraction grating handbook. Newport corporation, New York, 2005.
[94] M. Peters, M. Rüdiger, B. Bläsi, and W. Platzer. Electro-optical simulation of diffraction in solar cells. Opt. Express, 18(S4):A584-A593, 2010.
[95] R. Petit. Electromagnetic theory of gratings. Springer, Berlin, 1980.
[96] C. Pflaum. Convergence of the combination technique for the second-order elliptic differential equations. SIAM J. Numer. Anal., 34(6):2431-2455, 1997.
[97] E. Popov. Light diffraction by relief gratings, volume 31. Elsevier, Amsterdam, 1993.
[98] R. A. Potyrailo, H. Ghiradella, A. Vertiatchikh, J. R. Cournoyer K. Dovidenko, and E. Olson. Morpho butterfly wing scales demonstrate highly selective vapour response. Nature Photonics, 1(2):123-128, 2007.
[99] A. Y. Poyedinchuk, Y. A. Tuchkin, and V. P. Shestopalov. New numericalanalytical methods in diffraction theory. Math. Comput. Modelling, 32(9):1029-1046, 2000.
[100] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. Numerical recipes. Cambridge University Press, Cambridge, 1992.
[101] S. Prössdorf. On the super-approximation property of Galerkin's method with finite elements. Numerische Mathematik, 59(1):711-722, 1991.
[102] R. Prum, T. Quinn, and R. Torres. Anatomically diverse butterfly scales all produce structural colours by coherent scattering. J. Exp. Biol., 209(4):748765, 2006.
[103] B. D. Reddy. Introductory functional analysis. Springer, Berlin, 1998.
[104] L. F. Richardson. Measure and integration. A concise introduction to real analysis. Wiley, New Jersey, 2009.
[105] J. C. Robinson. Infinite-dimensional dynamical systems. Cambridge University Press, Cambridge, 2001.
[106] S. Sadov. Computation of quasiperiodic fundamental solution of Helmholtz equation. In Advances in Difference Equations (Gordon and Breach, 1997), pages 551-558. Veszprem, Amsterdam, 1995.
[107] S. Sauter. A refined finite element convergence theory for highly indefinite Helmholtz problems. Computing, 78(2):101-115, 2006.
[108] N. Savage. Acousto-optic devices. Nature Photonics, 4:728-729, 2010.
[109] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. Math. Comp, 28(128):959-962, 1974.
[110] A. H. Schatz and J. Wang. Some new error estimates for Ritz-Galerkin methods with minimal regularity assumptions. Math. Comp, 65:19-27, 1996.
[111] C. Schwab. p-and hp-Finite element methods. Oxford, New York, 1998. Theory and applications in solid and fluid mechanics.
[112] M. Sesay. Scattering properties from a periodic grating of two deformed cylindrical shapes. Journal Optical Society of America, 12(115706):1-6, 2010.
[113] P. Silvester and R. Ferrari. Finite elements for electrical engineers. Cambridge University Press, Cambridge, 3rd edition, 1996.
[114] H. C. Snorre. Discrete Fredholm properties and convergence estimates for the electric field integral equation. Maths. Comp, 73(245):143-167, 2003.
[115] D. H. Staelin, A. W. Morgenthaler, and J. A. Kong. Electromagnetic waves. Prentice Hall, New Jersey, 1994.
[116] I. Stakgold. Green's functions and boundary value problems. WileyInterscience, New York, 1998.
[117] E. Stein, W. Wunderlich, E. Ramm, E. Rank, R. Rannacher, K. Schweizerhof, W. Wendland, G. Wittum, and P. Wriggers. Error-controlled adaptive finite elements in solid mechanics. Wiley, West Sussex, 2003.
[118] B. Strycharz. An acoustic scattering problem for periodic, inhomogeneous media. Math. Meth. Appl. Sci., 21(10):969-983, 1998.
[119] E. Sulci and D. F. Mayers. An introduction to numerical analysis. Cambridge University press, Cambridge, 2003.
[120] J. Sun and C. Henge. Numerical scattering analysis of TE plane waves by a metallic diffraction grating with local defects. J. Opt. Soc. Am. A, 26(1):156-162, 2009.
[121] T. Suyama, Y. Okuno, A. Matsushima, and M. Ohtsu. A numerical analysis of stop band characteristics by multilayered dielectric gratings with sinusoidal profile. PIER B, 2:83-102, 2008.
[122] A. E. Taylor. Introduction to functional analysis. John Wiley and sons, New York, 1958.
[123] V. Turuned. Commutator of periodic pseudodifferential operators. J. Anal. Appl., 19:95-108, 2000.
[124] P. Vukusic. Structural colour in Lepidoptera. Curr. Biol., 16(16):R621R623, 2006.
[125] P. Vukusic, J. Sambles, C. Lawrence, and R. Wootton. Quantified interference and diffraction in single Morpho butterfly scales. Proc. R. Soc. Lond, 266(1427):1403-1411, 1999.
[126] W. R. Wade. An introduction to analysis. Pearson Prentice Hall, New Jersey, 2004.
[127] D. A. G Walker and R. E Taylor. Wave diffraction from linear arrays of cylinders. Ocean Engineering, 32:2053-2078, 2005.
[128] K. Yasumoto, T. Kushta, and H. Toyama. Accurate analysis of twodimensional electromagnetic scattering from multilayered periodic arrays of circular cylinders using lattice sums technique. IEEE transactions on antennas and propagation, 52(10):2603-2611, 2004.
[129] S. Yoshioka and S. Kinoshita. Wavelength-selective and anisotropic lightdiffusing scale on the wing of the Morpho butterfly. Proc. Biol. Sci, 271(1539):581-587, 2004.
[130] S. Yoshioka and S. Kinoshita. Single-scale spectroscopy of structurally colored butterflies: measurements of quantified reflectance and transmittance. J. Opt. Soc. Am. A, 23(1):140-141, 2006.
[131] S. Yoshioka, S. Kinoshita, Y. Fujii, and N. Okamoto. Photophysics of structural color in Morpho butterflies. Forma, 17:103-121, 2002.
[132] P. Zhu, K. Zhang, Z. Wang, Y. Liu, X. Liu, Z. Wu, S. A. McDonald, F. Marone, and M. Stampanoni. Low-dose, simple, and fast grating-based x-ray phase-contrast imaging. PNAS, 107(31):13756-13581, 2010.

## Appendix A

## Sobolev spaces

In order to make statements about the existence and uniqueness of solutions to our problem we will use results from functional analysis [68]. We therefore have to clearly define the space that the solution is a member of. To that end let us start by defining Sobolev spaces and associated norms. This section is concerned with establishing the notation which will be used in this thesis. More details on Sobolev spaces can be found in [35, 22, 68]. We start by giving the definition of a norm as described in [22].

## A. 1 Some useful definitions

Definition A-1. Let $V$ be a linear space, then, for all $v \in V$, a norm $\mathbf{k} . \mathbf{k}$ is a function which maps each element $v$ to a positive real value which satisfies the following properties

1. $\mathrm{k} v \mathrm{k} \geq 0$,
2. $\mathrm{k} v \mathrm{k}=0$ if and only if $v$ is equal to zero.
3. For $c \in \mathbb{R}, \mathbf{k} c v \mathbf{k}=|c| \mathbf{k} v \mathbf{k}$.
4. $\mathbf{k} v+w \mathbf{k} \leq \mathbf{k} v \mathbf{k}+\mathbf{k} w \mathbf{k}$, for all $w \in V$. This last property is called the triangle inequality.

Remark If only properties 1,3 and 4 hold then k . k is a semi-norm and is denoted by |.| instead.

Definition A-2. An inner product is a map

$$
(\cdot, \cdot)_{V}: V \times V \rightarrow \mathbb{C}
$$

that satisfies the following three axioms for all vectors $x, y, z \in V$ and all scalars $c \in \mathbb{C}$

1. Conjugate symmetry

$$
(x, y)_{V}=\overline{(y, x)_{V}}
$$

2. Linearity in the first argument

$$
\begin{align*}
(c x, y)_{V} & =c(x, y)_{V}  \tag{A.1}\\
(x+y, z)_{V} & =(x, z)_{V}+(y, z)_{V}
\end{align*}
$$

3. Positive-definiteness

$$
(x, x)_{V} \geq 0
$$

with equality only for $x=0$.
Definition A-3. A linear space $V$ with an inner product is defined as an inner product space and is denoted by $\left(V,(\cdot)_{V}\right)$.

Definition A-4. Let $\left(V,(\cdot)_{V}\right)$ be an inner-product space. If the associated normed linear space $(V, \mathbf{k} \cdot \mathbf{k})$ is complete, then $\left(V,(\cdot)_{V}\right)$ is called a Hilbert space [22].
Definition A-5. Let $V$ be a linear space. A bilinear form, $b(.,$.$) , is a mapping$ $b: V \times V \rightarrow \mathbb{R}$ such that for any $v, w \in V$, the maps $v \nrightarrow b(v, w)$ and $w \nrightarrow b(v, w)$ are linear forms on $V$.
Definition A-6. Let $\left(V,(\cdot)_{V}\right)$ be an inner-product space. A form $a: V \times V \rightarrow \mathbb{C}$ is sesquilinear if

$$
\begin{align*}
a(x+y, z+w) & =a(x, z)+a(x, w)+a(y, z)+a(y, w)  \tag{A.2}\\
a\left(c_{1} x, c_{2} y\right) & =\bar{c}_{1} c_{2} a(x, y)
\end{align*}
$$

for all $x, y, z, w \in V$ and for all $c_{1}, c_{2} \in \mathbb{C}$. In addition a sesquilinear form is continuous if there exists a constant $C_{\text {cont }}>0$, such that

$$
|a(u, v)|<C_{c o n t} \mathrm{k} u \mathrm{k}_{V} \mathrm{k} v \mathrm{k}_{V}
$$

for all $u \in V, v \in V$.
Definition A-7. Let $f$ be a function defined in a bounded domain $\digamma \subset \mathbb{R}^{m}$, for $m=1,2$ and let $p \in \mathbb{N}$. Then we define the space $L^{p}(\digamma)$ by

$$
L^{p}(\digamma)=\left\{f: \mathrm{k} f \mathrm{k}_{L^{p}(\digamma)}<\infty\right\} .
$$

Here

$$
\mathrm{k} f \mathrm{k}_{L^{p}(\digamma)}:=\left(\int_{\digamma}|f(x, y)|^{p} d x d y\right)^{\frac{1}{p}}, \quad 1 \leq p<\infty
$$

and for $p=\infty$

$$
\mathrm{k} f \mathrm{k}_{L^{\infty}(\digamma)}:=\operatorname{ess} \sup \{|f(x, y)|:(x, y) \in \digamma\}
$$

where
ess sup $\{|f(x, y)|:(x, y) \in \digamma\}=\inf \{C \geq 0:|f(x, y)| \leq C \quad$ for almost everywhere $\}$.

Lemma A-8. For $S \subset \mathbb{R}^{m}$, for $m=1,2$ and for any $f, g \in L^{1}(S)$

$$
\begin{equation*}
(f, g)_{S}=\int_{S} f \bar{g} d S \tag{A.3}
\end{equation*}
$$

is an inner product [22].
Proof. We need to check the three properties given in Definition A-2. We note that

$$
\begin{aligned}
(g, f)_{S} & =\int_{S} g \bar{f} d S \\
& =\int_{S} \overline{f \bar{g}} \\
& =\frac{(f, g)_{S}}{}
\end{aligned}
$$

we also have for $c \in \mathbb{C}$ that

$$
\begin{aligned}
((c f+h), g)_{S} & =\int_{S}(c f+h) \bar{g} d S \\
& =c \int_{S} f \bar{g} d S+\int_{S} h \bar{g} d S \\
& =c(f, g)_{S}+(h, g)_{S}
\end{aligned}
$$

We finish the proof by noting that

$$
\begin{aligned}
(f, f)_{S} & =\int_{S} f \bar{f} d S \\
& =\int_{S}|f|^{2} d S \\
& =\mathrm{k} f \mathrm{k}_{L^{2}(S)}^{2}
\end{aligned}
$$

from Definition A-7 for $p=2$ and so

$$
(f, f)_{S}=0
$$

if and only if $f=0$ from Definition A-1.
Definition A-9. Let $K$ be any compact subset of int $\digamma$, where int $\digamma$ is the interior of $\digamma$. Then we define by

$$
L_{l o c}^{1}(\digamma):=\left\{f: f \in L^{1}(K)\right\},
$$

the set of locally integrable functions.
We can now define the Sobolev norm and the associated Sobolev space as follows.

Definition A-10. Let $l$ be a positive integer, and let $f$ belong to $L_{l o c}^{1}(\digamma)$. Let $\gamma, \gamma_{1}$ and $\gamma_{2}$ be positive integers. Also let us assume that for all $|\gamma| \leq l$ the weak derivatives $D^{\gamma} f$ exist where

$$
D^{\gamma}=\partial_{x}{ }^{\gamma_{1}} \partial_{y}^{\gamma_{2}},
$$

such that

$$
|\gamma|:=\gamma_{1}+\gamma_{2} .
$$

Then the Sobolev space is defined as

$$
W_{p}^{l}(\digamma):=\left\{f: \mathrm{k} f \mathrm{k}_{W_{p}^{l}(\digamma)}<\infty\right\}
$$

where

$$
\mathrm{k} f \mathrm{k}_{W_{p}^{l}(\digamma)}:=\left(\sum_{|\gamma| \leq l} \mathrm{k} D^{\gamma} f \mathrm{k}_{L^{p}(\digamma)}^{p}\right)^{1 / p},
$$

for $1 \leq p<\infty$. If $l=\infty$, then

$$
\mathrm{k} f \mathrm{k}_{W_{\infty}^{l}(\digamma)}:=\max _{|\gamma| \leq l}\left\{\mathrm{k} D^{\gamma} f \mathrm{k}_{L^{\infty}(\digamma)}\right\}
$$

Definition A-11. Let $l$ be a positive integer, and let $f$ belong to $W_{p}^{l}(\digamma)$. The Sobolev semi-norm is defined as

$$
|f|_{W_{p}^{l}(\digamma)}:=\left(\sum_{|\gamma|=l} \mathrm{k} D^{\gamma} f \mathbf{k}_{L^{p}(\digamma)}^{p}\right)^{1 / p}
$$

for $1 \leq p<\infty$. If $l=\infty$, then

$$
|f|_{W_{\infty}^{l}(\digamma)}:=\max _{|\gamma|=l}\left\{\mathrm{k} D^{\gamma} f \mathrm{k}_{L^{\infty}(\digamma)}\right\} .
$$

This can be generalised for $l=s \in \mathbb{R}$. For the special case when $p=2$ we write $W_{2}^{s}(\digamma)=H^{s}(\digamma)$.

This leads to the following definition.
Definition A-12. Let $\zeta \subset \partial \digamma$. The space $H^{\frac{1}{2}}(\zeta)$ is defined as [35]

$$
H^{\frac{1}{2}}(\zeta)=\left\{v \in L^{2}(\zeta): \mathbf{k}^{2} \mathbf{k}_{H^{\frac{1}{2}}(\zeta)}<\infty\right\}
$$

with

$$
\mathbf{k} \mathrm{k}_{H^{\frac{1}{2}}(\zeta)}^{2}=\mathrm{k} v \mathrm{k}_{L^{2}(\zeta)}^{2}+|v|_{H^{\frac{1}{2}}(\zeta)}^{2},
$$

and the semi-norm

$$
|v|_{H^{\frac{1}{2}}(\zeta)}^{2}=\int_{\zeta}\left(\int_{\zeta} \frac{|v(x)-v(y)|^{2}}{|x-y|^{2}} d x\right) d y .
$$

Note that if $f \in H^{\frac{1}{2}}(\zeta)$ then $f \in L^{1}(\zeta)$ and $f \in L^{2}(\zeta)$ [35].

The following theorem will be used a lot later when we study the a priori error estimate of the problem.

Theorem A-13 (Trace theorem). Let $s \geq 1 / 2$, the trace operator $\tau: H^{s}(\digamma) \rightarrow$ $H^{s-\frac{1}{2}}(\partial \digamma)$ is a bounded operator [22, 35, 40] that is

$$
\mathrm{k} \tau g \mathrm{k}_{H^{s-\frac{1}{2}}(\partial \digamma)} \leq \mathrm{k} g \mathrm{k}_{H^{s}(\digamma)}
$$

The operator $\tau$ extends to a linear and continuous operator from $H^{s}(\operatorname{div}, \digamma)$ to $H^{s-1 / 2}(\partial \digamma)$ for $s \geq-1 / 2$ [26], such that

$$
H^{s}(\operatorname{div}, \digamma)=\left\{g \in H^{s}(\digamma): \operatorname{div} g \in H^{s}(\digamma)\right\} .
$$

Let $\digamma \subset \mathbb{R}^{2}$ and $s \in \mathbb{R}$, then we will denote $\mathrm{k} f \mathrm{k}_{H_{\#(\digamma)}^{s}}$ the norm of $f$ in $H^{s}(\digamma)$ when the function $f$ is periodic with respect to $x$. For the particular case where the function is periodic and $\digamma \subset \mathbb{R}$, we also have the following definition.

Definition A-14. Let $g(x)$ be a complex periodic function of period $\lambda$ where $x \in \mathbb{R}$. The space $H_{\#}^{s}(\digamma)$ for $s \in \mathbb{R}$ is [105]

$$
H_{\#}^{s}(\digamma)=\left\{g \in L^{2}(\digamma): g(0)=g(\lambda) \text { and } \mathbf{k} \mathbf{k}_{H^{s}(\digamma)}<\infty,\right\}
$$

with

$$
\begin{equation*}
\mathrm{k}_{\mathrm{g}} \mathrm{k}_{\#(F)}^{2}=\sum_{n \in \mathbb{Z}}\left(1+\left(\frac{2 \pi n}{\lambda}\right)^{2}\right)^{s}\left|g^{(n)}\right|^{2} \tag{A.4}
\end{equation*}
$$

and

$$
g^{(n)}=\frac{1}{\lambda} \int_{0}^{\lambda} g(x) e^{-i \frac{2 \pi n}{d} x} d x, \quad \text { for } \quad n \in \mathbb{Z}
$$

## A. 2 Quasi-periodicity

Definition A-15. A function $G(x, y)$ is $\alpha$-quasi-periodic of period $d$ if

$$
G(x+d, y)=e^{i \alpha d} G(x, y)
$$

for some $d \in \mathbb{R}$.
The periodicity of the grating combined with the presence of the incident wave make $U \alpha$-quasi-periodic [95, 82] and we have the following Lemma holds.

Lemma A-16. $G_{\alpha}$ is $\alpha$-quasi-periodic of period $d$ if and only if there exists a periodic function $G$ with the same period as $G_{\alpha}$ such that

$$
G_{\alpha}(x, y)=e^{i \alpha x} G(x, y)
$$

Proof. If we suppose that we have

$$
\begin{equation*}
G_{\alpha}(x, y)=e^{i \alpha x} G(x, y) \tag{A.5}
\end{equation*}
$$

then

$$
\begin{aligned}
G_{\alpha}(x+d, y) & =e^{i \alpha(x+d)} G(x+d, y) \\
& =e^{i \alpha d} e^{i \alpha x} G(x, y), \quad \text { since } \quad G \text { is periodic with period d, } \\
& =e^{i \alpha d} G_{\alpha}(x, y)
\end{aligned}
$$

by using equation (A.5) so $G_{\alpha}$ is $\alpha$-quasi-periodic. Now, let us suppose that $G_{\alpha}$ is $\alpha$-quasi-periodic

$$
\begin{aligned}
G_{\alpha}(x, y) & =e^{i \alpha d} G_{\alpha}(x-d, y) \\
& =e^{i \alpha(d+x-x)} G_{\alpha}(x-d, y) \\
& =e^{i \alpha x} e^{-i \alpha(x-d)} G_{\alpha}(x-d, y)
\end{aligned}
$$

Let us denote $G(x, y)=e^{-i \alpha x} G_{\alpha}(x, y)$ and let us show that $G$ is periodic with period $d$. We have

$$
\begin{equation*}
G(0, y)=G_{\alpha}(0, y) \tag{A.6}
\end{equation*}
$$

and

$$
G(d, y)=e^{-i \alpha d} G_{\alpha}(d, y)
$$

$G_{\alpha}$ is $\alpha$-quasi-periodic and so

$$
\begin{align*}
G_{\alpha}(0, y) & =e^{-i \alpha d} G_{\alpha}(d, y)  \tag{A.7}\\
& =G(d, y) \tag{A.8}
\end{align*}
$$

Using equations (A.6) and (A.7) gives

$$
G(0, y)=G(d, y)
$$

which finishes the proof.
From now on, if $\digamma \subset \mathbb{R}^{2}$ and $s \in \mathbb{R}$, then we will denote $\mathbf{k} f \mathrm{k}_{H_{\alpha \#}^{s}(\digamma)}$ the norm of $f$ in $H^{s}(\digamma)$ when the function $f$ is $\alpha$ - quasi periodic with respect $x$. For the particular case where $\digamma \subset \mathbb{R}$, and $g(x)$ an $\alpha$-quasi periodic function of period $\lambda$. We have the following definition.

Definition A-17. Let $g(x)$ be a complex $\alpha$-quasi-periodic function of period $d$ where $x \in \mathbb{R}$. The space $H_{\alpha \#}^{s}(\digamma)$ for $s \in \mathbb{R}$ is [118],

$$
H_{\alpha \#}^{s}(\digamma)=\left\{g \in L^{2}(\digamma): g(\lambda)=e^{i \alpha \lambda} g(0): \mathrm{kg}_{H^{s}([\digamma])}<\infty\right\}
$$

with

$$
\begin{equation*}
\mathrm{k}_{\mathrm{g}} \mathrm{k}_{H_{\alpha \#}^{s}(\digamma)}^{2}=\sum_{n \in \mathbb{Z}}\left(1+\left(\frac{2 \pi n}{\lambda}+\alpha\right)^{2}\right)^{s}\left|g^{\left(n_{\alpha}\right)}\right|^{2}, \tag{A.9}
\end{equation*}
$$

and

$$
g^{\left(n_{\alpha}\right)}=\frac{1}{\lambda} \int_{0}^{\lambda} g(x) e^{-i\left(\frac{2 \pi n}{\lambda}+\alpha\right) x} d x, \quad \text { for } \quad n \in \mathbb{Z}
$$

In light of these definitions, let us return to our problem and establish some notation that will allow us to define the various functions that will arise in each of the domains in Figure 2.3. When we study the grating problem analytically in Chapter 3, Chapter 4 and Chapter 6 we use both periodic and $\alpha$-quasi-periodic functions, therefore the function spaces that we will utilize on the boundaries and with the DtN maps are

$$
\begin{align*}
L_{\#}^{s}([0, d]) & =\left\{g \in L^{s}([0, d]): g(d)=g(0)\right\}  \tag{A.10}\\
L_{\alpha \#}^{s}([0, d]) & =\left\{g \in L^{s}([0, d]): g(d)=e^{i \alpha d} g(0)\right\},  \tag{A.11}\\
H_{\#}^{s}([0, d]) & =\left\{g \in H^{s}([0, d]): g(d)=g(0)\right\}  \tag{A.12}\\
H_{\alpha \#}^{s}([0, d]) & =\left\{g \in H^{s}([0, d]): g(d)=e^{i \alpha d} g(0)\right\}, \tag{A.13}
\end{align*}
$$

and the function spaces that we will utilize inside $\Omega$ are

$$
\begin{align*}
L_{\#}^{s}(\Omega) & =\left\{f \in L^{s}(\Omega): f(d, y)=f(0, y), \quad \forall y \in[-B, B]\right\},  \tag{A.14}\\
L_{\alpha \#}^{s}(\Omega) & =\left\{f \in L^{s}([0, d]): f(d, y)=e^{i \alpha d} f(0, y), \quad \forall y \in[-B, B]\right\},  \tag{A.15}\\
H_{\#}^{s}(\Omega) & =\left\{f \in H^{s}(\Omega): f(d, y)=f(0, y), \quad \forall y \in[-B, B]\right\},  \tag{A.16}\\
H_{\alpha \#}^{s}(\Omega) & =\left\{f \in H^{s}(\Omega): f(d, y)=e^{i \alpha d} f(0, y), \quad \forall y \in[-B, B]\right\} . \tag{A.17}
\end{align*}
$$

## Appendix B

## The Finite element method

In this appendix we recall some basics tools in the mathematical theory of finite elements that we need in our analysis. The finite element method is a numerical technique commonly used to approximate the solution to the partial differential equations that arise in many mathematical models. The finite element method reformulates the problem as a variational one which involves an integral of the partial differential equation over the spatial domain. This domain is broken into a finite number of pieces called elements, and the finite element method uses local basis functions to express the approximate solution in each element. By carefully matching this patchwork of local approximations the solution to the original problem is found. In this section we provide a brief review of the basic concepts used in the numerical analysis of finite element methods [35, 22, 21, 92].

