Explicit error bounds for the α-quasi-periodic Helmholtz problem

Natacha H. Lord and Anthony J. Mulholland*

Department of Mathematics and Statistics, Strathclyde University, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, UK
**Corresponding author: anthony.mulholland@strath.ac.uk*

Received May 16, 2013; revised August 27, 2013; accepted August 27, 2013;
posted August 28, 2013 (Doc. ID 190680); published September 25, 2013

This paper considers a finite element approach to modeling electromagnetic waves in a periodic diffraction grating. In particular, an *a priori* error estimate associated with the α-quasi-periodic transformation is derived. This involves the solution of the associated Helmholtz problem being written as a product of $e^{i\alpha x}$ and an unknown function called the α-quasi-periodic solution. To begin with, the well-posedness of the continuous problem is examined using a variational formulation. The problem is then discretized, and a rigorous *a priori* error estimate, which guarantees the uniqueness of this approximate solution, is derived. In previous studies, the continuity of the Dirichlet-to-Neumann map has simply been assumed and the dependency of the regularity constant on the system parameters, such as the wavenumber, has not been shown. To address this deficiency, in this paper an explicit dependence on the wavenumber and the degree of the polynomial basis in the *a priori* error estimate is obtained. Since the finite element method is well known for dealing with any geometries, comparison of numerical results obtained using the α-quasi-periodic transformation with a lattice sum technique is then presented.

OCIS codes: (290.0290) Scattering; (050.1950) Diffraction gratings; (050.1960) Diffraction theory; (050.2770) Gratings; (080.1753) Computation methods; (080.2720) Mathematical methods (general).
http://dx.doi.org/10.1364/JOSAA.30.002111

1. INTRODUCTION

Periodic diffraction gratings have been used recently, for example, in crystalline silicon solar cells [1], gas sensors [2], and medical x-ray imaging [3,4]. The problem of wave diffraction is based on solving Maxwell's equations in the diffraction grating region and on finding the resulting electromagnetic field when an incident wave interacts with the grating [5]. As there are two types of waves 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. These cases will be denoted by Case 1A/B (perfectly conducting) and Case 2A/B (transmitting dielectric), where A (B) denotes the TE (TM) wave [5]. For brevity, the main focus is on the TM mode for the transmitting dielectric grating (Case 2B) in this article. There are of course other numerical methods in the literature to solve the problem of diffraction of waves [5–7]. This paper is an extension of the work in [8], which mainly focuses on the finite element approach. In fact, a finite element method and the periodicity of the grating with respect to one direction are used to address the problem over one period. Since the domain is infinite in the other direction, some transparent boundary conditions are applied to truncate the domain. The advantage of the finite element method is its flexibility in dealing with complex geometries. It also naturally gives rise to a variational formulation that provides a platform to rigorously derive existence and uniqueness results and regularity bounds. Hence, the well-posedness of the problem and an *a priori* error estimate can be derived.

In Section 2, the geometry, a statement of the Helmholtz problem to be solved, and the associated function spaces are presented. There have been a 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 [5,6,8,9]. In these studies the continuity of the Dirichlet-to-Neumann (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 in the analysis, and so these results are derived here for the first time in Section 2. The α-quasi-periodic method is studied in Section 3 with an examination of the continuous problem, its variational formulation, and its well-posedness. The problem is then discretized to approximate its solution and a new *a priori* error estimate is derived. This result guarantees the uniqueness of the approximate solution and shows an explicit dependence on the wavenumber k, the mesh size h, and the degree of the polynomial basis p. In order to keep this paper to a manageable size, the proofs of the majority of the results are relegated to online reports [10,11].

2. PHYSICAL AND MATHEMATICAL DESCRIPTION OF THE PROBLEM

The aim of this paper is to solve the Helmholtz equation for a periodic grating of period d (with respect to x), as shown in Fig. 1. In order to formulate the scattering problem as a boundary value problem, an appropriate radiation condition (outgoing wave condition) must be included. In this paper, electromagnetic waves interacting with a periodic diffraction