## B. 1 Variational formulations of elliptic boundary value problems

In the following, let $\digamma \subset \mathbb{R}^{2}$ be a bounded domain and let $\partial \digamma$ denote the boundary of $\digamma$. Also, let $L$ be a partial differential operator of second order and let $u \in$ $C^{2}(D)$ be such that

$$
\begin{equation*}
L u:=-\sum_{j, k=1}^{2} \partial_{j}\left(a_{j k} \partial_{k} u\right)+a_{0} u, \tag{B.1}
\end{equation*}
$$

where $a_{j k}(x, y)$ and $a_{0}$ belong to $L^{\infty}(D)$. The associated boundary value problem is given as follow. Find $u \in C^{2}(D)$ such that for $f \in L^{2}(D)$ and $g \in L^{2}\left(\tau_{2}\right)$, we have

$$
\begin{align*}
L u & =f \text { in } D, \\
u & =0 \text { in } \tau_{1}, \\
\frac{\partial u}{\partial n} & =g \text { in } \tau_{2}, \tag{B.2}
\end{align*}
$$

where $\tau_{j} \subset \partial \digamma$ for $j=1,2$.
Lemma B-1. Let us denote $C_{0}^{1}(\digamma)=\left\{v \in C^{1}(\digamma): v(x, y)=0,(x, y) \in \tau_{1}\right\}$. Then for any $v \in C_{0}^{1}(\digamma)$, the given boundary value problem (B.2) can be formulated as a variational problem often referred as weak formulation as follows. Find $u \in C^{1}(\digamma)$ that satisfies

$$
\begin{equation*}
a(u, v)=(l, \bar{v})_{V}, \tag{B.3}
\end{equation*}
$$

where

$$
\begin{align*}
a(u, v) & =\int_{\digamma}\left(\sum_{j, k} a_{j k} \partial_{k} u \partial_{j} v+a_{0} u v\right) d x d y \\
(l, \bar{v})_{V} & =\int_{\digamma} f \bar{v}-\int_{\tau_{2}} \sum_{j, k} a_{j k} g \bar{v} n_{j} d s \tag{B.4}
\end{align*}
$$

and $f \in L^{2}(\digamma)$ and $g \in L^{2}\left(\tau_{2}\right)$. Here $n_{j}$ is the $j$-th component of the outwardpointing normal derivative $n$ for $j=1,2$.

Proof. We use Green's formula in our proof [22]

$$
\begin{equation*}
\int_{\digamma} \partial_{j} w \bar{v} d x d y=-\int_{\digamma} w \partial_{j} \bar{v} d x d y+\int_{\partial \digamma} w \bar{v} n_{j} d s \tag{B.5}
\end{equation*}
$$

for $j=1,2$ where $v, w$ is a $C^{1}$ function. We insert $w=a_{j k} \partial_{k} u$ in (B.5) which leads us to

$$
\begin{align*}
\int_{\digamma} \partial_{j}\left(a_{j k} \partial_{k} u\right) \bar{v} d x d y & =-\int_{\digamma} a_{j k} \partial_{k} u \partial_{j} \bar{v} d x d y+\int_{\partial \digamma} a_{j k} \partial_{k} u \bar{v} n_{j} d s  \tag{B.6}\\
& =-\int_{\digamma} a_{j k} \partial_{k} u \partial_{j} \bar{v} d x d y+\int_{\tau_{2}} a_{j k} g \bar{v} n_{j} d s
\end{align*}
$$

because $v=0$ on $\tau_{1}$ and we use equation (B.2), which finish the proof.

## B. 2 Analysis background

In this part we will introduce a brief review of the analysis, which is used to establish existence and uniqueness of solutions to variational formulations based on elliptic boundary value problems. Let $V$ be a Hilbert space with induced norm $\|$.$\| and let a: V \times V \rightarrow \mathbb{C}$ be a bounded, sesquilinear form, which is continuous (see Definition A-6). The dual of $V$, which we denote by $V^{\prime}$, is defined as the set of all linear maps

$$
\psi: V \rightarrow \mathbb{C} .
$$

We want to show that our problem is well posed by proving that the solution exists and is unique. We need one of the following results.

1. Lax-Milgram

Definition B-2. A sesquilinear form $a: V \times V \rightarrow \mathbb{C}$ is called $V$ - elliptic (positive definite) if there exists a positive constant $C$ such that

$$
C \mathbf{k} u \mathbf{k}_{V}^{2} \leq|a(u, u)|
$$

for all $u \in V$ [58][p. 46].
We prove the existence and uniqueness of solutions for positive definite problems by using the following Lax-Miligram theorem.

Theorem B-3. (Lax-Miligram) Assume that a sesquilinear form $a: V \times$ $V \rightarrow \mathbb{C}$, defined on a Hilbert space $V$, satisfies
(a) continuity which means there exists $M>0$ such that

$$
|a(u, v)| \leq M \mathbf{k} u \mathbf{k}_{V} \mathbf{k} v \mathbf{k}_{V}
$$

for all $u, v \in V$,
(b) $V$-ellipticity as given in Definition B-2 and let $f$ be a continuous linear functional defined on $V$. Then there exists a unique element $u_{0} \in V$ such that for all $v \in V$ we have [58][p. 46]

$$
a\left(u_{0}, v\right)=(f, v)_{V}
$$

## 2. Fredholm alternative

Definition B-4. A sesquilinear form $a: V \times V \rightarrow \mathbb{C}$ is called $V$-coercive if it satisfies for all $v \in V$ the Gårding inequality [58] [p. 51]

$$
M \mathrm{k} v \mathrm{k}_{V}^{2} \leq C \mathrm{k} v \mathrm{k}^{2}+|a(v, v)|, \quad \text { for all } \quad v \in V
$$

where $M$ and $C$ are positive constants and

$$
\mathbf{k} v \mathbf{k}^{2}=(v, v)_{V} .
$$

The following result is used to prove the existence of solutions to the variational equation from their uniqueness.

Lemma B-5. If the sesquilinear form $a: V \times V \rightarrow \mathbb{C}$ is $V$-coercive. Then the problem corresponding to $a(u, v)=(f, v)$ satisfies the Fredholm alternative and either the problem has a solution $u \in V$ for all continuous linear functionals $f$ or there exists a nontrivial solution of the homogeneous problem. Hence the existence of the solution follows if we can show uniqueness [58][p. 51].

## B. 3 Finite element space

We assume that the reader is familiar with finite element analysis but we include parts of this section to establish notation using [35, 22, 21, 92].

1. Finite element partitions.

In our study, we use the polynomial space $\mathbb{P}_{p}$ to construct finite elements on a triangular mesh. The space $\mathbb{P}_{p}$ is the set of polynomials of total degree at most $p$ in the variables $x$ and $y$,

$$
\mathbb{P}_{p}=\operatorname{span}\left\{x^{l} y^{m}, 0 \leq l, m, l+m \leq p, \text { for } l, m, p \in \mathbb{N}\right\}
$$

We will use $D$ to denote a polygonal domain and $\partial D$ to denote its boundary. A finite element partition $\zeta$ of $D$ is a collection $\{K\}$ of elements such that

- the elements form a partition of the domain, that is

$$
\bar{\Omega}=\mathrm{U}_{K \in \zeta} \bar{K},
$$

where $\bar{\Omega}(\bar{K})$ is the closure of $\Omega$ (is the closure of $K$ ).

- each triangular element is contained in $D$,
- the intersection of two adjacent elements is always nonempty and either it is a single common vertex or a single common edge of both elements

2. Finite element spaces on triangles.

Let $p \in \mathbb{N}$ and let $\zeta$ be a regular partition of the domain $D$ into triangular elements. The finite element subspace of order $p$ associated with $\zeta$ is given by

$$
V=\left\{v \in C(\bar{D}): \forall K \in \zeta,\left.v\right|_{K} \in \mathbb{P}_{p}\right\}
$$

where $\bar{D}$ is the closure of $D$. The mesh size of an element (triangle) is defined as the diameter of the triangle.
3. Degrees of freedom.

Let $p \in \mathbb{N}$ and let $\zeta$ be a regular partition of the domain $D$ into triangular elements. The total number of triangular elements $K$ inside the partition $\zeta$ is denoted $N_{K}$. To construct each triangular element we need 3 nodes (nodal degrees of freedom) and the total nodes needed for all the triangular elements inside $\zeta$ is denoted by $d o f_{h}$. To construct the polynomial basis of degree $p$ on each triangular element $K$ we need to add $3(p-1)+p-2$ nodes. If we denote by $d o f_{p}$, the total number of nodes needed for all the polynomial bases of degree $p$ on all the triangular elements then the total number of degrees of freedom, denoted by dof, is defined as

$$
\begin{equation*}
d o f=d o f_{h}+d o f_{p} . \tag{B.7}
\end{equation*}
$$

## Appendix C

## A regularity estimate for the inhomogeneous Helmholtz problem for periodic gratings for Case 1B, Case 2A and Case 2B

For completeness of Chapter 3, regularity results for Case 1B and Case 2B will be derived in this chapter.

## C. 1 General case

Let $f(x, y)$ represent the forcing term in the inhomogeneous Helmholtz equation. For perfectly conducting gratings, we have $f(x, y) \in L^{2}\left(\mathbb{R}_{+}^{2}\right)$ and we have $f(x, y) \in L^{2}\left(\mathbb{R}^{2}\right)$ for transmitting dielectric gratings. In this chapter, we want to study the regularity of the solution $U(x, y)$ of the inhomogeneous Helmholtz problem depending on the function $f(x, y)$ for Case 1B and 2B. The regularity of the solution $U(x, y)$ will enable us to study the a priori error estimation of the approximate solution when we solve the Helmholtz problem numerically.
We again use the same notation as in Chapter 2 for the spatial domains as shown in Figure 2.1. We also assume that $f$ is local with respect to the $y$ direction which means that $\operatorname{supp} f \subset \mathbb{R} \times[-B, B]$ (see Figure 2.3).

## C.1.1 The inhomogeneous Helmholtz equation

Similar to the derivation of equation (3.3), we have

- Case 1B: Perfectly conducting grating: TM case

Here we are solving for $U=H_{z}$ for a given function $f(x, y) \in L^{2}\left(\mathbb{R}_{+}^{2}\right)$. The
inhomogeneous Helmholtz problem is to find $U(x, y) \in C^{2}\left(\mathbb{R}_{+}^{2}\right)$ such that

$$
\begin{array}{rlrl}
\Delta U(x, y)+k_{1}^{2} U(x, y) & = & f(x, y), & \\
\frac{\partial U(x, y)}{\partial n} & = & 0, &  \tag{C.1}\\
& (x, y) \in \partial \mathbb{R}_{+}^{2} \\
\hline
\end{array}
$$

subject to the radial condition

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{C.2}
\end{equation*}
$$

where $\frac{\partial}{\partial n}$ denotes the normal derivative operator on the boundary of each scatterer (shaded region in Figure 2.1). Putting $\underline{H}=\left(0,0, H_{z}\right)$ and $\underline{J}=$ $\left(J_{x}, J_{y}, J_{z}\right)$ in equation (3.2) we find that $f(x, y)=\partial_{y}\left(J_{x}\right)-\partial_{x}\left(J_{y}\right)$.
Similar to the derivation of equation (3.24), we have for Case 2B.

## - Case 2B: Transmitting dielectric grating: TM case

Here we are solving for $U=H_{z}$ for a given function $f(x, y) \in L^{2}\left(\mathbb{R}^{2}\right)$. The inhomogeneous Helmholtz problem is to find $U(x, y) \in C^{2}\left(\mathbb{R}^{2}\right)$ such that

$$
\begin{equation*}
\nabla\left(\frac{1}{k^{2}} \nabla U(x, y)\right)+U(x, y)=f(x, y), \quad(x, y) \in \mathbb{R}^{2} \tag{C.3}
\end{equation*}
$$

subject to the radiation condition

$$
\begin{equation*}
\lim _{|y| \rightarrow \infty} U(x, y)=0 \tag{C.5}
\end{equation*}
$$

and the interface conditions given by equations (3.20), (3.21), (3.22) and (3.23). Putting $\underline{H}=\left(0,0, H_{z}\right)$ and $\underline{E}=\left(E_{x}, E_{y}, 0\right)$, equation (3.2) gives

$$
f(x, y)=\partial_{y}\left(\frac{1}{i w^{2} \varepsilon(x, y)} J_{x}\right)-\partial_{x}\left(\frac{1}{i w^{2} \varepsilon(x, y)} J_{y}\right)
$$

We now utilize the periodicity of the grating and restrict our problem to the vertical single strip $S=[0, d] \times \mathbb{R}$ as shown in Figure 2.2. Hence the wavenumber $k$ is given by equation (2.33) for Case 1B and equation (3.27) for Case 2B.

In a similar way to Case $1 \mathrm{~A}, U$ and $f$ are $\alpha$-quasi-periodic with respect to $x$ on this vertical single strip. We again choose our solution $U$ to satisfy the upward propagating radiation condition [28]. To study the regularity of our solution $U$, we also need to study the $\alpha$-quasi-periodic fundamental solution of the problems (C.1) and (C.3).

## C.1.2 The $\alpha$-quasi-periodic Green functions of the Helmholtz equation

In a same way to derive Theorem 23, we have the following results.
Theorem C-1. Case 1B: Let $f(x, y) \in L_{\alpha \#}^{2}\left(S \backslash \Omega_{3}\right)$ and let $U \in C^{2}\left(S \backslash \Omega_{3}\right)$ satisfy the inhomogeneous Helmholtz equation given by equation (C.1). Then, the solution $U$ of equation (C.1) is given by

$$
\begin{equation*}
U(x, y)=\int_{S \backslash \Omega_{3}} G\left(x-x_{0}, y-y_{0}\right) f\left(x_{0}, y_{0}\right) d x_{0} d y_{0} . \tag{C.6}
\end{equation*}
$$

Proof. We replace $S \backslash \overline{\Omega_{3}}$ with $S \backslash \Omega_{3}$ and we can use the same process as in Theorem 23 to show that

$$
\Delta U(x, y)+k_{1}^{2} U(x, y)=f(x, y) .
$$

Also, $U$ is $\alpha$-quasi -periodic because $G$ and $f$ are both $\alpha$-quasi-periodic. In addition,

$$
\lim _{|y| \rightarrow \infty} U(x, y)=0
$$

because $G$ is composed of bounded outgoing waves and $f$ has a locally compact support with respect to $y$. We finish the proof by noting that

$$
\left.\frac{\partial U}{\partial n}\right|_{\partial \Omega_{3}}=\int_{\partial \Omega_{3}} \frac{\partial G\left(x-x_{0}, y-y_{0}\right)}{\partial n} f\left(x_{0}, y_{0}\right) d x_{0} d y_{0}
$$

In our case, for perfectly conducting gratings, the profile of the grating $\partial \Omega_{3}$ can be presented either by Figure 3.1 or Figure 3.2. First, let $\partial \Omega_{3}$ be a closed curve as shown in Figure 3.1, we can see that $\frac{\partial G\left(x-x_{0}, y-y_{0}\right)}{\partial n} f\left(x_{0}, y_{0}\right)$ is regular inside $\partial \Omega_{3}$, therefore we can apply Cauchy integral theorem [88, p. 4] and we have

$$
\begin{equation*}
\left.\frac{\partial U}{\partial n}\right|_{\partial \Omega_{3}}=0 . \tag{C.7}
\end{equation*}
$$

Next, if $\partial \Omega_{3}$ is an open curve as shown in Figure 3.2, we can see that $\frac{\partial G\left(x-x_{0}, y-y_{0}\right)}{\partial n} f\left(x_{0}, y_{0}\right)$ is regular inside the closed contour $C$, where $C$ consists of $\cup_{j=1}^{3} L_{j} \cup \partial \Omega_{3}$ such that $L_{1}$ and $L_{2}$ are a period $d$ apart. We can apply Green's theorem and we have

$$
\begin{aligned}
\left.\frac{\partial U}{\partial n}\right|_{C} & =\int_{C} \frac{\partial G\left(x-x_{0}, y-y_{0}\right)}{\partial n} f\left(x_{0}, y_{0}\right) d x_{0} d y_{0}, \\
& =0 .
\end{aligned}
$$

We note that the contributions from $L_{1}$ and $L_{2}$ cancel each other and the contribution from $L_{3}$ also vanishes, by UPRC since $G$ consists of waves traveling away from the grating. Hence

$$
\begin{align*}
\left.\frac{\partial U}{\partial n}\right|_{\partial \Omega_{3}} & =\int_{\partial \Omega_{3}} \frac{\partial G\left(x-x_{0}, y-y_{0}\right)}{\partial n} f\left(x_{0}, y_{0}\right) d x_{0} d y_{0}  \tag{C.8}\\
& =0 \tag{C.9}
\end{align*}
$$

which finishes the proof.
For the TM case, we have the following result
Theorem C-2. Case 2B: Let $f(x, y) \in L_{\alpha \#}^{2}(S)$ and let $U \in C^{2}(S)$ satisfy the inhomogeneous Helmholtz equation given by equation (C.3) in $S$. Then, we have

$$
\begin{equation*}
U(x, y)=\int_{S} G_{j}\left(x-x_{0}, y-y_{0}\right) f\left(x_{0}, y_{0}\right) d x_{0} d y_{0} \tag{C.10}
\end{equation*}
$$

for $j \in\{0,1,2,3\}$ with $G_{j}$ as defined by equation (3.29) where $d_{j}^{n}=0$ for $j=1,2$ (radiation condition) and such that for $l \in\{0,1,2,3\}$, the coefficients $c_{j}^{n}, c_{l}^{n}$ and $d_{l}^{n}$ are chosen such that the boundary conditions on the interface separating $S_{j}$ and $S_{l}$, given by equations (3.20), (3.21), (3.22) and (3.23) are satisfied.

Proof. The proof is similar to Case 2A .
Remark The coefficients $c_{j}^{n}$ and $d_{j}^{n}$ are different for Case 2A and Case 2B because when we apply the interface conditions we use $\underline{E}=\left(0,0, E_{z}\right)$ and $\underline{H}=$ $\left(H_{x}, H_{y}, 0\right)$ for Case 2A whereas we use $\underline{H}=\left(0,0, H_{z}\right)$ and $\underline{E}=\left(E_{x}, E_{y}, 0\right)$ for Case 2B. From physical considerations, $c_{j}^{n}$ and $d_{j}^{n}$ equals to zero for some $|n|>$ $N_{0} \in \mathbb{N} ; N_{0}$ depends on the complexity of the profile of the grating. Denoting $C_{j}=$ $\left(c_{j}^{n}\right)_{|n|=0, \cdots, N_{0}}, D_{j}=\left(c_{j}^{n}\right)_{|n|=0, \cdots, N_{0}}$ for $j=0,1,2,3$ then we have the same number of unknowns and equations using equations (3.20), (3.21), (3.22) and (3.23).

## C.1.3 Regularity of the solution of Helmholtz problem for periodic grating

In this section, we use the $\alpha$-quasi-periodic fundamental solution $G$ (or $G_{j}$ ) to establish the regularity of each solution which means that we will try to bound the norm of each solution and its partial derivative by using some constants times the norm of the forcing term. From Theorem 27, we have derived the regularity of the solution $U$ in terms of the given function $f$ for Case 1A. We can proceed similarly for Case 1B and we have the following regularity result.

- Case 1B: Perfectly conducting grating: TM case

Theorem C-3. For any $\gamma=\left(\gamma_{1}, \gamma_{2}\right)$ such that $\gamma_{j} \in \mathbb{N}$, for $j=1,2$, and for $x \in[0, d], y \in S_{y}$ there exists a constant $C_{\text {reg }}$ which is independent of the wave number $k$ such that the solution $U$ of equation (C.1) given by Theorem C-1 satisfies

$$
\mathrm{k} D^{\gamma} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} \leq C_{r e g} \mathrm{k} k \mathbf{k}_{\infty}^{|\gamma|-1} \mathrm{k} f \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}
$$

where $\mathbf{k} D^{\gamma} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}$ is given in Definition A-10.

## - Case 2B: Transmitting dielectric grating: TM case

From Theorem 33, we have derived the regularity of the solution $U$ in terms of the given function $f$ for the Case 2A. We proceed similarly for Case 2B and we have the following results.

Lemma C-4. For any $\gamma=\left(\gamma_{1}, \gamma_{2}\right)$ such that $\gamma_{q} \in \mathbb{N}$, and $x \in[0, d]$ and $y \in \mathbb{R}$, there exists $N_{0} \in \mathbb{N}$, with $c_{j}^{n}$ and $d_{j}^{n}$ equal to zero when $|n|>N_{0}$, and there exists $k_{r e f}<\left|k_{j}\right|$ such that $U$ given by equation (C.10) in Theorem C-1 satisfies

$$
\begin{aligned}
\mathrm{k} \partial_{x}^{\gamma_{1}} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)} & \leq \sup _{n \in \mathbb{Z}} \mathrm{k} n_{\alpha} \mathrm{k}_{\infty}^{\gamma_{1}} \mathrm{k} U \mathrm{k}_{L^{2}\left(S \backslash \Omega_{3}\right)}, \\
\mathrm{k} \partial_{y}^{\gamma_{2}} U \mathrm{k}_{L^{2}(S)} & \leq \sup _{n \in \mathbb{Z}, j} \mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}^{\gamma_{2}^{2}} \mathrm{k} U \mathrm{k}_{L^{2}(S)}+C\left(k_{0}, k_{3}\right) C_{s} \sup _{n \in \mathbb{Z}, j}\left|\beta_{j}^{n}\right|^{\gamma_{2}-2} \mathrm{k} f \mathrm{k}_{L^{2}(\Omega)},
\end{aligned}
$$

for $j \in\{0,1,2,3\}$ where $C\left(k_{0}, k_{3}\right)$ is as given by equation (3.36) and

$$
\begin{equation*}
C_{s}=\sup _{n, j \in\{0,1,2,3\}}\left(\gamma_{2}\right)\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right) \tag{C.11}
\end{equation*}
$$

for $n \in Z$ with $\beta_{j}^{n}$ and $z_{n}$ as given by equations (2.43) and (2.44). The coefficients $c_{j}^{n}$ and $d_{j}^{n}$ are defined in Theorem C-2.