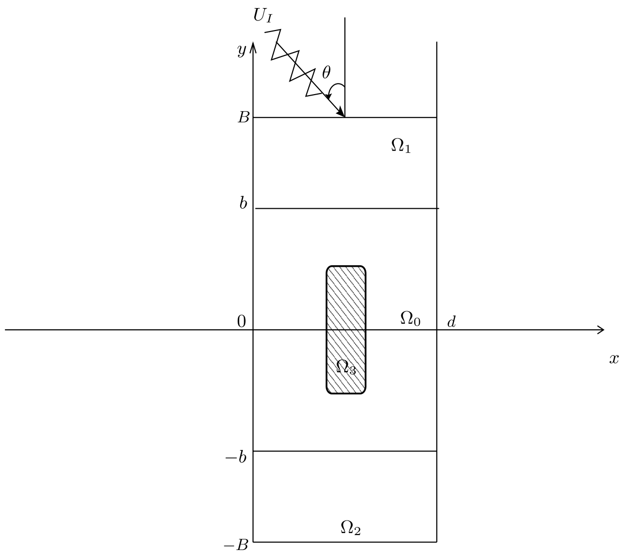


Fig. 1. Diagram showing the truncated periodic grating domain. Define Ω_1 to be the region above the scattering region $\{(x,y):0 \le x < d, b \le y \le B\}$, and the substrate Ω_2 to be $\{(x,y):0 \le x < d, -B \le y \le -b\}$.

grating are considered and hence the usual Sommerfeld radiation condition is not appropriate [12]; the radiating energy does not diminish in the direction of periodicity. The so-called upward propagating radiation condition (UPRC) is therefore utilized. It has been used previously to establish the uniqueness of the solution to the continuous problem associated with scattering from a periodic grating [12]. The periodicity of the grating is utilized to restrict the problem to a single vertical strip as shown in Fig. 1. The effects of the scatterers (the grating) are restricted to a horizontal strip $\Omega_0 = [0,d] \times [-b,b]$; we denote by $\Omega_3 \subset \Omega_0$ the scatterer (see Fig. 1), and the wavenumber k is

$$k(x,y) = \begin{cases} k_1 \in \mathbb{R}, \text{for } (x,y) \in \{0 \le x \le d, y \ge B\}, \\ k_1 \in \mathbb{R}, \text{for } (x,y) \in \Omega_1, \\ k_0 \in \mathbb{C}, \text{for } (x,y) \in \Omega_0 \setminus \Omega_3, \\ k_3 \in \mathbb{C}, \text{for } (x,y) \in \Omega_3, \\ k_2 \in \mathbb{C}, \text{for } (x,y) \in \Omega_2, \\ k_2 \in \mathbb{C}, \text{for } (x,y) \in \{0 \le x \le d, y \le -B\}, \end{cases}$$

with $k^2 = w^2\epsilon\mu$, where ϵ is the electric permittivity, μ is the magnetic permeability, w is the angular frequency, and subdomains $\Omega_1 = [0,d] \times [b,B]$ and $\Omega_2 = [0,d] \times [-B,-b]$. The incident wave is denoted by U_I, which is given by $U_I = e^{i\alpha x - i\beta_1^0 y}$, where $\alpha = k_1 \sin\theta$, $\beta_1^0 = k_1 \cos\theta$, and θ is the angle of incidence of the wave as shown in Fig. 1. By demanding that $\Re(k_j) > 0$ and $\Im(k_j) \ge 0$, where $\Re(k_j)$ ($\Im(k_j)$) denotes the real (imaginary) part of k_j, the scattered and diffracted waves are composed of bounded outgoing waves. The periodicity of the grating combined with the presence of the incident wave makes any solution α-quasi-periodic; that is, there exists a periodic function U_α with the same period as a solution U such that [7,8,13,14]

$$U(x,y) = e^{i\alpha x} U_\alpha(x,y). \quad (1)$$

The α-quasi-periodic method applies transformation (1) to the Helmholtz problem and solves the resulting scattering problem for the function U_α. It is more straightforward to implement periodic rather than quasi-periodic constraints using the finite element method since the quasi-periodic case involves the extra term $e^{i\alpha x}$. In addition, when there are high wavenumbers, the $e^{i\alpha x}$ term will oscillate rapidly and increase the computational error. There is motivation therefore to find U_α and not to solve the original Helmholtz problem directly.

A. Transparent Boundary Conditions (Dirichlet-to-Neumann Maps)

To solve the grating problem numerically for a wide range of grating geometries, a finite element method is used here. It is therefore necessary to truncate the domain to render it finite. To provide suitable boundary conditions for the finite element solver, an analytical solution (known as the Rayleigh expansion) in the adjacent domains is used. Transparent boundary conditions that match this analytical solution continuously and smoothly with the finite element solution inside the truncated region are employed. These transparent boundary conditions are captured by the DtN operators $T_\pm$ to match the Rayleigh expansion of the electromagnetic field on the boundary of the truncated region with the finite element solution inside the truncated domain. Denote the interfaces by $\Gamma_+ = \{(x,y){:}\ 0 \le x \le d, y = B\}$, $\Gamma_- = \{(x,y){:}\ 0 \le x \le d, y = -B\}$.