Proof. We proceed similarly as in the proof of Lemma 31, but we note that $\nabla U$ is no longer continuous across the interface separating two media therefore for $\gamma_{2}>0$ there is a jump condition across the interface and $C_{s}=C_{s 0}\left(\gamma_{2}\right)$.

Next, we give an approximation of $U$ using $f$.
Lemma C-5. For any $x \in[0, d]$ and for any $y \in \mathbb{R}$, we have for any function $f \in L_{\alpha \#}^{2}([0, d] \times \mathbb{R})$ and for $U$ as given by equation (C.10)

$$
\mathrm{k} U(x, y) \mathrm{k}_{L^{2}(S)} \leq \sup _{j \in\{0,1,2,3\}, n \in \mathbb{Z}} \frac{\sup _{n, j}\left(\left|c_{j}^{n}\right|,\left|d_{j}^{n}\right|\right)}{\mathrm{k} \beta_{j}^{n} \mathrm{k}_{\infty}} \mathrm{k} f(x, y) \mathrm{k}_{L^{2}(S)}
$$

for any $n \in \mathbb{Z}$ and $j=0,1,2,3$

Proof. We proceed similarly as we have done for Lemma 32.
This leads to the following regularity result for Case 2B .
Theorem C-6. For any $\gamma=\left(\gamma_{1}, \gamma_{2}\right)$ such that $\gamma_{j} \in \mathbb{N}$, for $j=1,2$, and for $x \in[0, d], y \in[-B, B] \subset \mathbb{R}$ there exists a constant $C_{\text {reg }}$ which is independent of the wave number $k$ such that the solution $U$ of equation (C.3) given by Theorem C-2 satisfies

$$
\mathbf{k} D^{\gamma} U \mathbf{k}_{L^{2}(\Omega)} \leq C_{\text {reg }}\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right) \mathbf{k} k \mathbf{k}_{\infty}^{|\gamma|-1} \mathbf{k} f \mathbf{k}_{L^{2}(\Omega)},
$$

with $C_{s}$ and $C\left(k_{0}, k_{3}\right)$ as given by equation (C.11) in Lemma C-4 and equation (3.36) in Lemma 31 and $\mathrm{k} D^{\gamma} U \mathrm{k}_{L^{2}(\Omega)}$ is given in Definition A-10.

Proof. We follow the same procedure as we have done in Theorem 33 for Case 2A.

## Appendix D

## A priori error estimates using the $\alpha, 0$-quasi periodic transformation Case 1B, Case 2A and Case 2B

For completeness of Chapter 4, we have relegated the analysis of Cases 1B, 2A and 2B to this chapter. Similar to Case 1A, $U$ is $\alpha$-quasi periodic then we can use $\alpha, 0$-quasi periodic transformation and define a function denoted by $U_{\alpha, 0}$ which is periodic with respect to $x$ where $U=e^{i \alpha x} U_{\alpha, 0}$. The subscripts $\alpha, 0$ indicate the transformation used to make risen the function. We start by looking for the wave equations satisfied by $U_{\alpha, 0}$, study the continuity properties corresponding to $U_{\alpha, 0}$, and then examine the variational formulation. We show that the problem corresponding to the variational formulation is well-posed. We then use the finite element method to discretize the problem, and provide a rigorous study of the a priori error estimation. Following the same approach in Case 1A, we derive regularity results for the scattering problem in periodic space $H_{\#}^{l}(\Omega)$ for the transmitting dielectric ( $H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ for perfectly conducting) gratings for $l \geq 1$. We then show that solving the variational formulation for $U$ is equivalent to solving the variational formulation for $U_{\alpha, 0}$. Since, we have already derived regularity results in the quasi-periodic case in Appendix C and Chapter 3, and since the variational formulation for $U$ is much simpler than that of $U_{\alpha, 0}$, we investigate the a priori error estimation in $H_{\alpha \#}^{l}(\Omega)$ for transmitting dielectric ( $H_{\alpha \#}^{l}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ for perfectly conducting) gratings. This will then allow us to prove new error estimates which give an explicit dependence on the wavenumber.

## D. 1 Restatement of the boundary value problems for the periodic solution

We proceed similarly as in Section 4.1 when we study Case 1A and we derive

- Case 1B: Perfectly conducting grating: TM case

Lemma D-1. Let $U_{\alpha, 0} \in C^{2}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ satisfy equation (4.1), then $U_{\alpha, 0}$ is the solution of the following problem set in the truncated domain $\Omega \backslash$ int $\Omega_{3}$ (see Figure 2.3)

$$
\begin{equation*}
\Delta U_{\alpha, 0}+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0}=0 \tag{D.1}
\end{equation*}
$$

with the DtN map interface conditions at the boundaries of the truncated region given by

$$
\begin{array}{ll}
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}, & \text { on } \Gamma_{+} \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=0, & \text { on } \Gamma_{-} . \tag{D.3}
\end{array}
$$

The Neumann boundary conditions at the surface of the diffraction grating is

$$
\begin{equation*}
\partial_{n} U_{\alpha, 0}(x, y)=-i \alpha n_{x} U_{\alpha, 0}(x, y), \text { on } \partial \Omega_{3}, \tag{D.4}
\end{equation*}
$$

and the periodic condition

$$
\begin{equation*}
U_{\alpha, 0}(d, y)=U_{\alpha, 0}(0, y), \quad y \in[-B, B], \tag{D.5}
\end{equation*}
$$

holds where $U(x, y)$ is the solution of the original Helmholtz problem given by equation (2.102) where $T_{ \pm}^{\alpha, 0}$ is given by equation (4.7) in Lemma 45 and $n=\left(n_{x}, n_{y}\right)$ is the normal unit vector exterior to $\partial \Omega_{3}$.

Proof. We omit most of the proof because it is very similar to Lemma 45. The boundary condition on the interface $\partial \Omega_{3}$ given by equation (2.103) becomes

$$
\begin{aligned}
\partial_{n} U & =\partial_{n}\left(e^{i \alpha x} U_{\alpha, 0}\right), \\
& =U_{\alpha, 0} \nabla\left(e^{i \alpha x}\right) \cdot n+e^{i \alpha x} \partial_{n} U_{\alpha, 0} \\
& =i \alpha n_{x} e^{i \alpha x} U_{\alpha, 0}+e^{i \alpha x} \partial_{n} U_{\alpha, 0}=0
\end{aligned}
$$

and since $e^{i \alpha x} \in 0$ then $\partial_{n} U=0$ if and only if $i \alpha n_{x} U_{\alpha, 0}+\partial_{n} U_{\alpha, 0}=0$ on $\partial \Omega_{3}$.

## - Case 2A: Transmitting dielectric grating: TE case

Lemma D-2. Let $U_{\alpha, 0} \in C^{2}(\Omega)$ satisfy equation (4.1), then $U_{\alpha, 0}$ is the solution of the following problem in the truncated domain $\Omega$ (see Figure 2.3)

$$
\begin{equation*}
\Delta U_{\alpha, 0}+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0}=0 \tag{D.6}
\end{equation*}
$$

with the DtN map interface conditions at the boundaries of the truncated region given by

$$
\begin{array}{ll}
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}, & \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=0, & \text { on } \Gamma_{-} . \tag{D.8}
\end{array}
$$

The periodic condition

$$
\begin{equation*}
U_{\alpha, 0}(d, y)=U_{\alpha, 0}(0, y), \quad y \in[-B, B], \tag{D.9}
\end{equation*}
$$

holds where $U(x, y)$ is the solution of the original Helmholtz problem given by equation (2.104) where $T_{ \pm}^{\alpha, 0}$ is given by equation (4.7) in Lemma 45.

Proof. The proof is similar to Lemma 45.

## - Case 2B: Transmitting dielectric grating: TM case

Lemma D-3. Let $U_{\alpha, 0} \in C^{2}(\Omega)$ satisfy equation (4.1), then $U_{\alpha, 0}$ is the solution of the following problem set in the truncated domain $\Omega$ (see Figure 2.3)

$$
\begin{equation*}
\nabla_{\alpha} \cdot\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0}\right)+U_{\alpha, 0}=0 \tag{D.10}
\end{equation*}
$$

with the DtN map at the boundaries of the truncated region given by

$$
\begin{array}{ll}
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B}, & \text { on } \Gamma_{+} \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial n}\right) U_{\alpha, 0}=0, & \text { on } \Gamma_{-} . \tag{D.12}
\end{array}
$$

The periodic condition

$$
\begin{equation*}
U_{\alpha, 0}(d, y)=U_{\alpha, 0}(0, y), \quad y \in[-B, B], \tag{D.13}
\end{equation*}
$$

holds where $U(x, y)$ is the solution of the original Helmholtz problem given by equation (2.105) where $T_{ \pm}^{\alpha, 0}$ is given by equation (4.7) in Lemma 45 and $\nabla_{\alpha}=\nabla+i(\alpha, 0)$.

Proof. Since $U=e^{i \alpha x} U_{\alpha, 0}$ then

$$
\begin{aligned}
\nabla \cdot\left(\frac{1}{k^{2}} \nabla U\right) & =\nabla \cdot\left(\frac{1}{k^{2}} \nabla\left(e^{i \alpha x} U_{\alpha, 0}\right)\right), \\
& =\nabla \cdot\left(\frac{1}{k^{2}}\left(i(\alpha, 0) e^{i \alpha x} U_{\alpha, 0}+e^{i \alpha x} \nabla U_{\alpha, 0}\right)\right), \\
& =\nabla \cdot\left(\frac{1}{k^{2}} e^{i \alpha x} \nabla_{\alpha} U_{\alpha, 0}\right), \\
& =i \alpha e^{i \alpha x}\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0}\right)+e^{i \alpha x} \nabla \cdot\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0}\right), \\
& =e^{i \alpha x} \nabla_{\alpha \cdot} \cdot\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0}\right),
\end{aligned}
$$

and so equation (2.105) becomes

$$
\begin{aligned}
0 & =\nabla \cdot\left(\frac{1}{k^{2}} \nabla U\right)+U \\
& =e^{i \alpha x} \nabla_{\alpha} \cdot\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0}\right)+e^{i \alpha x} U_{\alpha, 0} .
\end{aligned}
$$

We have $e^{i \alpha x} 60$ therefore

$$
\nabla_{\alpha \cdot}\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0}\right)+U_{\alpha, 0}=0 .
$$

The rest of the proof is similar to Lemma 45.

## D.1.1 Variational formulation

To obtain a numerical method for computing an approximation to $U_{\alpha, 0}$ we start by deriving a variational statement of each scattering problem. Similar to the derivation of Lemma 46 for Case 1A, we have the following results.

## - Case 1B: Perfectly conducting grating: TM case

Lemma D-4. The variational form of the boundary value problem given by equation (D.1) to equation (D.5) is given by the following statement. Find $U_{\alpha, 0} \in H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$, for all $v \in H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v\right)=(f, v)_{\Gamma_{+}}, \tag{D.14}
\end{equation*}
$$

where

$$
\begin{align*}
a(w, v)= & \int_{\Omega \backslash \operatorname{int} \Omega_{3}} \nabla w \cdot \nabla \bar{v}-\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) w \bar{v}-2 i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x} w\right) \bar{v} \\
& +i \int_{\partial \Omega_{3}} \alpha n_{x} w \bar{v}-\int_{\Gamma_{+}} T_{+}^{\alpha, 0} w \bar{v}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} w \bar{v}  \tag{D.15}\\
(f, v)_{\Gamma_{+}}= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \bar{v} \tag{D.16}
\end{align*}
$$

for $w \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$.
Proof. We proceed in a similar way as in Lemma 46 and get from equation (D.1),

$$
\begin{align*}
& \int_{\Omega \backslash i n t} \nabla \Omega_{3} \\
& \nabla U_{\alpha, 0} \cdot \nabla \bar{v}-\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0} \bar{v}-2 i \alpha \int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \bar{v}  \tag{D.17}\\
&-\int_{\partial \Omega_{3}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}-\int_{\Gamma_{L} \cup \Gamma_{R} \cup \Gamma_{ \pm}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}=0,
\end{align*}
$$

for all $v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$. Since $v$ and $U_{\alpha, 0}$ are periodic then $\int_{\Gamma_{L} \cup \Gamma_{R}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}=$ 0 . Using equations (D.2), (D.3) and (D.4) gives

$$
\begin{aligned}
& \int_{\Omega \backslash \text { int } \Omega_{3}} \nabla U_{\alpha, 0} \cdot \nabla \bar{v}-\int_{\Omega \backslash \text { int } \Omega_{3}}\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0} \bar{v}-2 i \alpha \int_{\Omega \backslash \text { int } \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \bar{v} \\
+ & i \int_{\partial \Omega_{3}} \alpha n_{x} U_{\alpha, 0} \bar{v}-\int_{\Gamma_{+}} T_{+}^{\alpha, 0} U_{\alpha, 0} \bar{v}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} U_{\alpha, 0} \bar{v}=-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \bar{v}
\end{aligned}
$$

which finishes the proof.

## - Case 2A: Transmitting dielectric grating: TE case

Lemma D-5. The variational formulation corresponding to the boundary problem given by equation (D.6) to equation (D.9) can be formulated as follows.
Find $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$, for all $v \in H_{\#}^{1}(\Omega)$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v\right)=(f, v)_{\Gamma_{+}}, \tag{D.18}
\end{equation*}
$$

with

$$
\begin{align*}
a(w, v)= & \int_{\Omega} \nabla w \cdot \nabla \bar{v}-\int_{\Omega}\left(k^{2}-\alpha^{2}\right) w \bar{v}-2 i \alpha \int_{\Omega}\left(\partial_{x} w\right) \bar{v}  \tag{D.19}\\
& -\int_{\Gamma_{+}} T_{+}^{\alpha, 0} w \bar{v}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} w \bar{v}
\end{align*}
$$

and

$$
\begin{equation*}
(f, v)_{\Gamma_{+}}=-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \bar{v} \tag{D.20}
\end{equation*}
$$

for $w \in H_{\#}^{1}(\Omega)$.
Proof. The proof is similar to that of Lemma 46.

## - Case 2B: Transmitting dielectric grating: TM case

Lemma D-6. The variational formulation corresponding to the boundary problem given by equation (D.10) to equation (D.13) can be formulated as follows.
Find $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$, for all $v \in H_{\#}^{1}(\Omega)$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v\right)=(f, v)_{\Gamma_{+}}, \tag{D.21}
\end{equation*}
$$

where

$$
\begin{align*}
a(w, v)= & \int_{\Omega} \frac{1}{k^{2}} \nabla w \cdot \nabla \bar{v}-\int_{\Omega} \frac{1-\alpha^{2}}{k^{2}} w \bar{v}-i \alpha \int_{\Omega} \frac{1}{k^{2}}\left(\partial_{x} w\right) \bar{v}  \tag{D.22}\\
& +i \alpha \int_{\Omega} \frac{1}{k^{2}}(w) \overline{\partial_{x} v}-\int_{\Gamma_{+}} \frac{1}{k_{1}^{2}} T_{+}^{\alpha, 0} w \bar{v}-\int_{\Gamma_{-}} \frac{1}{k_{2}^{2}} T_{-}^{\alpha, 0} w \bar{v},
\end{align*}
$$

and

$$
(f, v)_{\Gamma_{+}}=-\int_{\Gamma_{+}} \frac{2 i \beta_{1}^{0}}{k_{1}^{2}} e^{-i \beta_{1}^{0} B} \bar{v}
$$

for $w \in H_{\#}^{1}(\Omega)$.
Proof. Multiplying both sides of equation (2.105) by $\overline{v_{\alpha}}$, such that $\overline{v_{\alpha}}=e^{-i \alpha x} \bar{v}$, and integrating gives

$$
\int_{\Omega} \nabla \cdot\left(\frac{1}{k^{2}} \nabla U\right) \overline{v_{\alpha}}+\int_{\Omega} U \overline{v_{\alpha}}=0,
$$

for all $v_{\alpha} \in H_{\alpha \#}^{1}(\Omega)$. We integrate by parts to get

$$
\int_{\Omega} \frac{1}{k^{2}} \nabla U \cdot \nabla \overline{v_{\alpha}}-\int_{\Omega} U \overline{v_{\alpha}}-\int_{\partial \Omega} \frac{1}{k^{2}} \frac{\partial U}{\partial n} \overline{v_{\alpha}}=0 .
$$

Then,

$$
\begin{equation*}
\int_{\Omega} \frac{1}{k^{2}} \nabla U . \nabla \overline{v_{\alpha}}-\int_{\Omega} U \overline{v_{\alpha}}-\int_{\Gamma_{L} \cup \Gamma_{R} \cup \Gamma_{ \pm}} \frac{1}{k^{2}} \frac{\partial U}{\partial n} \overline{v_{\alpha}}=0 . \tag{D.23}
\end{equation*}
$$

with $\Gamma_{L}, \Gamma_{R}$ as defined in equations (3.47) and (3.48). Let $n_{\Gamma_{L}}\left(n_{\Gamma_{R}}\right)$ denote the exterior unit normal vector on $\Gamma_{L}$ (the exterior unit normal vector on $\Gamma_{R}$ ) and note that $n_{\Gamma_{L}}=-n_{\Gamma_{R}}$, and also $k_{\Gamma_{L}}=k_{\Gamma_{R}}$ from the geometry of our scattering problem (see Figure 2.3). Since $v$ and $U_{\alpha, 0}$ are periodic then

$$
\begin{aligned}
\int_{\Gamma_{L} \cup \Gamma_{R}} \frac{1}{k^{2}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v} & =\int_{\Gamma_{L}} \frac{1}{k^{2}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}+\int_{\Gamma_{R}} \frac{1}{k^{2}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v} \\
& =\int_{\Gamma_{L}} \frac{1}{k^{2}}\left(\frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}-\frac{\partial U_{\alpha, 0}}{\partial n} \bar{v}\right) \\
& =0 .
\end{aligned}
$$

Hence, equation (D.23) becomes

$$
\int_{\Omega} \frac{1}{k^{2}} \nabla U \cdot \nabla \overline{v_{\alpha}}-\int_{\Omega} U \overline{v_{\alpha}}-\int_{\Gamma_{ \pm}} \frac{1}{k^{2}} \frac{\partial U}{\partial n} \overline{v_{\alpha}}=0 .
$$

Since $U=e^{i \alpha x} U_{\alpha, 0}$ and $v_{\alpha}=e^{i \alpha x} v$, we have

$$
\begin{aligned}
& \int_{\Omega} \frac{1}{k^{2}} \nabla U \cdot \nabla \overline{v_{\alpha}}-\int_{\Omega} U \overline{v_{\alpha}} \\
= & \int_{\Omega} \frac{1}{k^{2}} \nabla U_{\alpha, 0} \cdot \nabla \bar{v}+\int_{\Omega} \frac{\alpha^{2}}{k^{2}} U_{\alpha, 0} \bar{v}-i \alpha \int_{\Omega} \frac{1}{k^{2}}\left(\partial_{x} U_{\alpha, 0}\right) \bar{v} \\
& +i \alpha \int_{\Omega} \frac{1}{k^{2}} U_{\alpha, 0} \overline{\partial_{x} v}-\int_{\Omega} U_{\alpha, 0} \bar{v}-\int_{\Gamma_{ \pm}} \frac{1}{k^{2}} \frac{\partial U_{\alpha, 0}}{\partial n} \bar{v} \\
= & 0
\end{aligned}
$$

and we finish the proof by using equations (D.11) and (D.12).

## D.1.2 Equivalence of the variational forms for the periodic and $\alpha$-quasi periodic problems

We want to show that the periodic problems are well posed. We also want to establish an upper bound on the error that arises when we solve the scattering problem numerically. For these reasons, we need to study the equivalence of the variational forms for the periodic and $\alpha$-quasi periodic problems. Hence, the three cases can be described as follows.

## - Case 1B: Perfectly conducting grating: TM case

Similarly to Lemma D-4, for the periodic function $U_{\alpha, 0}$, let

$$
\begin{align*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)= & \left(\nabla U_{\alpha, 0}, \nabla v_{\alpha}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-2 i \alpha\left(\partial_{x} U_{\alpha, 0}, v_{\alpha}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}+\left(i \alpha n_{x} U_{\alpha, 0}, v_{\alpha}\right)_{\partial \Omega_{3}} \\
& -\left(\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}, v_{\alpha}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, v_{\alpha}\right)_{\Gamma_{ \pm}},  \tag{D.24}\\
\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}}= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{\alpha}} .
\end{align*}
$$

The variational problem is to find $U_{\alpha, 0} \in H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ for all $v_{\alpha} \in H_{\#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)=\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}} . \tag{D.25}
\end{equation*}
$$

Similarly, for the $\alpha$-quasi periodic function $U$ let

$$
\begin{align*}
a(U, v) & =(\nabla U, \nabla v)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} U, v\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}-\left(T_{ \pm} U, v\right)_{\Gamma_{ \pm}}  \tag{D.26}\\
(f, v)_{\Gamma_{+}} & =-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{v} . \tag{D.27}
\end{align*}
$$

The variational problem is to find $U \in H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$ such that

$$
\begin{equation*}
a(U, v)=(f, v)_{\Gamma_{+}} . \tag{D.28}
\end{equation*}
$$

We have the following result.
Lemma D-7. Finding $U_{\alpha, 0} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ for all $v_{\alpha} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ such that a $\left(U_{\alpha, 0}, v_{\alpha}\right)=\left(f_{\alpha}, v_{\alpha}\right)$ as given in equation (D.25) is equivalent to finding $U \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ such that $a(U, v)=$ $(f, v)_{\Gamma_{+}}$using equation (D.28).

Proof. The proof is similar to Lemma 47.

## - Case 2A: Transmitting dielectric grating: TE case

Similarly to Lemma D-5, for the periodic function $U_{\alpha, 0}$, let

$$
\begin{align*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)= & \left(\nabla U_{\alpha, 0}, \nabla v_{\alpha}\right)_{\Omega}-2 i \alpha\left(\partial_{x} U_{\alpha, 0}, v_{\alpha}\right)_{\Omega} \\
& -\left(\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}, v_{\alpha}\right)_{\Omega}-\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, v_{\alpha}\right)_{\Gamma_{ \pm}},  \tag{D.29}\\
\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}}= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{\alpha}} .
\end{align*}
$$

The variational problem is to find $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$ for all $v_{\alpha} \in H_{\#}^{1}(\Omega)$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)=\left(f_{\alpha}, v_{\alpha}\right) \tag{D.30}
\end{equation*}
$$

Similarly, for the $\alpha$-quasi periodic function $U$ let

$$
\begin{align*}
a(U, v) & =(\nabla U, \nabla v)_{\Omega}-\left(k^{2} U, v\right)_{\Omega}-\left(T_{ \pm} U, v\right)_{\Gamma_{ \pm}}  \tag{D.31}\\
(f, v)_{\Gamma_{+}} & =-\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{v} . \tag{D.32}
\end{align*}
$$

The variational problem is to find $U \in H_{\alpha \#}^{1}(\Omega)$ for all $v \in H_{\alpha \#}^{1}(\Omega)$ such that

$$
\begin{equation*}
a(U, v)=(f, v)_{\Gamma_{+}} . \tag{D.33}
\end{equation*}
$$

We then have the following result.

Lemma D-8. Finding $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$ for all $v_{\alpha} \in H_{\#}^{1}(\Omega)$ such that $a\left(U_{\alpha, 0}, v_{\alpha}\right)=$ $\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}}$as given in equation (D.30) is equivalent to finding $U \in H_{\alpha \#}^{1}(\Omega)$ for all $v \in H_{\alpha \#}^{1}(\Omega)$ such that $a(U, v)=(f, v)_{\Gamma_{+}}$using equation (D.33).

Proof. By using $\Omega$ instead of $\Omega \backslash$ int $\Omega_{3}$ we can use the same process as in the proof of Lemma 47.

- Case 2B: Transmitting dielectric grating: TM case

Similarly to Lemma D-6, for the periodic function $U_{\alpha, 0}$, let

$$
\begin{align*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)= & \left(\frac{1}{k^{2}} \nabla U_{\alpha, 0}, \nabla v_{\alpha}\right)_{\Omega}-i \alpha\left(\frac{1}{k^{2}} \partial_{x} U_{\alpha, 0}, v_{\alpha}\right)_{\Omega}  \tag{D.34}\\
& +i \alpha\left(\frac{1}{k^{2}} U_{\alpha, 0}, \partial_{x} v_{\alpha}\right)_{\Omega}-\left(\left(\frac{1-\alpha^{2}}{k^{2}}\right) U_{\alpha, 0}, v_{\alpha}\right)_{\Omega}-\left(\frac{1}{k^{2}} T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, v_{\alpha}\right)_{\Gamma_{ \pm}} \\
\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}}= & -\int_{\Gamma_{+}} \frac{2 i \beta_{1}^{0}}{k_{1}^{2}} e^{-i \beta_{1}^{0} B} \overline{v_{\alpha}} .
\end{align*}
$$

The variational problem is to find $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$ for all $v_{\alpha} \in H_{\#}^{1}(\Omega)$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0}, v_{\alpha}\right)=\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}} . \tag{D.35}
\end{equation*}
$$

Similarly, for the $\alpha$-quasi periodic function $U$ let

$$
\begin{align*}
a(U, v) & =\left(\frac{1}{k^{2}} \nabla U, \nabla v\right)_{\Omega}-(U, v)_{\Omega}-\left(\frac{1}{k^{2}} T_{ \pm} U, v\right)_{\Gamma_{ \pm}}  \tag{D.36}\\
(f, v)_{\Gamma_{+}} & =-\int_{\Gamma_{+}} \frac{2 i \beta_{1}^{0}}{k_{1}^{2}} e^{i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{v} .
\end{align*}
$$

The variational problem is to find $U \in H_{\alpha \#}^{1}(\Omega)$ for all $v \in H_{\alpha \#}^{1}(\Omega)$ such that

$$
\begin{equation*}
a(U, v)=(f, v)_{\Gamma_{+}} \tag{D.37}
\end{equation*}
$$

and we have the following result.
Lemma D-9. Finding $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$ for all $v_{\alpha} \in H_{\#}^{1}(\Omega)$ such that a $\left(U_{\alpha, 0}, v_{\alpha}\right)=$ $\left(f_{\alpha}, v_{\alpha}\right)_{\Gamma_{+}}$as given in equation (D.35) is equivalent to finding $U \in H_{\alpha \#}^{1}(\Omega)$ for all $v \in H_{\alpha \#}^{1}(\Omega)$ such that $a(U, v)=(f, v)_{\Gamma_{+}}$using equation (D.37).

Proof. From equation (4.21), the first term of the sesquilinear form given by equation (D.34) becomes

$$
\begin{equation*}
\left(\frac{1}{k^{2}} \nabla U_{\alpha, 0}, \nabla v_{\alpha}\right)_{\Omega}=\int_{\Omega} \frac{\left(\alpha^{2} U \bar{v}-i \alpha U \partial_{x} \bar{v}+i \alpha \bar{v} \partial_{x} U+\nabla U . \nabla \bar{v}\right)}{k^{2}} . \tag{D.38}
\end{equation*}
$$

Using equation (4.23) and examining the next term in the sesquilinear form given by equation (D.34) we see

$$
\begin{equation*}
-i \alpha\left(\frac{1}{k^{2}} \partial_{x} U_{\alpha, 0}, v_{\alpha}\right)_{\Omega}=\int_{\Omega} \frac{\left(-\alpha^{2} U \bar{v}-i \alpha \bar{v} \partial_{x} U\right)}{k^{2}} . \tag{D.39}
\end{equation*}
$$

We also note that

$$
\begin{align*}
i \alpha\left(v_{\alpha}, \partial_{x} w_{\alpha}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}} & =i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(e^{-i \alpha x} v\right)\left(e^{i \alpha x} \partial_{x} \bar{w}+i \alpha e^{i \alpha x} \bar{w}\right) \\
& =\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(-\alpha^{2} v \bar{w}+i \alpha v \partial_{x} \bar{w}\right) . \tag{D.40}
\end{align*}
$$

From equation (D.40), the third term of the sesquilinear form given by equation (D.34) becomes

$$
\begin{equation*}
i \alpha\left(\frac{1}{k^{2}} U_{\alpha, 0}, \partial_{x} v_{\alpha}\right)_{\Omega}=\int_{\Omega} \frac{\left(-\alpha^{2} U \bar{v}+i \alpha U \partial_{x} \bar{v}\right)}{k^{2}} . \tag{D.41}
\end{equation*}
$$

We also use equation (4.24) to get

$$
\begin{equation*}
\left(\left(1-\frac{\alpha^{2}}{k^{2}}\right) U_{\alpha, 0}, v_{\alpha}\right)_{\Omega}=\int_{\Omega}\left(1-\frac{\alpha^{2}}{k^{2}}\right) U \bar{v} \tag{D.42}
\end{equation*}
$$

and from equation (4.25), the last term of the sesquilinear form given by equation (D.34) is

$$
\begin{equation*}
\left(\frac{1}{k^{2}} T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, v_{\alpha}\right)_{\Gamma_{ \pm}}=\left(\frac{1}{k^{2}} T_{ \pm} U, \bar{v}\right)_{\Gamma_{ \pm}} \tag{D.43}
\end{equation*}
$$

Substituting equations (D.38),(D.39),(D.41),(D.42) and (D.43) into equation (D.34) we get

$$
\begin{aligned}
a\left(U_{\alpha, 0}, v_{\alpha}\right) & =\int_{\Omega}\left(\frac{1}{k^{2}} \nabla U \cdot \nabla \bar{v}-U \bar{v}\right)-\int_{\Gamma_{ \pm}} \frac{1}{k^{2}} T_{ \pm} U \bar{v} . \\
& =a(U, v)
\end{aligned}
$$

and we finish the proof of Lemma D-9 in a similar way to that in Lemma 47.

## D.1.3 Well posedness of the variational problem

Before solving the variational formulation numerically, we show that our problem is well posed like we have done for Case 1A in Section 4.1.3.

## D.1.3.1 Existence and uniqueness of the solution

## - Case 1B: Perfectly conducting grating: TM case

Lemma D-10. For all $v \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, the solution $U_{\alpha, 0} \in H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ which satisfies equation (D.14) exists and is unique.

Proof. Since the TE case and TM case for perfectly conducting only differ on the boundary condition on $\partial \Omega_{3}$, and they both have the same sesquilinear form, then we can use the same arguments as in Lemma 48 to show Lemma D10.

- Case 2A: Transmitting dielectric grating: TE case

Lemma D-11. Let $k_{r e f}>0$ such that $|k|>k_{\text {ref }}$. For all $v \in H_{\#}^{1}(\Omega)$, the solution $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$ which satisfies equation (D.18) exists and is unique.

Proof. Note that we can obtain the sesquilinear form for the TE case for the transmitting dielectric grating from the sesquilinear form for the TE case for the perfectly conducting grating by replacing $\Omega \backslash$ int $\Omega_{3}$ with $\Omega$. We can then use the same arguments as in Lemma 48 to show Lemma D-11.

## - Case 2B: Transmitting dielectric grating: TM case

Lemma D-12. Let $k_{\text {ref }}>0$ such that $|k|>k_{\text {ref }}$. For all $v \in H_{\#}^{1}(\Omega)$, the solution $U_{\alpha, 0} \in H_{\#}^{1}(\Omega)$ which satisfies equation (D.21) exists and is unique.

Proof. We note by using Cauchy-Schwarz inequality [22, p. 50] that

$$
\begin{align*}
\left|\left(\frac{1}{k^{2}} \nabla U, \nabla v\right)_{\Omega}\right| & \leq \frac{1}{k_{r e f}^{2}} \int_{\Omega}|\nabla U \cdot \nabla \bar{v}| d x d y  \tag{D.44}\\
& \leq \frac{1}{k_{r e f}^{2}} \mathrm{k} \nabla U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k} \nabla v \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}, \tag{D.45}
\end{align*}
$$

and from equation (4.30) we get

$$
\begin{equation*}
\left|(U, v)_{\Omega}\right| \leq \mathrm{k} U \mathrm{k}_{L_{\alpha \neq 1}^{2}(\Omega)} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} . \tag{D.46}
\end{equation*}
$$

From equation (4.33), we have

$$
\begin{equation*}
\left|\int_{\Gamma_{ \pm}} \frac{1}{k^{2}} T_{ \pm} U \bar{v} d x\right|^{2} \leq C d \frac{1}{k_{r e f}^{2}}\left(\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{L_{\alpha \neq \#}^{2}(\Omega)}^{2}+\mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{1}(\Omega)}^{2}\right) \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{1}(\Omega)}^{2} . \tag{D.47}
\end{equation*}
$$

Hence, we have from equation (D.36)

$$
\begin{aligned}
|a(U, v)| \leq & \frac{1}{k_{r e f}^{2}}|U|_{H_{\alpha \#}^{1}(\Omega)}|v|_{H_{\alpha \#}^{1}(\Omega)}+{\mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}{\mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}}}+C d \frac{1}{k_{r e f}^{2}}\left(\left|k_{j}^{2}\right| \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}^{2}+\mathrm{k} U \mathrm{k}_{H_{\alpha \#}^{2}(\Omega)}^{2}\right) \mathrm{k} v \mathrm{k}_{H_{\alpha \#}^{2}(\Omega)}^{2},
\end{aligned}
$$

and so

Hence, $a(U, U)$ is continuous using Definition A-6. Taking the real part of $a(U, U)$ and we get

$$
\begin{equation*}
\mathfrak{R}(a(U, U))=\int_{\Omega} \mathfrak{R}\left(\frac{1}{k^{2}}\right)|\nabla U|^{2}-\int_{\Omega}|U|^{2}-\int_{\Gamma_{ \pm}} \mathfrak{R}\left(\frac{1}{k^{2}} T U \bar{U}\right) . \tag{D.49}
\end{equation*}
$$

Hence,

$$
\mathfrak{R}(a(U, U))+\int_{\Gamma_{ \pm}} \Re\left(\frac{1}{k^{2}} T_{ \pm} U \bar{U}\right)+\int_{\Omega}|U|^{2} \geq \frac{1}{\mathrm{k} k \mathrm{k}_{L^{\infty}(\Omega)}^{2}}|U|_{H_{\alpha \#}^{1}(\Omega)}^{2}
$$

and we use equation (D.47) to get

$$
\begin{aligned}
\int_{\Gamma_{ \pm}} \Re\left(\frac{1}{k^{2}} T_{ \pm} U \bar{U}\right) & \leq\left|\int_{\Gamma_{ \pm}}\left(\frac{1}{k^{2}} T_{ \pm} U \bar{U}\right)\right| \\
& \leq \sqrt{C d} \frac{1}{k_{r e f}}\left(\left|k_{j}\right| \mathrm{k} U \mathrm{k}_{L_{\alpha \neq}^{2}(\Omega)}+\mathrm{k} U \mathrm{k}_{H_{\alpha \neq}^{1}(\Omega)}\right) \mathrm{k} U \mathrm{k}_{H_{\alpha \neq}^{1}(\Omega)} \\
& \leq C \bar{C} \overline{C d} \frac{\left|k_{j}\right|}{k_{\text {ref }}} \mathrm{k} U \mathrm{k}_{L_{\alpha \neq \#}^{2}(\Omega)}^{2}
\end{aligned}
$$

from the equivalence of the norm in $H_{\alpha \#}^{l}(\Omega)$ for $l \geq 0$. Hence
from the equivalence of the norm in $H_{\alpha \#}^{l}(\Omega)$ for $l \geq 0$. Then, $a(U, U)$ is $H_{\alpha \#}^{1}(\Omega)$-coercive from Definition B-4. We can then use Lemma B-5 to show the existence of a solution from its uniqueness. The rest of the proof is similar to Lemma 48. Let us suppose that we have two solutions $U_{1}$ and $U_{2}$ and let us denote $w=U_{1}-U_{2}$. We have

$$
\begin{equation*}
a(w, w)=\int_{\Omega} \frac{1}{k^{2}}|\nabla w|^{2}-\int_{\Omega}|w|^{2}-\int_{\Gamma_{ \pm}} \frac{1}{k^{2}}\left(T_{ \pm} w\right) \bar{w}=0 . \tag{D.50}
\end{equation*}
$$

Since $\Re\left(\frac{1}{k^{2}}\right)$ can be positive or negative, we have to deal separately with each case. First, if $\mathfrak{R}\left(\frac{1}{k^{2}}\right)>0$, we can take the imaginary part of equation (D.50) to get

$$
\begin{align*}
& \int_{\Omega} \mathfrak{I}\left(\frac{1}{k^{2}}\right)|\nabla w|^{2}-\int_{\Gamma_{ \pm}} \mathfrak{J}\left(\frac{1}{k^{2}}\right) \mathfrak{R}\left(\left(T_{ \pm} w\right) \bar{w}\right) \\
& -\int_{\Gamma_{ \pm}} \mathfrak{R}\left(\frac{1}{k^{2}}\right) \mathfrak{J}\left(\left(T_{ \pm} w\right) \bar{w}\right)=0 . \tag{D.51}
\end{align*}
$$

By noting that $\mathfrak{J}\left(\frac{1}{k^{2}}\right)<0$ and using equation (2.72), we have

$$
\begin{align*}
& \int_{\Omega} \mathfrak{I}\left(\frac{1}{k^{2}}\right)|\nabla w|^{2}-\int_{\Gamma_{ \pm}} \mathfrak{I}\left(\frac{1}{k^{2}}\right) \mathfrak{R}\left(\left(T_{ \pm} w\right) \bar{w}\right) \\
& -\int_{\Gamma_{ \pm}} \mathfrak{R}\left(\frac{1}{k^{2}}\right) \mathfrak{J}\left(\left(T_{ \pm} w\right) \bar{w}\right) \leq 0 . \tag{D.52}
\end{align*}
$$

From equations (D.51) and (D.52) we have $w=0$ and so $U_{1}=U_{2}$. Secondly, if $\mathfrak{R}\left(\frac{1}{k^{2}}\right)<0$, we can take the real part of equation (D.50) to get

$$
\begin{align*}
& \int_{\Omega} \mathfrak{R}\left(\frac{1}{k^{2}}\right)|\nabla w|^{2}-\int_{\Omega}|w|^{2}-\int_{\Gamma_{ \pm}} \mathfrak{R}\left(\frac{1}{k^{2}}\right) \mathfrak{R}\left(\left(T_{ \pm} w\right) \bar{w}\right) \\
& +\int_{\Gamma_{ \pm}} \mathfrak{I}\left(\frac{1}{k^{2}}\right) \mathfrak{J}\left(\left(T_{ \pm} w\right) \bar{w}\right)=0 . \tag{D.53}
\end{align*}
$$

Once more by noting that $\mathfrak{J}\left(\frac{1}{k^{2}}\right)<0$ and using equation (2.72), we have

$$
\begin{align*}
& \int_{\Omega} \mathfrak{R}\left(\frac{1}{k^{2}}\right)|\nabla w|^{2}-\int_{\Omega}|w|^{2}-\int_{\Gamma_{ \pm}} \mathfrak{R}\left(\frac{1}{k^{2}}\right) \mathfrak{R}\left(\left(T_{ \pm} w\right) \bar{w}\right) \\
& +\int_{\Gamma_{ \pm}} \mathfrak{J}\left(\frac{1}{k^{2}}\right) \mathfrak{J}\left(\left(T_{ \pm} w\right) \bar{w}\right) \leq 0 . \tag{D.54}
\end{align*}
$$

From equations (D.53) and (D.54) we have $w=0$ and so $U_{1}=U_{2}$.
We finish the proof by noting that $U_{1}=e^{i \alpha x} U_{\alpha, 0_{1}}$ and $U_{2}=e^{i \alpha x} U_{\alpha, 0_{2}}$ and since $e^{i \alpha x} 60$ then $U_{\alpha, 0_{1}}=U_{\alpha, 0_{2}}$.

To show the continuous dependence of the variational formulation on the data, it is necessary to investigate the regularity estimate if the variational formulation. This is done below.

## D.1.3.2 Regularity estimate of the exact solution

In a similar fashion to that proposed in Section 4.1.3.2, we derive an explicit dependency on $k$ in the proof of the regularity of the solution and in the a priori error estimate. In this latter, we also derive an explicit dependency on the degree of the polynomial basis $p$. In the following theorem, we obtain a new result on the regularity estimate for the solution $U_{\alpha, 0}$ for Cases $1 \mathrm{~B}, 2 \mathrm{~A}$ and 2 B .

- Case 1B: Perfectly conducting grating: TM case

Theorem D-13. Let $f_{\alpha} \in H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ and let $U_{\alpha, 0} \in H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be the solution of

$$
\begin{array}{rlrl}
\Delta U_{\alpha, 0}+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0} & =f_{\alpha}, & & \text { in } \Omega \backslash \text { int } \Omega_{3},  \tag{D.55}\\
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, & & \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, & \text { on } \Gamma_{-}, \\
\partial_{n} U_{\alpha, 0} & =-i \alpha U_{\alpha, 0} n_{x}, & & \text { on } \partial \Omega_{3} .
\end{array}
$$

Then there exists a constant $C_{\text {stab }}$ which is independent of the wavenumber $k$ such that

$$
\mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq C_{\text {stab }} \mathrm{k} f_{\alpha} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} .
$$

Proof. We proceed similarly as in Theorem 51 but we use the regularity estimate given by Theorem C-3.

## - Case 2A: Transmitting dielectric grating: TE case

Theorem D-14. Let $f_{\alpha} \in H_{\#}^{l}(\Omega)$ and let $U_{\alpha, 0} \in H_{\#}^{l}(\Omega)$ be the solution of

$$
\begin{align*}
\Delta U_{\alpha, 0}+\left(k^{2}-\alpha^{2}\right) U_{\alpha, 0}+2 i \alpha \partial_{x} U_{\alpha, 0} & =f_{\alpha}, \text { in } \Omega,  \tag{D.56}\\
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, \text { on } \Gamma_{+}, \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, \text { on } \Gamma_{-} .
\end{align*}
$$

Then there exists a constant $C_{\text {stab }}$ which is dependent of the wavenumbers $k_{1}$ and $k_{3}$ such that

$$
\mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq C_{s t a b} \mathrm{k} f_{\alpha} \mathrm{k}_{L_{\#}^{2}(\Omega)}
$$

where $C_{\text {stab }}=C_{\text {reg }}\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right)$ and $C_{s} C\left(k_{0}, k_{3}\right)$ is as defined in Lemma 31.
Proof. We proceed similarly as in Theorem 51 but we use the regularity estimate given by Theorem 33.

## - Case 2B: Transmitting dielectric grating: TM case

Theorem D-15. Let $f_{\alpha} \in H_{\#}^{l}(\Omega)$ and let $U_{\alpha, 0} \in H_{\#}^{l}(\Omega)$ be the solution of

$$
\begin{align*}
\nabla_{\alpha \cdot}\left(\frac{1}{k^{2}} \nabla_{\alpha} U_{\alpha, 0}\right)+U_{\alpha, 0} & =f_{\alpha}, \text { in } \Omega  \tag{D.57}\\
\left(T_{+}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, \text { on } \Gamma_{+} \\
\left(T_{-}^{\alpha, 0}-\frac{\partial}{\partial \eta}\right) U_{\alpha, 0} & =0, \text { on } \Gamma_{-}
\end{align*}
$$

Then there exists a constant $C_{\text {stab }}$ which is dependent on the wave numbers $k_{0}$ and $k_{3}$ such that

$$
\mathrm{k} U_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq C_{s t a b} \mathrm{k} f_{\alpha} \mathrm{k}_{L_{\#}^{2}(\Omega)}
$$

where $C_{\text {stab }}=C_{\text {reg }}\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right)$ and $C_{s} C\left(k_{0}, k_{3}\right)$ is as defined in Theorem C-6.

Proof. We proceed similarly as in Theorem 51 but we use the regularity estimate given by Theorem C-6.

## D.1.4 Efficiency of the diffraction grating

We can use the definition of the grating efficiency given in Section 4.1.4 and use the variational formulation given by equations (D.15), (D.19) and (D.22) to show that the energy is conserved for each of these cases. Before doing so, we need the following results.

Lemma D-16. For Case 1B, we have

$$
\begin{equation*}
\mathfrak{R}\left(\int_{\Omega \backslash \mathrm{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}}\right)=0 \tag{D.58}
\end{equation*}
$$

and we also have for Case 2 that

$$
\begin{equation*}
\mathfrak{R}\left(\int_{\Omega}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}}\right)=0 . \tag{D.59}
\end{equation*}
$$

Proof. For Case 1B, let us define $\Omega_{x}$ where $\Omega_{x} \times[-B, B]=\Omega \backslash \operatorname{int} \Omega_{3}$. We have

$$
\int_{\Omega \backslash i \operatorname{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=\int_{-B}^{B}\left(\int_{\Omega_{x}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x\right) d y
$$

We note by integrating by parts that

$$
\int_{\Omega_{x}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x=\left[U_{\alpha, 0} \overline{U_{\alpha, 0}}\right]_{\partial \Omega_{x}}-\int_{\Omega_{x}} U_{\alpha, 0} \overline{\partial_{x} U_{\alpha, 0}} d x
$$

and so

$$
\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=\int_{-B}^{B}\left(\left[U_{\alpha, 0} \overline{U_{\alpha, 0}}\right]_{\partial \Omega_{x}}-\int_{\Omega_{x}} U_{\alpha, 0} \overline{\partial_{x} U_{\alpha, 0}} d x\right) d y
$$

If $\partial \Omega_{x}=0$ or $d$, then we can use the periodicity of $U_{\alpha, 0}$ and $I=0$. If $\partial \Omega_{x} \subset \partial \Omega_{3}$ and if $\partial \Omega_{3}$ is a closed curve as shown in Figure 3.1 then we can use Cauchy's theorem to show that $I=0$. If $\partial \Omega_{x} \subset \partial \Omega_{3}$ and if $\partial \Omega_{3}$ is an open curve then we can use the path $\partial \Omega_{3} \cup L_{j}$ as shown in Figure 3.2 with Cauchy's theorem to show
that $I=0$. Hence, we get equation (4.52) and for any $c \in \mathbb{C}$, if $c=-\bar{c}$ then $\mathfrak{R}(c)=0$. Hence we get equation (D.58). For Case 2, we note that

$$
\int_{\Omega}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=\int_{-B}^{B}\left(\int_{0}^{d}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x\right) d y
$$

We integrate by parts to get

$$
\int_{\Omega}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=\int_{-B}^{B}\left(\left[U_{\alpha, 0}{\overline{U_{\alpha, 0}}}_{0}^{d}-\int_{\Omega} U_{\alpha, 0} \overline{\partial_{x} U_{\alpha, 0}} d x\right) d y\right.
$$

Since $U_{\alpha, 0}$ is periodic then $\left[U_{\alpha, 0}{\overline{U_{\alpha, 0}}}_{0}^{d}=0\right.$ and so

$$
\int_{\Omega}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} d x d y=-\int_{\Omega} U_{\alpha, 0} \overline{\partial_{x} U_{\alpha, 0}} d x d y
$$

and the result follows.