Since both periodic and α-quasi-periodic functions are used, the function spaces on the domain boundaries are

$$L^s_{\alpha\#}([0,d]) = \{g \in L^s([0,d]){:}\ g(d) = e^{i\alpha d} g(0)\},$$
$$H^s_{\#}([0,d]) = \{g \in H^s([0,d]){:}\ g(d) = g(0)\},$$
$$H^s_{\alpha\#}([0,d]) = \{g \in H^s([0,d]){:}\ g(d) = e^{i\alpha d} g(0)\},$$

and the function spaces inside $\Omega = \{(x,y){:}\ 0 \le x \le d, -B \le y \le B\}$ are

$$H^s_{\#}(\Omega) = \{f \in H^s(\Omega){:}\ f(d,y) = f(0,y), \quad \forall\, y \in [-B,B]\},$$
$$H^s_{\alpha\#}(\Omega) = \{f \in H^s(\Omega){:}\ f(d,y) = e^{i\alpha d} f(0,y), \quad \forall\, y \in [-B,B]\},$$

with $s \in \mathbb{R}$ and $L^s(\Omega)$; $H^s(\Omega)$, $L^s([0,d])$, and $H^s([0,d])$ are Sobolev spaces [15]. The following norm is used later to simplify the algebra.

Definition 1. *Let $F \subset \mathbb{R}^2$ and $v \in H^1(\mathrm{F})$ ([15]); then define [16]*

$$\|v\|^2_{\mathcal{H}} = |v|^2_{H^1(\mathrm{F})} + \|k\|^2_\infty \|v\|^2_{L^2(\mathrm{F})}.$$

Note that $\|v\|_{\mathcal{H}}$ is equivalent to $\|v\|_{H^1(\mathrm{F})}$ since

$$\inf\left(1, \frac{1}{\|k\|_\infty}\right) \|v\|_{\mathcal{H}} \le \|v\|_{H^1(\mathrm{F})} \le \sup\left(1, \frac{1}{\|k\|_\infty}\right) \|v\|_{\mathcal{H}}$$

using the definition of Sobolev norms [15]. Also note that $\|v\|_{\mathcal{H}} \le |v|_{H^1(\mathrm{F})} + \|k\|_\infty \|v\|_{L^2(\mathrm{F})}$. Since $H^l_{\alpha\#}(\mathrm{F}) \subset H^l(\mathrm{F})$ and $H^l_{\#}(\mathrm{F}) \subset H^l(\mathrm{F})$ for any $l \ge 0$, then for any $v \in H^1_{\alpha\#}(\mathrm{F})$ ($v \in H^1_{\#}(\mathrm{F})$), $\|v\|_{\mathcal{H}}$ is well defined.

The following property holds.

Theorem 2. *Let $v \in H^l_{\alpha\#}(\Omega)$ and let $v_\alpha \in H^l_{\#}(\Omega)$ such that*

$$v = e^{i\alpha x} v_\alpha.$$

Then,

$$\|v_\alpha\|_{H^l_{\#}(\Omega)} \le l2^l \|v\|_{H^l_{\alpha\#}(\Omega)} \tag{2}$$

for $l \ne 0$ and

$$\|v_\alpha\|_{L^2_{\#}(\Omega)} = \|v\|_{L^2_{\alpha\#}(\Omega)}. \tag{3}$$

Proof [11, p. 61].

The DtN maps are defined below.

Definition 3. *Let $f \in H^{1/2}_{\alpha\#}(\Gamma_\pm)$; then the DtN maps [8] are defined by*

$$Tf(x) = T_\pm f(x) = \sum_{n\in Z} i\beta^n_j f^{(n_\alpha)}(\pm B) e^{in_\alpha x},$$

where $Tf \in H^{-(1/2)}_{\alpha\#}(\Gamma_\pm)$, $n_\alpha = \alpha + (2\pi n/d)$, $\beta^n_j = e^{iz_n/2}(|k^2_j - n^2_\alpha|)^{1/2}$, and

$$z_n = \arg(k^2_j - n^2_\alpha), \tag{4}$$

for $j \in \{1,2\}$ and $f^{(n_\alpha)}(\pm B) = 1/d \int_0^d f(x, \pm B) e^{-in_\alpha x} \mathrm{d}x$. The wavenumbers must satisfy $k^2_j \ne n^2_\alpha$, and at these discrete frequencies these guided modes propagate without loss along the grating and cause the phenomenon of resonance [7,8].