## D.1.5 Conservation of the energy or energy balance

We again define respectively by $E_{r}, E_{t}$ and $E_{a b s}$, the refracted energy, the transmitted energy and the absorbed energy by the grating. For Cases 1 B and 2A, their definitions are given in Definition 54.

- Case 1B: Perfectly conducting grating: TM case

Theorem D-17 (Conservation of energy). Let $E_{t}, E_{r}$, and $E_{\text {abs }}$ defined as in Definition 54. Then, we have the energy balance

$$
E_{t}+E_{r}=1
$$

Proof. Note that $\mathfrak{I} \int_{\partial \Omega_{3}} i \alpha n_{x}\left|U_{\alpha, 0}\right|^{2}=0$ since $\partial \Omega_{3}$ is an open curve (see Figure 2.4), $U_{\alpha, 0}$ is periodic and so the integral is zero. If $\partial \Omega_{3}$ is a closed curve (see Figure 2.3) then using Cauchy's theorem, the integral is zero and the rest of the proof is similar to Lemma 56.

- Case 2A: Transmitting dielectric grating: TE case

Theorem D-18 (Conservation of energy). Let $E_{t}, E_{r}$, and $E_{a b s}$ as given by Definition 54. Then, we have the energy balance

$$
E_{t}+E_{r}+E_{a b s}=1
$$

Proof. We take the imaginary part of equation (D.29), and we use equation (4.51) with Lemma 8 to get

$$
\begin{equation*}
\left(\mathfrak{I}\left(k^{2}\right) U_{\alpha, 0}, U_{\alpha, 0}\right)_{\Omega}+\mathfrak{I}\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, U_{\alpha, 0}\right)_{\Gamma_{ \pm}}-2 \beta_{1}^{0} \Re\left(e^{-i \beta_{1}^{0} B}, U_{\alpha, 0}\right)_{\Gamma_{+}}=0 \tag{D.60}
\end{equation*}
$$

We proceed similarly as in Lemma 56. Hence, we use Lemma 55 with equation (2.72) in Lemma 8 to get equation (4.55) which is

$$
\mathfrak{J}\left(T_{ \pm}^{\alpha, 0} U_{\alpha, 0}, U_{\alpha, 0}\right)_{\Gamma_{ \pm}}+2 \beta_{1}^{0} \Re\left(e^{-i \beta_{1}^{0} B}, U_{\alpha, 0}\right)_{\Gamma_{+}}=d \beta_{1}^{0} \sum_{n_{\alpha}^{2}>k^{2}} R_{m}+T_{m}-d \beta_{1}^{0} .
$$

We use equations (D.60) and (4.55) to get

$$
\begin{equation*}
d \beta_{1}^{0}\left(\sum_{n_{\alpha}^{2}>k^{2}} R_{m}+T_{m}-1\right)+\left(\mathfrak{J}\left(k^{2}\right) U_{\alpha, 0}, U_{\alpha, 0}\right)_{\Omega}=0 \tag{D.61}
\end{equation*}
$$

which finishes the proof.

## - Case 2B: Transmitting dielectric grating: TM case

Definition D-19. For Case 2B, the refracted energy, the transmitted energy and the absorbed energy by the grating are given respectively by

$$
\begin{aligned}
E_{r}= & \sum_{\left(\frac{2 \pi m}{d}+\alpha\right)^{2}<k_{1}^{2}} R_{m}, \\
E_{t}= & \frac{k_{1}^{2}}{\mathfrak{R}\left(k_{2}^{2}\right)} \sum_{\left(\frac{2 \pi m}{d}+\alpha\right)^{2}<k_{2}^{2}} T_{m}, \\
E_{a b s}= & \int_{\Omega} \frac{1}{\bar{J}\left(k^{2}\right)}\left|\nabla U_{\alpha, 0}\right|^{2}+\int_{\Omega} \frac{2 \alpha}{\bar{J}\left(k^{2}\right)}\left(\partial_{x} U_{\alpha, 0}\right) \overline{U_{\alpha, 0}} \\
& +\int_{\Omega} \frac{\alpha^{2}}{\mathfrak{J}\left(k^{2}\right)}\left|U_{\alpha, 0}\right|^{2}+d \sum_{(m+\alpha)^{2}>k_{2}^{2}}\left|U_{\alpha, 0}(m)(-B)\right|^{2}
\end{aligned}
$$

such that $R_{m}$ and $T_{m}$ are as given by equations (4.49) and (4.50).
Theorem D-20 (Conservation of energy). Let $E_{r}, E_{t}$ and $E_{a b s}$ be defined as in Definition D-19. The energy is conserved and we have the energy balance

$$
E_{t}+E_{r}+E_{a b s}=1
$$

The theorem can be proved by following the same process given for the Case 2A in Theorem D-18. We just need to use equation (D.22) instead of equation (D.19).

## D. 2 The discrete problem

In order to solve numerically the scattering problem, we need to discretize the variational formulation corresponding to the continuous problem as we did for Case 1A in Section 4.2.

## D.2.1 Variational formulation

## - Case 1B: Perfectly conducting grating: TM case

Let $X \subset H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be a finite element space with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, \ldots, N$, be a basis of $X$. We discretize the variational form given by the equation (D.14) and we want to find $U_{\alpha, 0_{h}} \in X$, for all $v_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0_{h}}, v_{h}\right)=\left(f, v_{h}\right), \tag{D.62}
\end{equation*}
$$

where

$$
\begin{align*}
a\left(w_{h}, v_{h}\right)= & \int_{\Omega \backslash \operatorname{int} \Omega_{3}} \nabla w_{h} \cdot \nabla \overline{v_{h}}-\int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(k^{2}-\alpha^{2}\right) w_{h} \overline{v_{h}}-2 i \alpha \int_{\Omega \backslash \operatorname{int} \Omega_{3}}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& +i \int_{\partial \Omega_{3}} \alpha n_{x} w_{h} \overline{v_{h}}-\int_{\Gamma_{+}} T_{+}^{\alpha, 0} w_{h} \overline{v_{h}}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} w_{h} \overline{v_{h}}  \tag{D.63}\\
\left(f, v_{h}\right)= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{h}}, \tag{D.64}
\end{align*}
$$

for $w_{h} \in X$. Similarly as Case 1A , there exists $U_{j}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, 0_{h}}=\sum_{j=1}^{N} U_{j} \psi_{j}(x, y)$. Hence, the discrete problem given by equation (D.62) is equivalent to the following linear algebraic system

$$
\begin{equation*}
A U=L \tag{D.65}
\end{equation*}
$$

with $U=U_{j}$ for $j=1, \cdots, N$,

$$
A=a\left(\psi_{i}, \psi_{j}\right)
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

- Case 2A: Transmitting dielectric grating: TE case

Let $X \subset H_{\#}^{1}(\Omega)$ be a finite element space with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, . ., N$, be a basis of $X$. We discretize the variational form given by the equation (D.18) and we want to find $U_{\alpha, 0_{h}} \in X$, for all $v_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0_{h}}, v_{h}\right)=\left(f, v_{h}\right), \tag{D.66}
\end{equation*}
$$

where

$$
\begin{align*}
a\left(w_{h}, v_{h}\right)= & \int_{\Omega} \nabla w_{h} . \nabla \overline{v_{h}}-\int_{\Omega}\left(k^{2}-\alpha^{2}\right) w_{h} \overline{v_{h}}-2 i \alpha \int_{\Omega}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& -\int_{\Gamma_{+}} T_{+}^{\alpha, 0} w_{h} \overline{v_{h}}-\int_{\Gamma_{-}} T_{-}^{\alpha, 0} w_{h} \overline{v_{h}}  \tag{D.67}\\
\left(f, v_{h}\right)= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{h}} \tag{D.68}
\end{align*}
$$

for $w_{h} \in X$. Once more there exists $U_{j}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, 0_{h}}=$ $\sum_{j=1}^{N} U_{j} \psi_{j}(x, y)$. Hence, the discrete problem given by equation (D.66) is equivalent to the following linear algebraic system

$$
\begin{equation*}
A U=L \tag{D.69}
\end{equation*}
$$

with $U=U_{j}$ for $j=1, \cdots, N$,

$$
A=a\left(\psi_{i}, \psi_{j}\right)
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

## - Case 2B: Transmitting dielectric grating: TM case

Let $X \subset H_{\#}^{1}(\Omega)$ be a finite element space with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, \ldots, N$, be a basis of $X$. We discretize the variational form given by the equation (D.21) and we want to find $U_{\alpha, 0_{h}} \in X$, for all $v_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{\alpha, 0_{h}}, v_{h}\right)=\left(f, v_{h}\right), \tag{D.70}
\end{equation*}
$$

where

$$
\begin{align*}
& a\left(w_{h}, v_{h}\right) \\
& =\int_{\Omega} \frac{1}{k^{2}} \nabla w_{h} \cdot \nabla \overline{v_{h}}-\int_{\Omega} \frac{1-\alpha^{2}}{k^{2}} w_{h} \overline{v_{h}}-i \alpha \int_{\Omega} \frac{1}{k^{2}}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& +i \alpha \int_{\Omega} \frac{1}{k^{2}}\left(w_{h}\right) \overline{\partial_{x} v_{h}}-\int_{\Gamma_{+}} \frac{1}{k_{1}^{2}} T_{+}^{\alpha, 0} w_{h} \overline{v_{h}}-\int_{\Gamma_{-}} \frac{1}{k_{2}^{2}} T_{-}^{\alpha, 0} w_{h} \overline{v_{h}}, \tag{D.71}
\end{align*}
$$

and

$$
\begin{equation*}
\left(f, v_{h}\right)=-\int_{\Gamma_{+}} \frac{2 i \beta_{1}^{0}}{k_{1}^{2}} e^{-i \beta_{1}^{0} B} \overline{v_{h}}, \tag{D.72}
\end{equation*}
$$

for $w_{h} \in X$. Once more there exists $U_{j}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, 0_{h}}=$ $\sum_{j=1}^{N} U_{j} \psi_{j}(x, y)$. Hence, the discrete problem given by equation (D.70) is equivalent to the following linear algebraic system

$$
\begin{equation*}
A U=L \tag{D.73}
\end{equation*}
$$

with $U=U_{j}$ for $j=1, \cdots, N$,

$$
A=a\left(\psi_{i}, \psi_{j}\right)
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

## D.2.2 Truncation of the DtN map

The DtN operators that we use as transparent boundary conditions are truncated for computational purposes as shown in Section 4.2.2and we approximate $T_{ \pm}^{\alpha, 0}$ in equation (4.7) with equation (4.61). Therefore, instead of solving directly equations (D.62), (D.66) and (D.70) for Cases 1B, 2A and 2B, we approximate $U_{\alpha, 0}$ by $U_{\alpha, 0_{h}}^{M}$ and we solve numerically the following problems.

## - Case 1B: Perfectly conducting grating: TM case

Let $X \subset H_{\#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be a finite element subspace with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, . ., N$, be a basis of $X$. We want to find $U_{\alpha, 0_{h}}^{M} \in X$, for all $v_{h} \in X$ such that

$$
\begin{equation*}
a^{M}\left(U_{\alpha, 0_{h}}^{M}, v_{h}\right)=\left(f, v_{h}\right), \tag{D.74}
\end{equation*}
$$

where

$$
\begin{align*}
a^{M}\left(w_{h}, v_{h}\right)= & \int_{\Omega \backslash \text { int } \Omega_{3}} \nabla w_{h} . \nabla \overline{v_{h}}-\int_{\Omega \backslash \text { int } \Omega_{3}}\left(k^{2}-\alpha^{2}\right) w_{h} \overline{v_{h}}-2 i \alpha \int_{\Omega \backslash \text { int } \Omega_{3}}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& +i \int_{\partial \Omega_{3}} \alpha n_{x} w_{h} \overline{v_{h}}-\int_{\Gamma_{ \pm}} T_{ \pm}^{\alpha, 0^{M}} w_{h} \overline{v_{h}}, \\
\left(f, v_{h}\right)= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{h}}, \tag{D.75}
\end{align*}
$$

for $w_{h} \in X$. This leads to the following linear algebraic system

$$
\begin{equation*}
A^{M} U^{M}=L \tag{D.76}
\end{equation*}
$$

with $U^{M}=U_{j}^{M}$ for $j \in\{1, \ldots, N\}$, such that $U_{\alpha, 0_{h}}^{M}=\sum_{j=1}^{N} U_{j}^{M} \psi_{j}(x, y)$,

$$
A^{M}=a^{M}\left(\psi_{i}, \psi_{j}\right),
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

## - Case 2A: Transmitting dielectric grating: TE case

Let $X \subset H_{\#}^{1}(\Omega)$ be a finite element subspace with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, . ., N$, be a basis of $X$. We want to find $U_{\alpha, 0_{h}}^{M} \in X$, for all $v_{h} \in X$ such that

$$
\begin{equation*}
a^{M}\left(U_{\alpha, 0_{h}}^{M}, v_{h}\right)=\left(f, v_{h}\right), \tag{D.77}
\end{equation*}
$$

where

$$
\begin{align*}
a^{M}\left(w_{h}, v_{h}\right)= & \int_{\Omega} \nabla w_{h} . \nabla \overline{v_{h}}-\int_{\Omega}\left(k^{2}-\alpha^{2}\right) w_{h} \overline{v_{h}}-2 i \alpha \int_{\Omega}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& -\int_{\Gamma_{ \pm}} T_{ \pm}^{\alpha, 0^{M}} w_{h} \overline{v_{h}} \\
\left(f, v_{h}\right)= & -\int_{\Gamma_{+}} 2 i \beta_{1}^{0} e^{-i \beta_{1}^{0} B} \overline{v_{h}}, \tag{D.78}
\end{align*}
$$

for $w_{h} \in X$. There exists $U^{M}=U_{j}^{M}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, 0_{h}}^{M}=$ $\sum_{j=1}^{N} U_{j}^{M} \psi_{j}(x, y)$. Hence, we get the following linear algebraic system

$$
\begin{equation*}
A^{M} U^{M}=L \tag{D.79}
\end{equation*}
$$

with

$$
A^{M}=a^{M}\left(\psi_{i}, \psi_{j}\right),
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

- Case 2B: Transmitting dielectric grating: TM case

Let $X \subset H_{\#}^{1}(\Omega)$ be a finite element subspace with $\operatorname{dim}(X)=N<\infty$ and let $\psi_{i}$ for $i=1, . ., N$, be a basis of $X$. We want to find $U_{\alpha, 0_{h}}^{M} \in X$, for all $v_{h} \in X$ such that

$$
\begin{equation*}
a^{M}\left(U_{\alpha, 0_{h}}, v_{h}\right)=\left(f, v_{h}\right), \tag{D.80}
\end{equation*}
$$

where

$$
\begin{align*}
a^{M}\left(w_{h}, v_{h}\right)= & \int_{\Omega} \frac{1}{k^{2}} \nabla w_{h} \nabla \overline{v_{h}}-\int_{\Omega} \frac{1-\alpha^{2}}{k^{2}} w_{h} \overline{v_{h}}-i \alpha \int_{\Omega} \frac{1}{k^{2}}\left(\partial_{x} w_{h}\right) \overline{v_{h}} \\
& +i \alpha \int_{\Omega} \frac{1}{k^{2}}\left(w_{h}\right) \overline{\partial_{x} v_{h}}-\int_{\Gamma_{ \pm}} \frac{1}{k^{2}} T_{ \pm}^{\alpha, 0^{M}} w_{h} \overline{v_{h}} \\
\left(f, v_{h}\right)= & -\int_{\Gamma_{+}} \frac{2 i \beta_{1}^{0}}{k_{1}^{2}} e^{-i \beta_{1}^{0} B} \overline{v_{h}}, \tag{D.81}
\end{align*}
$$

for $w_{h} \in X$. We are looking for $U^{M}=U_{j}^{M}$ for $j \in\{1, ., N\}$, such that $U_{\alpha, 0_{h}}^{M}=\sum_{j=1}^{N} U_{j}^{M} \psi_{j}(x, y)$. We get the following linear algebraic system

$$
\begin{equation*}
A^{M} U^{M}=L \tag{D.82}
\end{equation*}
$$

with

$$
A^{M}=a^{M}\left(\psi_{i}, \psi_{j}\right),
$$

and

$$
L=\left(f, \psi_{j}\right)
$$

for $\{i, j\} \in\{1, . ., N\}$.

## D. 3 A priori error estimates for the exact solution.

We derive an a priori error estimate for $U_{\alpha, 0}$ corresponding to Cases $1 \mathrm{~B}, 2 \mathrm{~A}$ and 2B in similar fashion to that of Case 1 A in Section 4.3. We again have the following three key results

- an estimate of the error from the discretisation of the continuous problem.
- an estimate of the error from truncating the DtN operator corresponding to the continuous problem.
- an estimate of the total error


## D.3.1 A priori error estimation of the discretized problem

In this section we will derive an upper bound on the error between the exact periodic solution $U_{\alpha, 0}$ and that found numerically by discretizing the problem $U_{\alpha, 0_{h}}$. In each case we will state the discretized periodic problem in its variational form, find a regularity bound for the $\alpha$-quasi periodic exact solution $U$ in terms of the norm in Definition 49, examine the discretisation error for $U$, and then use this to derive an a priori bound on the discretisation error for the periodic solution $U_{\alpha, 0}$.
Let $X \subset H_{\alpha \#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be a finite element subspace of order $p$ with $l \geq 1$, and let $\zeta_{h}$ be any regular partition of $X$ as described in Section B.3. We denote by $h$ the maximum mesh size after partitioning $\Omega \backslash \operatorname{int} \Omega_{3}$ using $\zeta_{h}$. We make the standard assumption given by equation (4.74) on the subspace $X$. Similarly, let $X^{\alpha}$ be a finite element subspace of order $p$ of $H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$.

- Case 1B: Perfectly conducting grating: TM case

The discretized problem corresponding to equation (D.25) is given below. Find $U_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{h}, \phi\right)=(f, \phi) \tag{D.83}
\end{equation*}
$$

with

$$
\begin{align*}
a\left(U_{h}, \phi\right) & =\left(\nabla U_{h}, \nabla \phi\right)_{\Omega \backslash i n t} \Omega_{3}-\left(k^{2} U_{h}, \phi\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm} U_{h}, \phi\right)_{\Gamma_{ \pm}},  \tag{D.84}\\
(f, \phi)_{\Gamma_{+}} & =-2 i \beta_{1}^{0} \int_{\Gamma_{+}} e^{-i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{\phi} . \tag{D.85}
\end{align*}
$$

for all $\phi \in X$, where $T_{ \pm}$is given by Definition 3 .
Lemma D-21. Let $U \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$, then for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ we have

$$
|a(U, v)| \leq C_{c}{\mathbf{k} U \mathrm{k}_{\mathcal{H}}} \mathrm{k}^{2} \mathbf{k}_{\mathcal{H}}
$$

such that $C_{c}=C d+1$ depends only on the grating period $d$.
Proof. The proof is similar to that of Lemma 57.
We also have the following
Lemma D-22. For $U \in H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$, we have

$$
|U|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2}-\Re(k) \mathbf{k} U \mathbf{k}_{L_{\alpha \# t}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \mathbf{k U} \mathrm{k}_{\mathcal{H}} \leq|a(U, U)|
$$

such that $a(u, v)$ is given by equation (D.26).
Proof. Since the only difference for Case 1A and Case 1B is the boundary condition on the interface, they both satisfy the same sesquilinear form $a(u, v)$. Hence, the proof of Lemma D-22 is similar to that of Lemma 58.

We also have the following result.
Lemma D-23. Let $U \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ be the solution of equation (D.28), and let $U_{h}$ be the corresponding discretized solution which satisfies equation (D.83). If we call $e_{h}=U-U_{h}$ then there exists a constant $C_{1}=C(C d+1) \mathbf{k} k \mathbf{k}_{\infty} \frac{h}{p} C_{\text {reg }}$, where $C_{\text {reg }}$ is the constant defined in Theorem C-3, such that

$$
\mathrm{k}_{\mathrm{e}_{h}} \mathrm{k}_{L_{\alpha \neq}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \leq C_{1} \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}} .
$$

Proof. Using the duality argument [22, p. 137], let $w$ be the solution of equation (4.86) then

$$
\begin{equation*}
\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}=\sup _{\phi \in C_{\infty}(\Omega)} \frac{\left|a\left(e_{h}, w-\psi\right)\right|}{\mathrm{k}_{\phi} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}} \tag{D.86}
\end{equation*}
$$

for all $\psi \in X$, by using Galerkin orthogonality (analogous to the derivation of equation (4.88)). So, from equation (D.84)

$$
\begin{aligned}
& \left|a\left(e_{h}, w-\psi\right)\right|=\left|\left(\nabla e_{h}, \nabla(w-\psi)\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} e_{h}, w-\psi\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm} e_{h}, w-\psi\right)_{\Gamma_{ \pm}}\right| \\
& \leq\left|e_{h}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash i n t \Omega_{3}\right)}|w-\psi|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}+\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{~K}_{h} \mathrm{k}_{L_{\alpha \# 1}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \mathrm{k} w-\psi \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \mathrm{int} \Omega_{3}\right)} \\
& +C \mathbf{k} e_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w-\psi \mathbf{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}
\end{aligned}
$$

using Cauchy's inequality [22, page. 50] and from equation (4.82). Hence,

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq(C d+1) \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w-\psi \mathbf{k}_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}
$$

from equation (4.84). We use the standard approximation estimate in finite element space given by equation (4.74) and we have

$$
\begin{equation*}
\left|a\left(e_{h}, w-\psi\right)\right| \leq C(C d+1) \frac{h}{p} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}} \mathbf{k} w \mathrm{k}_{H_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \tag{D.87}
\end{equation*}
$$

We use the result from Theorem C-3 and we have the regularity estimate

$$
\left.\mathrm{k} w \mathrm{k}_{H_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq\left(C_{r e g} \mathrm{k} k \mathrm{k}_{\infty}\right) \mathbf{k} \phi \mathbf{k}_{L_{\alpha \#}^{2}(\Omega \backslash i n t} \Omega_{3}\right)
$$

Using this in equation (D.87),

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq \frac{h}{p} C(C d+1) C_{r e g} \mathbf{k} k \mathbf{k}_{\infty} \mathrm{K}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} \phi \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)},
$$

and equation (D.86) gives

$$
\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \neq \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq C(C d+1) \mathbf{k} k \mathbf{k}_{\infty} \frac{h}{p} C_{r e g} \mathrm{k}_{h} \mathbf{k}_{\mathcal{H}}
$$

and we finish the proof by letting $C_{1}=C(C d+1) C_{r e g} \mathrm{k} k \mathrm{k}_{\infty} \frac{h}{p}$.
The previous three lemmas now allow us to derive the following a priori error estimate for the periodic solution $U_{\alpha, 0}$.

Theorem D-24. Let the wavenumber $|k| \geq k_{r e f}>0$, let the maximum mesh size $h \in\left[0, h_{0}\right]$, and let the order of the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$, and $C_{3}=1-\left(\Re(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}$ with $C_{1}$ as given in Lemma D-23. Let $U_{\alpha, 0}$ be the continuous solution of equation (D.55). Then $U_{\alpha, 0_{h}} \in X^{\alpha}$ the corresponding discretized solution exists and is unique. In addition, if $e_{\alpha, 0_{h}}=U_{\alpha, 0}-U_{\alpha, 0_{h}}$ then

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{C_{c}}{C_{3}} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}},
$$

and

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq 2 C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}}
$$

for all test functions $\psi_{\alpha, 0} \in X^{\alpha}$, where $C_{1}$ is given in Lemma D-23 and $C_{c}$ is defined in Lemma D-21.

Proof. We leave the proof because it is similar to that of Theorem 60.

For transmitting dielectric gratings, let $X \subset H_{\alpha \#}^{l}(\Omega)$ be a finite element subspace of order $p$ with $l \geq 1$, let $\zeta_{h}$ any regular partition of $X$ as described in Section B.3. Let $h$ be the maximum mesh size after partitioning $\Omega$ using $\zeta_{h}$. Once more, let us make the following standard assumption on the subspace $X$ as given by equation (4.39), [35]

$$
\begin{align*}
& \inf _{\psi \in X}\left\{\mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}+\frac{h}{p} \mathbf{k} \nabla v-\nabla \psi \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}+\left(\frac{h}{p}\right)^{\frac{1}{2}} \mathbf{k} v-\psi \mathbf{k}_{L_{\alpha \#}^{2}\left(\Gamma_{ \pm}\right)}\right. \\
&\left.+\frac{h}{p} \mathbf{k} v-\psi \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)}\right\} \leq C\left(\frac{h}{p}\right)^{l} \mathbf{k}^{l} \mathbf{k}_{H_{\alpha \#}^{l}(\Omega)} . \tag{D.88}
\end{align*}
$$

## - Case 2A: Transmitting dielectric grating: TE case

The discretized problem corresponding to equation (D.33) is given below. Find $U_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{h}, \phi\right)=(f, \phi) \tag{D.89}
\end{equation*}
$$

with

$$
\begin{align*}
a\left(U_{h}, \phi\right) & =\left(\nabla U_{h}, \nabla \phi\right)_{\Omega}-\left(k^{2} U_{h}, \phi\right)_{\Omega}-\left(T_{ \pm} U_{h}, \phi\right)_{\Gamma_{ \pm}},  \tag{D.90}\\
(f, \phi)_{\Gamma_{+}} & =-2 i \beta_{1}^{0} \int_{\Gamma_{+}} e^{i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{\phi} . \tag{D.91}
\end{align*}
$$

for all $\phi \in X$, where $T_{ \pm}$is given by Definition 3 .
Lemma D-25. Let $U \in H_{\alpha \#}^{1}(\Omega)$, then for all $v \in H_{\alpha \#}^{1}(\Omega)$ we have

$$
|a(U, v)| \leq C_{c} \mathbf{k} U \mathrm{k}_{\mathcal{H}}{\mathrm{k} v \mathbf{k}_{\mathcal{H}}}
$$

such that $C_{c}=C d+1$ depends only on the grating period $d$.
Proof. Similarly as in the proof of Lemma 57, we can use Cauchy- Schwarz inequality [22, p. 50] to get

$$
\begin{equation*}
\left|(\nabla U, \nabla v)_{\Omega}\right| \leq \mathrm{k} \nabla U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k} \nabla v \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}, \tag{D.92}
\end{equation*}
$$

and also that

$$
\begin{equation*}
\left|\left(k^{2} U, v\right)_{\Omega}\right| \leq \mathrm{k} k^{2} \mathbf{k}_{\infty} \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2} \Omega} . \tag{D.93}
\end{equation*}
$$

Hence, using together equations (D.92), (D.93), (4.83) we have

$$
|a(U, v)| \leq|U|_{H_{\alpha \#}^{1}(\Omega)}|v|_{H_{\alpha \#}^{1}(\Omega)}+\mathbf{k} k \mathbf{k}_{\infty}^{2} \mathbf{k} U \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)} \mathbf{k}^{2} \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}+C d \mathbf{k} U \mathbf{k}_{\mathcal{H}} \mathbf{k} v \mathbf{k}_{\mathcal{H}}
$$

and so

$$
\begin{equation*}
|a(U, v)| \leq C_{c} \mathbf{k} U \mathbf{k}_{\mathcal{H}} \mathbf{k}_{v} \mathbf{k}_{\mathcal{H}}, \tag{D.94}
\end{equation*}
$$

we let $C_{c}=(C d+1)$ to finish the proof.

We also have the following property.
Lemma D-26. For $U \in H_{\alpha \#}^{1}(\Omega)$, we have

$$
|U|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\Re(k) \mathbf{k} U \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)} \mathbf{k} U \mathbf{k}_{\mathcal{H}} \leq|a(U, U)|
$$

such that $a(u, v)$ is given by equation (D.31).
Proof. By taking the real part of $a(U, U)$, we can write

$$
\begin{aligned}
\mathfrak{R}(a(U, U)) & =\left(|U|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\Re\left(k^{2}\right){\left.\mathbf{k} U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}^{2}-\Re\left(T_{ \pm} U, U\right)\right),} \quad \geq|U|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\Re(k)^{2} \mathbf{k} U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}^{2}\right.
\end{aligned}
$$

using Lemma 8 . Since $(\Re(k))^{2} \geq \Re\left(k^{2}\right)$ and because $\mathfrak{R}(k){\mathbf{k} U \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)} \leq k \mathbf{k} U \mathbf{k}_{L_{\alpha \neq \#}^{2}(\Omega)} \leq, ~}$ $k U \mathrm{k}_{\mathcal{H}}$ then

$$
\mathfrak{R}(a(U, U)) \geq\left(|U|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\Re(k){\mathbf{k} U \mathbf{k}_{\mathcal{H}}} \mathbf{k} U \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}\right) .
$$

We finish the proof by noting that $\mathfrak{R}(a(U, U)) \leq|a(U, U)|$.
We also have the following result which gives the relationship between the norm in H and the norm in $L^{2}$ of the error estimate that we will use later to give an upper bound of the error estimate by discretizing the problem.

Lemma D-27. Let $U \in H_{\alpha \#}^{1}(\Omega)$ be the solution of equation (D.33) and $U_{h}$ be the solution of equation (D.89). If we call $e_{h}=U-U_{h}$ then there exists a constant $C_{1}=C(C d+1) \mathbf{k} k \mathbf{k}_{\infty} \frac{h}{p} C_{\text {stab }}$, where $C_{\text {stab }}=\left(1+C_{S} C\left(k_{0}, k_{3}\right)\right) C_{\text {reg }}$ is the constant defined in Lemma 31, such that

$$
\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq C_{1} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}
$$

Proof. We use the duality argument [22, p. 137], and let $w$ be the dual solution of

$$
\begin{align*}
\Delta w+k^{2} w & =\phi & & (x, y) \in \Omega  \tag{D.95}\\
T_{ \pm}^{*}-\partial_{n} w & =0 & & \text { on } \Gamma_{ \pm}
\end{align*}
$$

for all $\phi, w \in H_{\alpha \#}^{1}(\Omega)$ where $T_{ \pm}^{*}$ are the dual operators of $T_{ \pm},[61, \mathrm{p} .476]$. Then, similarly as in the proof of Lemma 59, we have for any $\psi \in X$

$$
\begin{equation*}
\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \neq \#}^{2}(\Omega)}=\sup _{\phi \in C_{\infty}(\Omega)} \frac{\left|a\left(e_{h}, w-\psi\right)\right|}{\mathrm{k} \phi \mathrm{k}_{L_{\alpha \neq}^{2}(\Omega)}} \tag{D.96}
\end{equation*}
$$

and

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq(C d+1) \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w-\psi \mathbf{k}_{H_{\alpha \# 1}^{1}(\Omega)} .
$$

We use the standard approximation estimate in finite element space given by equation (D.88) and so

$$
\begin{equation*}
\left|a\left(e_{h}, w-\psi\right)\right| \leq C(C d+1) \frac{h}{p} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}} \mathbf{k} w \mathrm{k}_{H_{\alpha \#}^{2}(\Omega)} \tag{D.97}
\end{equation*}
$$

We use the result from Theorem 33 and we have the regularity estimate

$$
\frac{\mathbf{k} w \mathbf{k}_{H_{\alpha \#}^{2}(\Omega)}}{\mathbf{k} \phi \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}} \leq\left(1+C_{S} C\left(k_{0}, k_{3}\right)\right) C_{r e g} \mathrm{k} k \mathbf{k}_{\infty}
$$

Using this in equation (D.97), we see

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq \frac{h}{p} C(C d+1) C_{r e g}\left(1+C_{S} C\left(k_{0}, k_{3}\right)\right) \mathbf{k} k \mathbf{k}_{\infty} \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} \phi \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)},
$$

and so equation (D.105) gives

$$
\mathrm{k}_{e_{h}} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq C(C d+1) \mathbf{k} k \mathbf{k}_{\infty} \frac{h}{p} C_{s t a b} \mathrm{ke}_{h} \mathbf{k}_{\mathcal{H}}
$$

with $C_{\text {stab }}=\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right) C_{\text {reg }}$. We finish the proof by letting

$$
C_{1}=C(C d+1) C_{s t a b} \mathrm{k} k \mathrm{k}_{\infty} \frac{h}{p}
$$

The previous three lemmas now allow us to derive the following a priori error estimate for the periodic solution $U_{\alpha, 0}$.

Theorem D-28. Let the wavenumber $|k| \geq k_{r e f}>0$, let the maximum mesh size $h \in\left[0, h_{0}\right]$, and let the order of the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$, and $C_{3}=1-\left(\Re(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}>0$ with $C_{1}$ as given in Lemma D-27. Let $U_{\alpha, 0}$ be the continuous solution of equation (4.47). Then $U_{\alpha, 0_{h}} \in X^{\alpha}$ the corresponding discretized solution exists and is unique. In addition, if $e_{\alpha, 0_{h}}=U_{\alpha, 0}-U_{\alpha, 0_{h}}$ then we have

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathbf{k}_{\mathcal{H}} \leq 4 \frac{C_{c}}{C_{3}} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}},
$$

and

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{0} \leq 2 C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}},
$$

for all test functions $\psi_{\alpha, 0} \in X^{\alpha}$, where $C_{1}$ as given in Lemma D-27 and $C_{c}$ as defined in Lemma D-25.

Proof. Let us denote $e_{h}=U-U_{h}$, and let $\psi=e^{i \alpha x} \psi_{\alpha, 0}$. We can use Lemma D-26 with Galerkin orthogonality (analogously to equation (4.88)) to get

$$
\left(\left|e_{h}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\Re(k) \mathrm{k}_{h_{h}} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}\right) \leq\left|a\left(e_{h}, U-\psi\right)\right| .
$$

From Lemma D-25 and since $\left|e_{h}\right|_{H_{\alpha \#}^{1}(\Omega)} \leq \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}}$ we have

$$
\left|e_{h}\right|_{H_{\alpha \#}^{1}(\Omega)}-\Re(k) \operatorname{ke}_{h} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq C_{c} \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}}
$$

and so

$$
\begin{aligned}
& \left|e_{h}\right|_{H_{\alpha \#}^{1}(\Omega)}+\mathrm{k} k \mathrm{k}_{\infty} \mathrm{ke}_{h} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \\
& \quad-\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) \mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq C_{c} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}} .
\end{aligned}
$$

Using Lemma D-27, we have

$$
\mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}}-\left(\mathcal{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1} \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}} \leq C_{c} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}} .
$$

Since $C_{3}=1-\left(\Re(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}>0$ then

$$
\begin{equation*}
\mathrm{k}_{e_{h}} \mathrm{k}_{\mathcal{H}} \leq \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}} . \tag{D.98}
\end{equation*}
$$

From Theorem 50 and equation (4.93), we have

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 2 \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}},
$$

then we use equation (D.98) to get

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 2 \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} .
$$

We use equation (4.95) and this becomes

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{C_{c}}{C_{3}} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} .
$$

We use once more Lemma D-27 with equation (D.98) and we get

$$
\mathrm{k} e_{h} \mathrm{k}_{L_{\alpha \neq 1}^{2}(\Omega)} \leq C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}} .
$$

From Theorem 50 and from equation (4.95), we have

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}(\Omega)} \leq 2 C_{1} \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}}
$$

The rest of the proof is similar to that of Theorem 60.

## - Case 2B: Transmitting dielectric grating: TM case

The discretized problem corresponding to equation (D.37) is given below. Find $U_{h} \in X$ such that

$$
\begin{equation*}
a\left(U_{h}, \phi\right)=(f, \phi)_{\Gamma_{+}} \tag{D.99}
\end{equation*}
$$

with

$$
\begin{align*}
a\left(U_{h}, \phi\right) & =\left(\frac{1}{k^{2}} \nabla U_{h}, \nabla \phi\right)_{\Omega}-\left(U_{h}, \phi\right)_{\Omega}-\left(\frac{1}{k^{2}} T_{ \pm} U_{h}, \phi\right)_{\Gamma_{ \pm}}  \tag{D.100}\\
(f, \phi)_{\Gamma_{+}} & =-\frac{2 i \beta_{1}^{0}}{k_{1}^{2}} \int_{\Gamma_{+}} e^{i\left(\alpha x-\beta_{1}^{0} B\right)} \bar{\phi}
\end{align*}
$$

for all $\phi \in X$, where $T_{ \pm}$is given by Definition 3 .
Lemma D-29. Let $U \in H_{\alpha \#}^{1}(\Omega)$, then for all $v \in H_{\alpha \#}^{1}(\Omega)$ we have

$$
|a(U, v)| \leq C_{c} \mathbf{k} U \mathrm{k}_{\mathcal{H}} \mathrm{k} \mathbf{k}_{\mathcal{H}}
$$

such that $C_{c}=\frac{1}{k_{r e f}^{2}}(C d+1)$ depends only on the period of the diffraction grating $d$ and a lower bound on the wavenumber $k_{r e f}$.

Proof. Similarly as in the proof of Lemma 57, we can use Cauchy- Schwarz inequality [22, p. 50] with the first term of equation (D.36) to get

$$
\begin{align*}
\left|\left(\frac{1}{k^{2}} \nabla U, \nabla v\right)_{\Omega}\right| & \leq \frac{1}{\mathrm{k} k^{2} \mathrm{k}_{\infty}} \mathrm{k} \nabla U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k} \nabla v \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}, \\
& \leq \frac{1}{k_{r e f}^{2}} \mathrm{k} \nabla U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k} \nabla v \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} . \tag{D.101}
\end{align*}
$$

We note for the second term in equation (D.36) that

$$
\begin{equation*}
\left|(U, v)_{\Omega}\right| \leq \mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k} v \mathrm{k}_{L_{\alpha \#}^{2} \Omega} . \tag{D.102}
\end{equation*}
$$

Hence, with the triangle inequality and putting equations (D.101), (D.102) and (4.83) in equation (D.36) we have

$$
|a(U, v)| \leq \frac{1}{k_{r e f}^{2}}|U|_{H_{\alpha \#}^{1}(\Omega)}|v|_{H_{\alpha \neq 1}^{1}(\Omega)}+{\mathrm{k} U \mathrm{k}_{L_{\alpha \#}^{2}}(\Omega)} \mathrm{k}^{2} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}+\frac{C d}{k_{r e f}^{2}} \mathrm{k} U \mathrm{k}_{\mathcal{H}} \mathrm{k} v \mathrm{k}_{\mathcal{H}} .
$$

Since $\frac{\left\|k^{2}\right\|_{\infty}}{k_{\text {ref }}^{2}} \geq 1$ then
$|a(U, v)| \leq \frac{1}{k_{r e f}^{2}}\left(|U|_{H_{\alpha \neq 1}^{1}(\Omega)}|v|_{H_{\alpha \#}^{1}(\Omega)}+\mathrm{k}^{2} \mathbf{k}_{\infty} \mathbf{k} U \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathrm{k}^{2} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}+C d \mathbf{k} U \mathrm{k}_{\mathcal{H}} \mathrm{k}^{2} \mathbf{k}_{\mathcal{H}}\right)$ and so by using equation (4.84)

$$
\begin{equation*}
|a(U, v)| \leq C_{c} \mathbf{k} U \mathbf{k}_{\mathcal{H}} \mathbf{k}_{u} \mathbf{k}_{\mathcal{H}}, \tag{D.103}
\end{equation*}
$$

where $C_{c}=\frac{1}{k_{r e f}^{2}}(C d+1)$ to finish the proof.

We also have the following property.
Lemma D-30. For $U \in H_{\alpha \#}^{1}(\Omega)$, we have

$$
\frac{1}{\mathrm{k} \Re\left(k^{2}\right) \mathrm{k}_{\infty}}|U|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathrm{k}^{2} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}^{2} \leq|a(U, U)|+\left|\left(\frac{1}{k^{2}} T_{ \pm} U, U\right)_{\Gamma_{ \pm}}\right|
$$

such that $a(u, v)$ is given by equation (D.36).
Proof. From equation (D.36), we have

$$
\left(\frac{1}{k^{2}} \nabla U, \nabla U\right)_{\Omega}-(U, U)_{\Omega}=a(U, U)+\left(\frac{1}{k^{2}} T_{ \pm} U, U\right)_{\Gamma_{ \pm}}
$$

Since $|\Re(c)| \leq|c|$ for any $c \in \mathbb{C}$ then

$$
\left|\Re\left(\frac{1}{k^{2}} \nabla U, \nabla U\right)_{\Omega}-(U, U)_{\Omega}\right| \leq\left|a(U, U)+\left(\frac{1}{k^{2}} T_{ \pm} U, U\right)_{\Gamma_{ \pm}}\right| .
$$

By noting that $|b-c| \geq|b|-|c|$ and with the triangle inequality we get

$$
\left|\Re\left(\frac{1}{k^{2}} \nabla U, \nabla U\right)_{\Omega}\right|-(U, U)_{\Omega} \leq|a(U, U)|+\left|\left(\frac{1}{k^{2}} T_{ \pm} U, U\right)_{\Gamma_{ \pm}}\right| .
$$

The proof is finished by noting that

$$
\left|\Re\left(\frac{1}{k^{2}} \nabla U, \nabla U\right)_{\Omega}\right| \geq \frac{1}{\mathrm{k} \Re\left(k^{2}\right) \mathrm{k}_{\infty}}|U|_{H_{\alpha \#}^{1}(\Omega)} .
$$

The following result shows the relationship between the norm in H and the norm in $L^{2}$ of the error estimate.

Lemma D-31. For $U \in H_{\alpha \#}^{1}(\Omega)$ be the solution of equation (D.37), and let $U_{h}$ the corresponding discretized solution of equation (D.99). If we denote $e_{h}=$ $U-U_{h}$, then there exists a constant $C_{1}=C C_{\text {stab }} \frac{h}{p} \frac{\|k\|_{\infty}}{k_{r e f}^{2}}(C d+1)$, where $C_{\text {stab }}=$ $\left(1+C_{S} C\left(k_{0}, k_{3}\right)\right) C_{\text {reg }}$ is the constant defined in Theorem C-6, such that

$$
\mathrm{k}_{\mathrm{e}_{h}} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq C_{1} \mathbf{k} U \mathrm{k}_{\mathcal{H}}
$$

Proof. We use the duality argument [22, p. 137], and let $w$ be the dual solution of

$$
\begin{align*}
\nabla \cdot\left(\frac{1}{k^{2}} \nabla w\right)+w & =\phi & & (x, y) \in \Omega,  \tag{D.104}\\
\left(T_{ \pm}^{*}-\partial_{n} w\right) & =0 & & \text { on } \Gamma_{ \pm},
\end{align*}
$$

for all $\phi, w \in H_{\alpha \#}^{1}(\Omega)$ where $T_{ \pm}^{*}$ are the dual operators of $T_{ \pm}[61, \mathrm{p} .476]$. Then, similarly as in the proof of Lemma 59, we have

$$
\begin{equation*}
\mathrm{k}_{e_{h}} \mathrm{k}_{L_{\alpha \neq}^{2}(\Omega)}=\sup _{\phi \in C_{\infty}(\Omega)} \frac{\left|a\left(e_{h}, w-\psi\right)\right|}{\mathrm{k} \phi \mathrm{k}_{L_{\alpha \neq \#}^{2}(\Omega)}^{2}} \tag{D.105}
\end{equation*}
$$

such that $\psi \in X$ and so from equation (D.100)

$$
\begin{aligned}
& \left|a\left(e_{h}, w-\psi\right)\right|=\left|\left(\frac{1}{k^{2}} \nabla e_{h}, \nabla(w-\psi)\right)_{\Omega}-\left(e_{h}, w-\psi\right)_{\Omega}-\left(\frac{1}{k^{2}} T_{ \pm} e_{h}, w-\psi\right)_{\Gamma_{ \pm}}\right| \\
& \leq\left|\left(\frac{1}{k^{2}} \nabla e_{h}, \nabla(w-\psi)\right)_{\Omega}\right|+\left|\left(e_{h}, w-\psi\right)_{\Omega}\right|+\left|\left(\frac{1}{k^{2}} T_{ \pm} e_{h}, w-\psi\right)_{\Gamma_{ \pm}}\right| \\
& \leq \frac{1}{k_{r e f}^{2}}\left(\left|e_{h}\right|_{H_{\alpha \neq 1}^{1}(\Omega)}|w-\psi|_{H_{\alpha \# \#}^{1}(\Omega)}+\mathbf{k} k \mathbf{k}_{\infty}^{2} \mathbf{k}_{h} \mathbf{k}_{L_{\alpha \# \#}^{2}(\Omega)} \mathbf{k} w-\psi \mathbf{k}_{L_{\alpha \neq \#}^{2}(\Omega)}\right. \\
& \left.+C d \mathbf{k} e_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w-\psi \mathbf{k}_{H_{\alpha \#( }^{1}(\Omega)}\right) .
\end{aligned}
$$

using equation (4.82) with Cauchy's inequality [22, p. 50]. Hence, using equation (4.84) we get

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq(C d+1) \frac{1}{k_{r e f}^{2}} \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w-\psi \mathbf{k}_{H_{\alpha \#}^{1}(\Omega)}
$$

We use the standard approximation estimate in finite element space given by equation (D.88) and so

$$
\begin{equation*}
\left|a\left(e_{h}, w-\psi\right)\right| \leq C(C d+1) \frac{1}{k_{r e f}^{2}} \frac{h}{p} \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} w \mathbf{k}_{H_{\alpha \#}^{2}(\Omega)} \tag{D.106}
\end{equation*}
$$

We use the result from Theorem C-6 and we have the regularity estimate

$$
\frac{\mathrm{k} w \mathrm{k}_{H_{\alpha \#}^{2}(\Omega)}}{\mathrm{k} \phi \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}} \leq\left(1+C_{S} C\left(k_{0}, k_{3}\right)\right) C_{r e g} \mathrm{k} k \mathrm{k}_{\infty} .
$$

Hence

$$
\left|a\left(e_{h}, w-\psi\right)\right| \leq C \mathbf{k} k \mathbf{k}_{\infty} \frac{h}{p k_{r e f}^{2}} C_{r e g}(C d+1)\left(1+C_{S} C\left(k_{0}, k_{3}\right)\right) \mathbf{k}_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} \phi \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}
$$

and so

$$
\mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq C C_{s t a b} \frac{h \mathbf{k} k \mathbf{k}_{\infty}}{p k_{\text {ref }}^{2}}(C d+1) \mathrm{k}_{h} \mathbf{k}_{\mathcal{H}}
$$

with $C_{s t a b}=\left(1+C_{s} C\left(k_{0}, k_{3}\right)\right) C_{r e g}$. We finish the proof by letting

$$
C_{1}=C C_{s t a b} \frac{h \mathbf{k} k \mathbf{k}_{\infty}}{p k_{r e f}^{2}}(C d+1) .
$$

We have the following result.
Theorem D-32. Let the wavenumber $|k| \geq k_{\text {ref }}>0$, let the maximum mesh size $h \in\left[0, h_{0}\right]$, and let the order of the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$, and $c_{4}=1-2 \mathbf{k} k \mathbf{k}_{\infty} C_{1}>0$ with $C_{1}$ as given in Lemma D-31. Let $U_{\alpha, 0}$ be the continuous solution of equation (D.57). Then $U_{\alpha, 0_{h}} \in X^{\alpha}$ the corresponding discretized solution exists and is unique. If $e_{\alpha, 0_{h}}=U_{\alpha, 0}-U_{\alpha, 0_{h}}$ then we have

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 4 c_{k}(2 C d+1) / c_{4} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}},
$$

and

$$
\mathbf{k} e_{\alpha, 0_{h}} \mathrm{k}_{0} \leq 2 c_{k} C_{1}(2 C d+1) / c_{4} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}},
$$

for all test functions $\psi_{\alpha, 0} \in X^{\alpha}$, where $C \in \mathbb{R}$, $d$ the period of the grating and $c_{k}=\mathrm{k} \Re\left(k^{2}\right) \mathrm{k}_{\infty} / k_{\text {ref }}^{2}$.