The following property holds for $T_\pm$.

Lemma 4. *The inner product of a function g and its normal derivative on the boundary $\Gamma_\pm$ satisfy*

$$\left.\begin{aligned} \Re(Tg, g)_{\Gamma_\pm} &= -d\sum_{n\in\mathbb{Z}} \sin(z_n/2)|\beta^n_j||g^{(n_\alpha)}|^2 \le 0 \\ \Im(Tg, g)_{\Gamma_\pm} &= d\sum_{n\in\mathbb{Z}} \cos(z_n/2)|\beta^n_j||g^{(n_\alpha)}|^2 \ge 0 \end{aligned}\right\}. \tag{5}$$

Proof [11, p. 25].

From Definition 3, the DtN map can be used so that the scattering domain can be truncated at $\Gamma_\pm$ and the problem studied in this paper can be stated as follows [11, p. 38, Lemma 17]. The homogeneous Helmholtz problem is to find $U(x,y) \in C^2(\Omega)$, where

$$\nabla \cdot \left(\frac{1}{k^2(x,y)} \nabla U(x,y)\right) + U(x,y) = 0, \qquad (x,y) \in \Omega, \tag{6}$$

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

$$\partial_n U(x,y)|_{\Gamma_+} = T_+ U + g(x) \qquad x \in \Gamma_+, \tag{7}$$

$$\partial_n U(x,y)|_{\Gamma_-} = T_- U \qquad x \in \Gamma_-, \tag{8}$$

subject to the α-quasi-periodic condition given by $U(d,y) = e^{i\alpha d} U(0,y)$, where $y \in [-B, B]$. Here

$$\begin{aligned} T_+ U(x) &= \sum_{n\in\mathbb{Z}} i\beta^n_1 U^{(n_\alpha)}(B) e^{in_\alpha x}, \\ T_- U(x) &= \sum_{n\in\mathbb{Z}} i\beta^n_2 U^{(n_\alpha)}(-B) e^{in_\alpha x}, \\ g(x) &= -2i\beta^0_1 e^{-i\beta^0_1 B + i\alpha x}, \end{aligned}$$

with $k^2_j \ne n^2_\alpha$ $U^{(n_\alpha)}(\pm B) = 1/d \int_0^d U(x, \pm B) e^{-in_\alpha x} \mathrm{d}x$.

Note that $(1/k^2(x,y))\nabla U(x,y) \in C^1(\Omega)$ is from the interface condition that corresponds to the TM case [17]. Hence, $\nabla.((1/k^2(x,y))\nabla U(x,y))$ is well defined if $U(x,y) \in C^2(\Omega)$. The α-quasi-periodic approach [11] is used to solve the scattering problem and so additional DtN maps are required as detailed below.

Definition 5. *Let $f \in H^{1/2}_{\#}(\Gamma_\pm)$. Then define the DtN maps [13]*

$$T^\alpha f(x) = T^\alpha_\pm f(x) = \sum_{n\in\mathbb{Z}} i\beta^n_j f^{(n)}(\pm B) e^{i\frac{2\pi n}{d}x}, \tag{9}$$

where $T^\alpha_\pm f \in H^{-(1/2)}_{\#}(\Gamma_\pm)$ and

$$f^{(n)}(\pm B) = \frac{1}{d}\int_0^d f(x, \pm B) e^{-i2\pi n x/d} \mathrm{d}x.$$

In order to show the uniqueness of the solution to the scattering problem, one of the prerequisites is that T is a continuous operator.