Proof. With $e_{h}=U-U_{h}$, and let $\psi=e^{i \alpha x} \psi_{\alpha, 0}$. We can use Lemma D-30 with Galerkin orthogonality similar to equation (4.88) to get

From Lemma D-29 and equation (4.83) and since $\mathbf{k} k \boldsymbol{k}_{\infty}^{2} \geq \mathbf{k} \mathfrak{R}\left(k^{2}\right) \mathrm{k}_{\infty}$, we see that

$$
\begin{align*}
& \frac{1}{\mathrm{k} \mathrm{\Re R}\left(k^{2}\right) \mathrm{k}_{\infty}}\left(\left|e_{h}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k}_{h} \mathrm{k}_{L_{\alpha \#}^{2}}^{2}(\Omega)\right. \\
& \leq C_{c} \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}}+\frac{C d}{k_{r e f}^{2}} \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} \tag{D.107}
\end{align*}
$$

using Céa's theorem [22, p. 64]. Since $C_{c}=\frac{1}{k_{r e f}^{2}}(C d+1)$ as given in the proof of Lemma D-29, we have

$$
\left.\begin{array}{l}
\frac{1}{\mathrm{k} \mathrm{\Re}\left(k^{2}\right) \mathrm{k}_{\infty}}\left(\left|e_{h}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k}_{h_{h}} \mathrm{k}_{L_{\alpha \#}^{2}}^{2}(\Omega)\right.
\end{array}\right)
$$

By letting $c_{k}=\mathbf{k R}\left(k^{2}\right) \mathrm{k}_{\infty} / k_{\text {ref }}^{2}$ and noting that $\mathrm{k} k \mathrm{k}_{\infty} \mathrm{ke}_{h} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq \mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}$, we get

$$
\begin{aligned}
& \left|e_{h}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathbf{k} k \mathbf{k}_{\infty} \mathrm{k}_{e_{h}} \mathbf{k}_{L_{\alpha \neq}^{2}(\Omega)} \mathrm{k}_{e_{h}} \mathbf{k}_{\mathcal{H}} \\
& \quad \leq c_{k}(2 C d+1) \mathbf{k} e_{h} \mathbf{k}_{\mathcal{H}} \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}} .
\end{aligned}
$$

We can use Definition 49 to substitute for the first term on the left hand side and divide by $\mathrm{k} e_{h} \mathrm{~K}_{\mathcal{H}}$

$$
\mathbf{k} e_{h} \mathbf{k}_{\mathcal{H}}-2 \mathbf{k} k \mathbf{k}_{\infty} \mathbf{k}_{e_{h}} \mathbf{k}_{L_{\alpha \neq}^{2}(\Omega)} \leq c_{k}(2 C d+1) \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}}
$$

Using Lemma D-31, we have

$$
\begin{equation*}
\mathrm{k}_{h} \mathrm{k}_{\mathcal{H}}-2 \mathbf{k} k \mathrm{k}_{\infty} C_{1} \mathrm{k}_{e_{h}} \mathrm{k}_{\mathcal{H}} \leq c_{k}(2 C d+1) \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}} . \tag{D.108}
\end{equation*}
$$

We suppose that $2 \mathbf{k} k \mathbf{k}_{\infty} C_{1}<1$ and so $c_{4}=1-2 \mathbf{k} k \mathrm{k}_{\infty} C_{1}>0$

$$
\begin{equation*}
\mathrm{k}_{e_{h}} \mathrm{k}_{\mathcal{H}} \leq \frac{c_{k}}{c_{4}}(2 C d+1) \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}} \tag{D.109}
\end{equation*}
$$

From Theorem 50 and equation (4.93), we have

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 2 \mathrm{k} e_{h} \mathrm{k}_{\mathcal{H}},
$$

then we use equation (D.109) to get

$$
\mathbf{k} e_{\alpha, 0_{h}} \mathbf{k}_{\mathcal{H}} \leq 2 c_{k}(2 C d+1) / c_{4} \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}} .
$$

We use equation (4.95) to get the first result

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{\mathcal{H}} \leq 4 c_{k}(2 C d+1) / c_{4} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} .
$$

To show the second result, we use once more Lemma D-27 with equation (D.109) and we get

$$
\mathbf{k}_{h} \mathbf{k}_{L_{\alpha \neq 1}^{2}(\Omega)} \leq c_{k} C_{1}(2 C d+1) \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}} .
$$

From Theorem 50 and from equation (4.95), we have

$$
\mathrm{k} e_{\alpha, 0_{h}} \mathrm{k}_{L_{\#}^{2}(\Omega)} \leq 2 c_{k} C_{1}(2 C d+1) \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} .
$$

The proof of the uniqueness of $U_{\alpha, 0_{h}}$ is similar to that of Theorem 60 .

## D.3.2 A priori error estimation of the continuous problem by truncating the DtN operator

As explained in Section 4.3.2, the parameter $b$ is introduced for computational efficiency, to derive an a priori error estimate, and to allows us to cope with more general problems involving layered geometry such as cladding or substrate. We have the following results for Cases 1B, 2A and 2B.

- Case 1B: Perfectly conducting grating: TM case

Let $U_{\alpha, 0}^{M}$ be the approximate solution of the continuous problem given by equation (D.55) when we truncate the DtN map by $T_{ \pm}^{\alpha, 0^{M}}$ with $M \in \mathbb{N}$, as given by equation (4.98). The error estimate from truncating the $\operatorname{DtN}$ operator is given in the following theorem.

Theorem D-33. Let us choose $M \in \mathbb{N}$ such that $M>M_{0}=|k|+|\alpha|$ and let us denote by

$$
e_{\alpha, 0}^{M}=U_{\alpha, 0}-U_{\alpha, 0}^{M} .
$$

If $\left(\mathfrak{R}(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}<1$, such that $C_{1}$ is as given in Lemma D-23, and so $C_{3}=$ $1-\left(\Re(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}>0$, then $U_{\alpha, 0}^{M}$ exists and is unique and we have

$$
\begin{align*}
\mathrm{k} e_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}} & \leq 4 \frac{d}{C_{3}} e^{-(B-b) \sin \left(z_{n} / 2\right)} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}  \tag{D.110}\\
\mathbf{k} & U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)},  \tag{D.111}\\
\mathrm{K} e_{\alpha, 0}^{M} \mathrm{k}_{L_{\#}^{2}(\Omega)} & \leq 2 \frac{d}{C_{3}} C_{1} e^{-(B-b) \sin \left(z_{n} / 2\right)} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
\mathrm{k} & U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}\left(\Gamma_{1, \pm}\right)}},
\end{align*}
$$

where $\Gamma_{1, \pm}=\{(x, \pm b) \in \Omega\}$ and $z_{n}$ is given by equation (2.44) with $b$ as shown in Figure 2.3.

Proof. Let $U \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ satisfy $a(U, v)=(f, v)_{\Gamma_{+}}$for all $v \in H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)$ as given by equation (D.28). We proceed similarly as in the proof of Theorem 60, so we are approximating the continuous problem by finding $U^{M} \in H_{\alpha \#}^{1}\left(\Omega \backslash\right.$ int $\left.\Omega_{3}\right)$, such that

$$
\begin{equation*}
a^{M}\left(U^{M}, v\right)=(f, v)_{\Gamma_{+}} \tag{D.112}
\end{equation*}
$$

with

$$
\begin{align*}
a^{M}\left(U^{M}, v\right) & =\left(\nabla U^{M}, \nabla v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} U^{M}, v\right)_{\Omega \backslash \mathrm{int} \Omega_{3}}-\left(T_{ \pm}^{M} U^{M}, v\right)_{\Gamma_{ \pm}}  \tag{D.113}\\
(f, v)_{\Gamma_{+}} & =\left(-2 i \beta_{1}^{0} e^{i\left(\alpha x-\beta_{1}^{0} B\right)}, v\right)_{\Gamma_{+}},
\end{align*}
$$

for all $v \in \Omega \backslash$ int $\Omega_{3}$ such that $T_{ \pm}^{M}(v)$ is given by equation (4.103). From equations (D.28) and (D.112) we have

$$
\begin{equation*}
a(U, v)-a^{M}\left(U^{M}, v\right)=0 \tag{D.114}
\end{equation*}
$$

We let $e^{M}=U-U^{M}$ and since $T_{ \pm}=T_{ \pm}^{M}+\left(T_{ \pm}-T_{ \pm}^{M}\right)$ then

$$
\begin{equation*}
\left(\nabla e^{M}, \nabla v\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} e^{M}, v\right)_{\Omega \backslash i \operatorname{int} \Omega_{3}}-\left(T_{ \pm}{ }^{M} e^{M}, v\right)_{\Gamma_{ \pm}}=\left(\left(T_{ \pm}-T_{ \pm}{ }^{M}\right) U, v\right)_{\Gamma_{ \pm}} . \tag{D.115}
\end{equation*}
$$

Similar to the proof of Theorem 61, we let $M \in \mathbb{N}$ such that for $M>M_{0}$ we have $M>|k|+|\alpha|$ then $n_{\alpha}^{2}>k^{2}$ for $\left|\frac{2 \pi n}{d}\right| \geq M$ and we can show that
such that $\Gamma_{1, \pm}=\{(x, \pm b) \in \Omega\}$ and $c_{\text {min }}=\inf |n|>\frac{M d}{2 \pi} \sin \left(z_{n} / 2\right)$.
We note again as in the proof of Theorem 61 that the left hand side of equation (D.115) satisfies

$$
\begin{gathered}
\left|\left(\nabla e^{M}, e^{M}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(k^{2} e^{M}, e^{M}\right)_{\Omega \backslash \operatorname{int} \Omega_{3}}-\left(T_{ \pm}{ }^{M} e^{M}, e^{M}\right)_{\Gamma_{ \pm}}\right|, \\
\quad \geq \mathfrak{R}\left(\left|e^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-k^{2} \mathbf{k} e^{M} \mathbf{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}\right),
\end{gathered}
$$

since $-\mathfrak{R}\left(T_{ \pm}{ }^{M} e^{M}, e^{M}\right)_{\Gamma_{ \pm}} \geq 0$ from equation (2.72). So,

$$
\left|e^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2}-\Re\left(k^{2}\right) \mathbf{k}^{M} \mathbf{k}_{L_{\alpha \neq 1}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \leq\left|a^{M}\left(e^{M}, e^{M}\right)\right| .
$$

Hence, using equation (D.115), we get

$$
\begin{aligned}
& \left|e^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}^{2}-\Re\left(k^{2}\right) \mathbf{k e}^{M} \mathrm{k}_{L_{\alpha \#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}^{2} \\
& \leq d e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathbf{k}^{-\left(\mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}\left(\Gamma_{1, \pm}\right)}} \mathrm{Ke}^{M} \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} .\right.} .
\end{aligned}
$$

Again we use the duality argument [22, p. 137] to approximate k. $_{L_{L_{\alpha \#}^{2}}(\Omega)}$, with the dual problem given by equation (4.86). Therefore, we can use Lemma D-23 and equation (4.91) to get

$$
\left|e^{M}\right|_{H_{\alpha \# t}^{1}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)}-\Re(k) C_{1} \mathrm{ke}^{M} \mathrm{k}_{\mathcal{H}} \leq d e^{-c_{\min (B-b)} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathbf{k U} \mathrm{k}_{H_{\alpha \# t}^{\frac{1}{\alpha}}\left(\Gamma_{1, \pm}\right)} .
$$

Since $\left.\mathbf{k} e^{M} \mathbf{k}_{\mathcal{H}} \geq \mathbf{k} e^{M} \mathbf{k}_{L_{\alpha \#}^{2}(\Omega \backslash i n t} \Omega_{3}\right)$ then

$$
\begin{gathered}
\left|e^{M}\right|_{H_{\alpha \#}^{1}\left(\Omega \backslash \text { int } \Omega_{3}\right)}+k \mathbf{k} e^{M} \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}-\left(\mathfrak{R}(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1} \mathbf{k}^{M} \mathbf{k}_{\mathcal{H}} \\
\leq d e^{-c_{m i n}(B-b)} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
\mathbf{k} U \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)}
\end{gathered}
$$

Since we have $C_{3}=1-\mathfrak{R}(k)+\mathbf{k} k \mathbf{k}_{\infty} C_{1}>0$, we can use equations (4.93) and (4.95) together with Theorem 50 and Definition 49 to get

$$
\begin{equation*}
\mathbf{k} e_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{d}{C_{3}} e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathbf{k} U_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)} . \tag{D.117}
\end{equation*}
$$

From Lemma 59 and equations (4.110), (4.93) and (4.95) together with Theorem 50 , we conclude that

$$
\begin{equation*}
\mathrm{ke}_{\alpha, 0}^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq 2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{\min }} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)} . \tag{D.118}
\end{equation*}
$$

The proof for the uniqueness is similar to that in Theorem 61.

## - Case 2A: Transmitting dielectric grating: TE case

Let $U_{\alpha, 0}^{M}$ be the approximate solution of the continuous problem given by equation (D.56) when we truncate the DtN map by $T_{ \pm}^{\alpha, 0^{M}}$, with $M \in \mathbb{N}$, as given by equation (4.98). The error estimate by truncating $T_{ \pm}^{\alpha, 0}$ is given by the following theorem.

Theorem D-34. Let us choose $M \in \mathbb{N}$ such that $M>M_{0}=|k|+|\alpha|$ and let us denote by

$$
e_{\alpha, 0}^{M}=U_{\alpha, 0}-U_{\alpha, 0}^{M} .
$$

If $\left(\mathfrak{R}(k)+\mathbf{k} k \mathrm{k}_{\infty}\right) C_{1}<1$, such that $C_{1}$ as given in Lemma D-27, so that $C_{3}=$ $1-\left(\Re(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}>0$ then $U_{\alpha, 0}^{M}$ exists and is unique and we have

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}} & \leq 4 \frac{d}{C_{3}} e^{-(B-b) c_{\text {min }}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)}, \\
\mathrm{K} e_{\alpha, 0}^{M} \mathrm{k}_{L_{\#}^{2}(\Omega)} & \leq 2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{\text {min }}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}
\end{aligned} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}\left(\Gamma_{1, \pm}\right)}} .
$$

The constant $c_{\text {min }}=\left.\left.\inf \right|_{n}\right|_{>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$ with $\Gamma_{1, \pm}=\{(x, \pm b) \in \Omega\}$ and $z_{n}$ is given by equation (2.44) where $b$ is described in Figure 2.3

Proof. The proof is similar to that of Theorem 60 and Theorem D-33. We just need $U \in H_{\alpha \#}^{1}(\Omega)$ to satisfy $a(U, v)=(f, v)_{\Gamma_{+}}$, for all $v \in H_{\alpha \#}^{1}(\Omega)$, as given by equation (D.33).

- Case 2B: Transmitting dielectric grating: TM case

Let $U_{\alpha, 0}^{M}$ be the approximate solution of the continuous problem given by equation (D.57) when truncate the $\operatorname{DtN}$ map by $T_{ \pm}^{\alpha, 0^{M}}$, with $M \in \mathbb{N}$, as given by equation (4.98). We have the following theorem.

Theorem D-35. Let us choose $M \in \mathbb{N}$ such that $M>M_{0}=|k|+|\alpha|$ and $2 \pi|n| / d>M$. Let $\mathbf{k} k \mathbf{k}_{\infty} \geq k_{\text {ref }}$ and let us denote by

$$
e_{\alpha, 0}^{M}=U_{\alpha, 0}-U_{\alpha, 0}^{M} .
$$

If $2 \mathbf{k} k \mathbf{k}_{\infty} C_{1}<1$, such that $C_{1}$ as given in Lemma D-31, $C \leq{ }^{\sqrt{ }} \overline{5}$ as defined in Lemma 7 then $C_{4}=1-2 \mathrm{k} k \mathrm{k}_{\infty} C_{1}>0$ and $U_{\alpha, 0}^{M}$ exists and is unique. In addition, we have

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}} & \leq 4 c_{k} \frac{d}{C_{4}}\left(C \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}}+e^{-(B-b) c_{\min }} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)}\right), \\
\mathrm{k} e_{\alpha, 0}^{M} \mathrm{k}_{L_{\#}^{2}(\Omega)} & \leq 2 c_{k} d \frac{C_{1}}{C_{4}}\left(C \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}}+e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right.}\right),
\end{aligned}
$$

for all test functions $\psi_{\alpha, 0} \in X^{\alpha}$ with $c_{k}=\frac{\left\|\Re\left(k^{2}\right)\right\|_{\infty}}{k_{r e f}^{2}}, c_{\text {min }}=\inf _{|n|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$ and $\Gamma_{1, \pm}=\{(x, \pm b) \in \Omega\}$ and $z_{n}$ is given by equation (2.44).
Proof. Let $U \in H_{\alpha \#}^{1}(\Omega)$ satisfy $a(U, v)=(f, v)_{\Gamma_{+}}$, for all $v \in H_{\alpha \#}^{1}(\Omega)$, as given by equation (D.37). We proceed similarly as in the proof of Theorem D-32, so we are approximating the continuous problem by finding $U^{M} \in H_{\alpha \#}^{1}(\Omega)$, such that

$$
\begin{equation*}
a^{M}\left(U^{M}, v\right)=(f, v)_{\Gamma_{+}} \tag{D.119}
\end{equation*}
$$

with

$$
\begin{align*}
a^{M}\left(U^{M}, v\right) & =\left(\frac{1}{k^{2}} \nabla U^{M}, \nabla v\right)_{\Omega}-\left(U^{M}, v\right)_{\Omega}-\left(\frac{1}{k^{2}} T_{ \pm}^{M} U^{M}, v\right)_{\Gamma_{ \pm}}  \tag{D.120}\\
(f, v)_{\Gamma_{+}} & =\left(-\frac{2 i \beta_{1}^{0}}{k_{1}^{2}} e^{i\left(\alpha x-\beta_{1}^{0} B\right), v}\right)_{\Gamma_{+}}
\end{align*}
$$

for all $v \in \Omega$ such that $T_{ \pm}^{M}(v)$ is given by equation (4.103). We note again from equations (D.37) and (D.119) that

$$
\begin{equation*}
a(U, v)-a^{M}\left(U^{M}, v\right)=0 \tag{D.121}
\end{equation*}
$$

By letting $e^{M}=U-U^{M}$ and noting that $T_{ \pm}=T_{ \pm}^{M}+\left(T_{ \pm}-T_{ \pm}^{M}\right)$, we get from equation (D.121)

$$
\begin{align*}
\left(\frac{1}{k^{2}} \nabla e^{M}, \nabla v\right)_{\Omega}-\left(e^{M}, v\right)_{\Omega}-\left(\frac{1}{k^{2}} T_{ \pm}{ }^{M} e^{M}, v\right)_{\Gamma_{ \pm}} & =\left(\frac{1}{k^{2}}\left(T_{ \pm}-T_{ \pm}{ }^{M}\right) U, v\right)_{\Gamma_{ \pm}} \\
a^{M}\left(e^{M}, v\right) & =\left(\frac{1}{k^{2}}\left(T_{ \pm}-T_{ \pm}^{M}\right) U, v\right)_{\Gamma_{ \pm}} .(\mathrm{D} . \tag{D.122}
\end{align*}
$$

Similarly as in the proof of Theorem 61, we let $M$ such that $M>k+|\alpha|$ so that $n_{\alpha}^{2}>k^{2}$ for $|n| \geq \frac{M d}{2 \pi}$ and we can show that
with $c_{\text {min }}=\inf _{\left|{ }_{n}\right|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$. Then, we note that $\mathfrak{R}\left(a^{M}\left(e^{M}, v\right)\right) \leq\left|a^{M}\left(e^{M}, v\right)\right|$. Also, we note that truncating $T_{ \pm}$does not affect the validity of Lemma D-30 and so

$$
\begin{aligned}
\frac{1}{\mathrm{k} \mathcal{R}\left(k^{2}\right) \mathrm{k}_{\infty}}\left(\left|e^{M}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} e^{M}{\mathrm{k}_{L_{\alpha \neq}^{2}(\Omega)}^{2}(\Omega)}^{2}\right. & \leq\left|\left(\frac{1}{k^{2}} T_{ \pm}^{M} e^{M}, e^{M}\right)_{\Gamma_{ \pm}}\right|+\left|a^{M}\left(e^{M}, e^{M}\right)\right| \\
& \leq\left|\left(\frac{1}{k^{2}} T_{ \pm}^{M} e^{M}, e^{M}\right)_{\Gamma_{ \pm}}\right|+\left|a^{M}\left(e^{M}, U-\psi\right)\right|
\end{aligned}
$$

since we have Galerkin orthogonality similar to equation (4.88), where
$\psi=e^{i \alpha x} \psi_{\alpha, 0}$ with $\psi_{\alpha, 0} \in X^{\alpha}$. By noting that $\left|\left(T_{ \pm}^{M} v, v\right)_{\Gamma_{ \pm}}\right| \leq \mid\left(T_{ \pm} v, v\right)_{\Gamma_{ \pm}}$, we can use equations (4.83), (4.104) and similar to (4.109) we get

$$
\begin{aligned}
& \frac{1}{\mathrm{k} \Re\left(k^{2}\right) \mathrm{k}_{\infty}}\left(\left|e^{M}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathrm{k} k \mathrm{k}_{\infty}^{2} \mathrm{k} e^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}^{2}\right) \text {, } \\
& \leq \frac{C d}{k_{r e f}^{2}} \mathrm{k} e^{M} \mathbf{k}_{\mathcal{H}} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}} \\
& +d e^{-(B-b) \sin \left(z_{n} / 2\right) \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}}{\mathrm{~K} U \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)} \mathrm{Ke}^{M} \mathrm{~K}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{ \pm}\right)} \mathrm{N}^{2}} \\
& \leq \frac{C d}{k_{r e f}^{2}} \mathrm{k} e^{M} \mathrm{k}_{\mathcal{H}} \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} \\
& +\frac{d}{k_{r e f}^{2}} e^{-(B-b) \sin \left(z_{n} / 2\right) \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathbf{k} U \mathbf{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)} \mathrm{Ke}^{M} \mathbf{k}_{\mathcal{H}}
\end{aligned}
$$

since $U_{h}^{M}$ minimizes $a^{M}$, we can use Céa's Lemma [22, p. 64] and Theorem A-13 with the equivalence of the norms in H and in $H_{\alpha \#}^{1}(\Omega)$. Once more, we use the duality argument $[22, \mathrm{p} .137]$ to approximate $\mathrm{k} . \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)}$, with the dual problem given by equation (4.86). Therefore, we can use Lemma D-31 and similar to equation (D.108), we get

$$
\begin{gathered}
\frac{1}{\left\|\Re\left(k^{2}\right)\right\|_{\infty}}\left(\mathbf{k} e^{M} \mathbf{k}_{\mathcal{H}}-2 \mathbf{k} k \mathbf{k}_{\infty} C_{1} \mathbf{k} e^{M} \mathbf{k}_{\mathcal{H}}\right) \leq \frac{C d}{k_{r e f}^{2}} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}} \\
\quad+\frac{d}{k_{r e f}^{2}} e^{-(B-b) \sin \left(z_{n} / 2\right) \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathbf{k} U \mathbf{k}_{H_{\alpha \#}^{\frac{1}{\alpha}}\left(\Gamma_{1, \pm}\right)}
\end{gathered}
$$

Letting $c_{k}=\frac{\left\|\Re\left(k^{2}\right)\right\|_{\infty}}{k_{r e f}^{2}}$ and supposing $2 \mathbf{k} k \mathbf{k}_{\infty} C_{1}<1$, we have $C_{4}=1-2 \mathbf{k} k \mathbf{k}_{\infty} C_{1}>0$ and

$$
\begin{equation*}
\mathrm{k} e^{M} \mathbf{k}_{\mathcal{H}} \leq \frac{c_{k}}{C_{4}} d\left(C \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}}+e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathrm{k}^{\left(\mathrm{k}_{H_{\alpha \#}^{\frac{1}{\alpha}}\left(\Gamma_{1, \pm)}\right)}\right)}\right. \tag{D.124}
\end{equation*}
$$

with $c_{\text {min }}=\inf _{|n|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$. We can use equations (4.93) and (4.95) together with Theorem 50 to get

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{c_{k}}{C_{4}} d\left(C \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}}+e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)}\right) \tag{D.125}
\end{equation*}
$$

From Lemma 59 and equations (D.125), (4.93) and (4.95) together with Theorem 50, we conclude that

$$
\begin{equation*}
\mathrm{ke}_{\alpha, 0}^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq 2 \frac{d}{C_{4}} c_{k} C_{1}\left(C \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}}+e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\alpha \#}^{\frac{1}{2}}\left(\Gamma_{1, \pm)}\right)}\right) . \tag{D.126}
\end{equation*}
$$

We can show that $U_{\alpha, 0}^{M}$ is unique in a similar way to that in Theorem 61.

## D.3.3 Total error made by solving numerically the problem

Similar to Section 4.3.3, we also note that the error made by solving numerically the Helmholtz equation for Cases $1 \mathrm{~B}, 2 \mathrm{~A}$ and 2 B for a periodic grating arises from discretising using finite elements and truncating the $\operatorname{DtN}$ operator when we use the transparent boundary conditions. Let us denote the total error by $e_{\alpha, 0}=U_{\alpha, 0}-U_{\alpha, 0_{h}}^{M}$ then it can be estimated as follows for the three cases.

## - Case 1B: Perfectly conducting grating: TM case

Theorem D-36. Let $|k| \geq k_{\text {ref }}>0$, the maximum mesh size $h \in\left[0, h_{0}\right]$, the degree of the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$ with $\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}<1$ where $C_{1}$ is defined in Lemma D-23. Let $C_{3}=1-\left(\Re(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}$. Let $M \in \mathbb{N}$ such that $M \geq M_{0}$ and let $U_{\alpha, 0}$ be the continuous solution of equation (D.55), $U_{\alpha, 0_{h}}^{M}$ be the corresponding discretized solution with the truncated DtN operator and the total error be $e_{\alpha, 0}=U_{\alpha, 0}-U_{\alpha, 0_{h}}^{M}$. Then we have

$$
\begin{aligned}
\mathrm{k}_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq & 4 \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \\
& +4 \frac{d}{C_{3}} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathrm{k} e_{\alpha, 0} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \leq 2 \frac{C_{c}}{C_{3}} C_{1} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \\
& +2 d \frac{C_{1}}{C_{3}} e^{-(B-b) c_{\text {min }} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}\left(\Gamma_{1, \pm}\right)}},
\end{aligned}
$$

for all test functions $\psi_{\alpha, 0} \in X^{\alpha}$ where $C_{c}$ is given in Lemma D-21, $c_{\min }=$ $\inf _{|n|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$ with $z_{n}$ as defined in equation (2.44).
Since we have equation (4.114), then we can use the standard finite element esti-
mate equation (4.74) and we can write

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \leq & 4 \sup \left(1, \mathrm{k} k \mathbf{k}_{\infty}\right)\left(\frac{h}{p}\right)^{l-1} \frac{C_{c}}{C_{3}} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{H_{\#}^{l}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \\
& +4 \frac{d}{C_{3}} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}
\end{aligned} \mathrm{k}_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{2}\left(\Gamma_{1, \pm}\right)}},
$$

and using the definition of $C_{1}$ from Lemma D-23

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{L_{\#}^{2}\left(\Omega \backslash \operatorname{int} \Omega_{3}\right)} \leq & 2 \mathbf{k} k \mathbf{k}_{\infty}\left(\frac{h}{p}\right)^{l} \frac{C_{c}}{C_{3}}(C d+1) C_{r e g} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{H_{\#}^{l}\left(\Omega \backslash \text { int } \Omega_{3}\right)} \\
& +2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
\mathrm{k} & U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

for any integer $l \geq 2$.
Proof. The proof is similar to that of Theorem 62.

## - Case 2A: Transmitting dielectric grating: TE case

Theorem D-37. Let $|k| \geq k_{\text {ref }}>0$, the maximum mesh size $h \in\left[0, h_{0}\right]$, the degree of the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$ with $\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}<1$ where $C_{1}$ is defined in Lemma D-27. Let $C_{3}=1-\left(\Re(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}$ and $M \in \mathbb{N}$ such that $M \geq M_{0}$. Let $U_{\alpha, 0}$ be the continuous solution of equation (D.56), $U_{\alpha, 0_{h}}^{M}$ be the corresponding discretized solution with the truncated DtN operator and the total error be $e_{\alpha, 0}=U_{\alpha, 0}-U_{\alpha, 0_{h}}^{M}$. Then we have

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq & 4 \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \\
& +4 \frac{d}{C_{3}} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H} \frac{\frac{1}{\#}\left(\Gamma_{1, \pm}\right)}{},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{L_{\#}^{2}(\Omega)} \leq & 2 \frac{C_{c}}{C_{3}} C_{1} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \\
& +2 d \frac{C_{1}}{C_{3}} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H}{ }_{H}^{\frac{1}{\#}\left(\Gamma_{1, \pm}\right)}
\end{aligned}
$$

for all $\psi_{\alpha, 0} \in X^{\alpha}$ where $C_{c}$ is given by Lemma D-25, $c_{\min }=\inf _{|n|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$ with $z_{n}$ as defined in equation (2.44).

Note that

$$
\begin{equation*}
\inf \left(1, \mathbf{k} k \mathbf{k}_{\infty}\right) \mathbf{k} e_{\alpha, 0} \mathbf{k}_{H_{\#}^{1}(\Omega)} \leq \mathrm{k} e_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \leq \sup \left(1, \mathrm{k} k \mathbf{k}_{\infty}\right) \mathrm{k} e_{\alpha, 0} \mathrm{k}_{H_{\#}^{1}(\Omega)} . \tag{D.127}
\end{equation*}
$$

Then we can use the standard finite element estimate equation (D.88) and we can write

$$
\begin{aligned}
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq & 4 \sup \left(1, \mathrm{k} k \mathrm{k}_{\infty}\right)\left(\frac{h}{p}\right)^{l-1} \frac{C_{c}}{C_{3}} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{H_{\#}^{l}(\Omega)} \\
& +4 \frac{d}{C_{3}} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm)}\right)},
\end{aligned}
$$

and using the definition of $C_{1}$ from Lemma D-27

$$
\begin{aligned}
\mathbf{k} e_{\alpha, 0} \mathbf{k}_{L_{\#}^{2}(\Omega)} \leq & 2 \mathbf{k} k \mathbf{k}_{\infty}\left(\frac{h}{p}\right)^{l} \frac{C_{c}}{C_{3}}(C d+1) C_{s t a b} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{H_{\#}^{l}(\Omega)} \\
& +2 \frac{d}{C_{3}} C_{1} e^{-(B-b) c_{\min } \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}}} \mathbf{k} U_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

for any integer $l \geq 2$.
Proof. We have

$$
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq \mathrm{k} U_{\alpha, 0}-U_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}}+\mathrm{k} U_{\alpha, 0}^{M}-U_{\alpha, 0_{h}}^{M} \mathrm{k}_{\mathcal{H}}
$$

Similar to the proof of Theorem 62, we derive an a priori error estimate for the second term. Let $e_{h}^{M}=U^{M}-U_{h}^{M}$ where $U^{M}=e^{i \alpha x} U_{\alpha, 0}^{M}$ and $U_{h}^{M}=e^{i \alpha x} U_{\alpha, 0_{h}}^{M}$. In a similar way to derive equation (4.115), we get

$$
\mathfrak{R}\left(a^{M}\left(e_{h}^{M}, e_{h}^{M}\right)\right) \geq \mathfrak{R}\left(\left|e_{h}^{M}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-k^{2} \mathbf{k} e_{h}^{M} \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}^{2}\right)
$$

since $-\mathfrak{R}\left(T_{ \pm}^{M} e_{h}^{M}, e_{h}^{M}\right)_{\Gamma_{ \pm}}>0$ using Lemma 8. So,

$$
\left|e_{h}^{M}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathfrak{R}(k) \mathbf{k} e_{h}^{M} \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)} \mathbf{k} e_{h}^{M} \mathbf{k}_{\mathcal{H}} \leq\left|a^{M}\left(e_{h}^{M}, e_{h}^{M}\right)\right|
$$

since $\mathfrak{R}(k) \mathrm{k} e_{h}^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq \mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}}$. Then

$$
\begin{aligned}
\left|e_{h}^{M}\right|_{H_{\alpha \#}^{1}(\Omega)}^{2}-\mathfrak{R}(k) \operatorname{ke}_{h}^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \mathfrak{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}} & \leq\left|a^{M}\left(e_{h}^{M}, U^{M}-\psi\right)\right| \\
& \leq C_{c} \mathbf{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}} \mathbf{k} U^{M}-\psi \mathbf{k}_{\mathcal{H}}
\end{aligned}
$$

using Galerkin orthogonality and Lemma D-26. Similar to the proof of Theorem 62 to derive equation (4.117), we have

$$
\mathrm{ke}_{h}^{M} \mathrm{k}_{L_{\alpha \neq}^{2}(\Omega)} \leq C_{1} \mathrm{ke}_{h}^{M} \mathrm{k}_{\mathcal{H}}
$$

. Hence,

$$
\mathrm{k}_{h}{ }_{h}^{M} \mathrm{k}_{\mathcal{H}}-\left(\Re(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1} \mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}} \leq C_{c} \mathrm{k} U^{M}-\psi \mathrm{k}_{\mathcal{H}} \leq C_{c} \mathrm{k} U-\psi \mathbf{k}_{\mathcal{H}}
$$

using Lemma $\mathrm{D}-25$ and by noting that when $M \geq M_{0}, U^{M}$ tends to $U$. Since we suppose that $\left(\mathfrak{R}(k)+\mathrm{k} k \mathrm{k}_{\infty}\right) C_{1}<1$ then

$$
\begin{equation*}
\mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}} \leq \frac{C_{c}}{C_{3}} \mathbf{k} U-\psi \mathbf{k}_{\mathcal{H}} \tag{D.128}
\end{equation*}
$$

where $C_{3}=1-\left(\Re(k)+\mathbf{k} k \mathbf{k}_{\infty}\right) C_{1}$. From equations (4.93), (4.95) and (D.128), we get

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0_{h}}^{M} \mathrm{k}_{\mathcal{H}} \leq 4 \frac{C_{c}}{C_{3}} \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} . \tag{D.129}
\end{equation*}
$$

From Lemma 59 and Theorem 50, we get

$$
\begin{equation*}
\mathfrak{k e} e_{\alpha, 0_{h}}^{M} \mathrm{k}_{L_{\#}^{2}(\Omega)} \leq 2 \frac{C_{1}}{C_{3}} C_{c} \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \tag{D.130}
\end{equation*}
$$

We use the result given by Theorem D-34 with equations (D.129) and(D.130) to finish the proof of the total error of discretizing and truncating the DtN operator.

## - Case 2B: Transmitting dielectric grating: TM case

Theorem D-38. Let $|k| \geq k_{\text {ref }}>0$, the maximum mesh size $h \in\left[0, h_{0}\right]$, the degree of the polynomial basis $p \in\left[p_{0}, \infty\right]$ such that $k \frac{h_{0}}{p_{0}}<1$ with $2 C_{1} \mathrm{k} k \mathrm{k}_{\infty}<1$ so that $C_{4}=1-2 C_{1} \mathbf{k} k \mathrm{k}_{\infty}>0$ where $C_{1}$ is defined in Lemma D-31. Let $M \in \mathbb{N}$ such that $M \geq M_{0}$ and let $U_{\alpha, 0}$ be the continuous solution of equation (D.57), $U_{\alpha, 0_{h}}^{M}$ be the corresponding discretized solution with the truncated DtN operator and the total error be $e_{\alpha, 0}=U_{\alpha, 0}-U_{\alpha, 0_{h}}^{M}$. Then we have

$$
\begin{aligned}
& \mathbf{k} e_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \leq 4 \frac{c_{k}}{C_{4}}(3 C d+1) \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \\
&+4 \frac{c_{k}}{C_{4}} d e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
& \mathbf{k} U_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathrm{k} e_{\alpha, 0} \mathrm{k}_{L_{\#}^{2}(\Omega)} \leq \frac{c_{k}}{C_{4}} C_{1}(3 C d+1) \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \\
&+2 d \frac{c_{k}}{C_{4}} C_{1} e^{-(B-b) c_{\text {min }}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
& \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H}{ }_{\#\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

for all $\psi_{\alpha, 0} \in X^{\alpha}$ where $c_{k}=\frac{\|k\|_{\infty}}{k_{r e f}}, C_{c}$ is given by Lemma D-29, $C \leq{ }^{\sqrt{ }} \overline{5}$ as in Lemma 7 and $c_{\text {min }}=\inf _{|n|>\frac{M d}{2 \pi}} \sin \left(z_{n} / 2\right)$ with $z_{n}$ as defined in equation (2.44). Since we have equation (D.127) then we have

$$
\begin{aligned}
& \mathbf{k} e_{\alpha, 0} \mathbf{k}_{\mathcal{H}} \leq 4 \sup \left(1, \mathbf{k} k \mathbf{k}_{\infty}\right)\left(\frac{h}{p}\right)^{l-1} \frac{c_{k}}{C_{4}}(3 C d+1) \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{H_{\#}^{l}(\Omega)} \\
&+4 \frac{c_{k}}{C_{4}} d e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
& \mathrm{k} U_{\alpha, 0} \mathrm{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)},
\end{aligned}
$$

and using the definition of $C_{1}$ from Lemma D-31

$$
\begin{aligned}
& \mathbf{k} e_{\alpha, 0} \mathbf{k}_{L_{\#}^{2}(\Omega)} \leq 2 \mathbf{k} k \mathbf{k}_{\infty}\left(\frac{h}{p}\right)^{l} \frac{c_{k}}{C_{4}} C_{s t a b}(3 C d+1) \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{H_{\#}^{l}(\Omega)} \\
&+2 d \frac{c_{k}}{C_{4}} C_{1} e^{-(B-b) c_{m i n}} \sqrt{(M-|\alpha|)^{2}-k_{j}^{2}} \\
& \mathbf{k} U_{\alpha, 0} \mathbf{k}_{H_{\#}^{\frac{1}{2}}\left(\Gamma_{1, \pm}\right)}
\end{aligned}
$$

Proof. We have

$$
\mathrm{k} e_{\alpha, 0} \mathrm{k}_{\mathcal{H}} \leq \mathrm{k} U_{\alpha, 0}-U_{\alpha, 0}^{M} \mathrm{k}_{\mathcal{H}}+\mathrm{k} U_{\alpha, 0}^{M}-U_{\alpha, 0_{h}}^{M} \mathrm{k}_{\mathcal{H}}
$$

where $\mathbf{k} U_{\alpha, 0}-U_{\alpha, 0}^{M} \mathbf{k}_{\mathcal{H}}$ has already been shown in Theorem D-35. An a priori error estimate for the second term can be derived in similar way to that performed in Theorem D-37. Let us denote $e_{h}^{M}=U^{M}-U_{h}^{M}$ where $U^{M}=e^{i \alpha x} U_{\alpha, 0}^{M}$ and $U_{h}^{M}=e^{i \alpha x} U_{\alpha, 0_{h}}^{M}$. We can show similarly as for Lemma D-30 that

$$
\frac{1}{\mathrm{k} \Re\left(k^{2}\right) \mathrm{k}_{\infty}}\left|e_{h}^{M}\right|_{H_{\alpha \neq 1}^{1}(\Omega)}^{2}-\mathrm{k} e_{h}^{M} \mathrm{k}_{L_{\alpha \neq}^{2}(\Omega)}^{2} \leq\left|a^{M}\left(e_{h}^{M}, e_{h}^{M}\right)\right|+\left|\left(\frac{1}{k^{2}} T_{ \pm} e_{h}^{M}, e_{h}^{M}\right)_{\Gamma_{ \pm}}\right| .
$$

Similar to that proof of Theorem D-28, we have

$$
\begin{aligned}
& \frac{1}{\left\|\Re\left(k^{2}\right)\right\|_{\infty}} \mathbf{k} e_{h}^{M} \mathbf{k}_{\mathcal{H}}^{2}-2 \mathbf{k} k \mathbf{k}_{\infty}^{2} \mathrm{k}_{h}^{M} \mathrm{k}_{L_{\alpha \neq}^{2}}^{2}(\Omega) \\
& \leq\left|a^{M}\left(e_{h}^{M}, e_{h}^{M}\right)\right|+\frac{1}{k_{\text {ref }}^{2}}\left|\left(T_{ \pm}^{M} e_{h}^{M}, e_{h}^{M}\right)\right| .
\end{aligned}
$$

Similarly as we did to get equation (D.107), we get

$$
\begin{aligned}
& \mathbf{k} e_{h}^{M} \mathbf{k}_{\mathcal{H}}-2 \mathbf{k} k \mathbf{k}_{\infty}{\mathbf{k} e_{h}^{M} \mathbf{k}_{L_{\alpha \#}^{2}(\Omega)}}^{\leq c_{k}(2 C d+1) \mathbf{k} U^{M}-\psi \mathbf{k}_{\mathcal{H}}}
\end{aligned}
$$

where $\psi=e^{i \alpha x} \psi_{\alpha, 0}$ such that $\psi_{\alpha, 0} \in X^{\alpha}$. In a similar way to derive Lemma D-31, we have $\mathrm{k} e_{h}^{M} \mathrm{k}_{L_{\alpha \#}^{2}(\Omega)} \leq C_{1} \mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}}$. We can divide by $\mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}}$ and so

$$
\begin{aligned}
& \mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}}-2 C_{1} \mathrm{k} k \mathrm{k}_{\infty} \mathrm{k}_{h}^{M} \mathrm{k}_{\mathcal{H}} \\
& \leq c_{k}(2 C d+1) \mathrm{k} U^{M}-\psi \mathrm{k}_{\mathcal{H}} .
\end{aligned}
$$

Since we suppose $2 C_{1} \mathrm{k} k \mathrm{k}_{\infty}<1$ then $C_{4}=1-2 C_{1} \mathrm{k} k \mathrm{k}_{\infty}>0$ and we have

$$
\mathbf{k} e_{h}^{M} \mathbf{k}_{\mathcal{H}} \leq \frac{c_{k}}{C_{4}}(2 C d+1) \mathbf{k} U^{M}-\psi \mathbf{k}_{\mathcal{H}} .
$$

For $M \geq M_{0}, U^{M}$ tends to $U$ therefore

$$
\begin{equation*}
\mathrm{k} e_{h}^{M} \mathrm{k}_{\mathcal{H}} \leq \frac{c_{k}}{C_{4}}(2 C d+1) \mathrm{k} U-\psi \mathrm{k}_{\mathcal{H}} . \tag{D.131}
\end{equation*}
$$

From equations (4.93) and (4.95), we get

$$
\begin{equation*}
\mathfrak{k} e_{\alpha, 0_{h}}^{M} \mathbf{k}_{\mathcal{H}} \leq 4 \frac{c_{k}}{C_{4}}(2 C d+1) \mathbf{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathbf{k}_{\mathcal{H}} . \tag{D.132}
\end{equation*}
$$

From Lemma 59 and Theorem 50, we get

$$
\begin{equation*}
\mathrm{k} e_{\alpha, 0_{h}}^{M} \mathrm{k}_{L_{\# \#}^{2}(\Omega)} \leq 2 \frac{c_{k}}{C_{4}} C_{1}(2 C d+1) \mathrm{k} U_{\alpha, 0}-\psi_{\alpha, 0} \mathrm{k}_{\mathcal{H}} . \tag{D.133}
\end{equation*}
$$

We use the result given by Theorem D-35 with equations (D.132) and(D.133) to finish the proof of the total error of discretizing and truncating the DtN operator.