Lemma 6. *The operator T: $H^{1/2}_{\alpha\#}(\Gamma_\pm) \to H^{-(1/2)}_{\alpha\#}(\Gamma_\pm)$ is a continuous linear form, for any $k^2_j \ne n^2_\alpha$, and there exists a positive constant $2 \le C^2 \le \sqrt{5}$ such that*

$$\|Tf\|^2_{H^{-\frac{1}{2}}_{\alpha\#}(\Gamma_\pm)} \le C^2 \sup(|k^2_j|, 1) \|f\|^2_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_\pm)}$$

for $j = 1, 2$. In addition, the bilinear form $T(f,g)$: $(f,g) \mapsto (Tf, g)_{\Gamma_\pm}$ is continuous on $H^{1/2}_{\alpha\#}(\Gamma_\pm) \times H^{1/2}_{\alpha\#}(\Gamma_\pm)$ and

$$|(Tf, g)_{\Gamma_\pm}| \le dC \sup(|k_j|, 1) \|f\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_\pm)} \|g\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_\pm)} \tag{10}$$

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

$$|\beta^n_j|^2 \le \begin{cases} C^2 |k^2_j|, & \text{if } |k^2_j| > n^2_\alpha, \\ C^2 n^2_\alpha & \text{if } |k^2_j| < n^2_\alpha. \end{cases}$$

Proof [10, p. 23].

To model the scattering problem, the homogeneous Helmholtz problem is used since the forcing term (incident wave) is part of the boundary conditions. But when it comes to the well-posedness of the problem, the inhomogeneous Helmholtz equation is needed to show the continuous dependence of the problem on the data (for any given forcing term). Hence, the regularity of the solution $U(x,y)$ is now investigated to enable an *a priori* error estimate to be derived. In general, the *inhomogeneous* Helmholtz problem is to find $(1/k^2(x,y))\nabla U(x,y) \in C^1(\mathbb{R}^2)$, such that the following is satisfied:

$$\nabla \cdot \left(\frac{1}{k^2(x,y)} \nabla U(x,y)\right) + U(x,y) = f(x,y), \tag{11}$$

subject to the radiation condition $\lim_{|y|\to\infty} U(x,y) = 0$. Utilizing the periodicity of the grating reduces the problem to one on the vertical single strip $S = [0,d] \times \mathbb{R}$, where f is a general given α-quasi-periodic with respect to x and a compact support in Ω; the solution U is also α-quasi-periodic with respect to x and must satisfy the UPRC [12, p. 30]. In the following theorem, the regularity of the solution to Eq. (11) is stated.

Theorem 7. *For any $\gamma = (\gamma_1, \gamma_2)$ 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_{reg} that is independent of the wavenumber k, with $k_j \neq n_\alpha$, such that the solution U of Eq. (11) satisfies*

$$\|D^\gamma U\|_{L^2(\Omega)} \leq C_{\mathrm{reg}}(1 + C_s C(k_0,k_3))\|k\|_\infty^{|\gamma|-1}\|f\|_{L^2(\Omega)},$$

where C_s is a constant independent of k, $C(k_0,k_3)$ is a constant that only depends on k_0 and k_3 [10, p. 219], $|\gamma| = \sqrt{\gamma_1^2+\gamma_2^2}$, and $\|D^\gamma U\|_{L^2(\Omega)}$ is an L^2-Sobolev norm of order γ [15,18,19].

Proof. Let U be the solution of the inhomogeneous equation (11); using the integral representation [17,20] of U leads to

$$U(x,y) = \int_S G_j(x-x_0, y-y_0) f(x_0,y_0)\mathrm{d}x_0\mathrm{d}y_0 \qquad (12)$$

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

$$G_j(x,y) = \frac{1}{2d}\sum_{n\in\mathbb{Z}} c_j^n \frac{e^{in_\alpha x + i\beta_j^n|y|}}{i\beta_j^n} + \frac{1}{2d}\sum_{n\in\mathbb{Z}} d_j^n \frac{e^{in_\alpha x - i\beta_j^n|y|}}{i\beta_j^n}, \qquad (13)$$

where c_j^n and d_j^n represent the coefficients of the Green's functions $G_j(x,y)$.

By introducing the notation

$$I_{mn}(y,y_0) = \int_{\mathbb{R}} \left(c_j^n \frac{e^{i\beta_j^n|y-y_0|}}{\beta_j^n} + d_j^n \frac{e^{-i\beta_j^n|y-y_0|}}{\beta_j^n} \right) f^{(m_\alpha)}(y_0)\mathrm{d}y_0,$$

we note by using the continuity of I_{mn} with respect to y_0 along with the outgoing wave boundary condition that

$$I_{mn}(y,y_0) = 0 \quad \text{if } y_0 \neq y, \qquad (14)$$

and

$$I_{mn}(y,y) = \frac{1}{\beta_j^n}(c_j^n + d_j^n) f^{(m_\alpha)}(y) \quad \text{if } y_0 = y. \qquad (15)$$

We have

$$U(x,y) = \frac{1}{d}\int_{[0,d]\times\mathbb{R}} \sum_{n\in\mathbb{Z}} e^{in_\alpha(x-x_0)} \left(c_j^n \frac{e^{i\beta_j^n|y-y_0|}}{2i\beta_j^n} + d_j^n \frac{e^{i\beta_j^n|y-y_0|}}{2i\beta_j^n} \right) \times \sum_{m\in\mathbb{Z}} e^{im_\alpha x_0} f^{(m_\alpha)}(y_0)\mathrm{d}x_0\mathrm{d}y_0.$$

Now $\sum_{n\in\mathbb{Z}}(c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{i\beta_j^n|y-y_0|})/(2i\beta_j^n)$ and f are continuous with compact support and the integral is well defined so we can use *Fubini*'s theorem [21, p. 110] to interchange the order of summation and integration to get

$$U(x,y) = \frac{1}{d}\sum_{n\in\mathbb{Z}}\int_{[0,d]\times\mathbb{R}} e^{in_\alpha(x-x_0)} \left(c_j^n \frac{e^{i\beta_j^n|y-y_0|}}{2i\beta_j^n} + d_j^n \frac{e^{i\beta_j^n|y-y_0|}}{2i\beta_j^n} \right) \times \sum_{m\in\mathbb{Z}} e^{im_\alpha x_0} f^{(m_\alpha)}(y_0)\mathrm{d}x_0\mathrm{d}y_0 \qquad (16)$$

$$= \frac{1}{d}\sum_{m,n\in\mathbb{Z}}\int_{[0,d]} \frac{e^{in_\alpha(x-x_0)}}{2i} e^{im_\alpha x_0} I_{mn}(y,y_0)\mathrm{d}x_0. \qquad (17)$$

Since $1/d\int_0^d e^{-in_\alpha x_0 + im_\alpha x_0}\mathrm{d}x_0 = 1$ if $m = n$ and the integral is 0 otherwise, then

$$U(x,y) = \sum_{n\in\mathbb{Z}} \frac{e^{in_\alpha(x)}}{2i} I_{nn}(y,y_0).$$

Using Eq. (15),

$$U(x,y) = \sum_{n\in\mathbb{Z}} \frac{e^{in_\alpha(x)}}{2i}\left(\frac{1}{\beta_j^n}(c_j^n + d_j^n) f^{(n_\alpha)}(y)\right)$$

and so

$$\|U\|_{L^2(S)} \leq \sup_{n\in\mathbb{Z},j} \frac{\{|c_j^n|, |d_j^n|\}}{|\beta_j^n|} \left\|\sum\nolimits_{n\in\mathbb{Z}} e^{in_\alpha x} f^{(n_\alpha)}(y)\right\|_{L^2(S)}.$$

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

$$\|U\|_{L^2(\Omega)} \leq \sup_{j\in\{0,1,2,3\},n\in\mathbb{Z}} \frac{\{|c_j^n|, |d_j^n|\}}{|\beta_j^n|} \|f(x,y)\|_{L^2(\Omega)}. \qquad (18)$$

Taking the weak derivative of U with respect to x, in Eq. (16), γ_1 times gives

$$\partial_x^{\gamma_1} U = \frac{1}{d}\sum_{m,n\in\mathbb{Z}}\int_0^d (in_\alpha)^{\gamma_1} e^{in_\alpha(x-x_0)} e^{im_\alpha(x_0)}, \times \int_{\mathbb{R}} \left(\frac{c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{-i\beta_j^n|y-y_0|}}{2i\beta_j^n} \right) f^{(m_\alpha)}(y_0)\mathrm{d}y_0\mathrm{d}x_0,$$

and so

$$\|\partial_x^{\gamma_1} U\|_{L^2(S)} \leq \frac{1}{d}\sup_{n\in\mathbb{Z}}\|n_\alpha\|_\infty^{\gamma_1} \left\| \sum_{m,n\in\mathbb{Z}}\int_0^d e^{in_\alpha(x-x_0)} e^{im_\alpha(x_0)}, \times \int_{\mathbb{R}} \left(\frac{c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{-i\beta_j^n|y-y_0|}}{2i\beta_j^n} \right) \times f^{(m_\alpha)}(y_0)\mathrm{d}y_0\mathrm{d}x_0 \right\|_{L^2(S)}.$$

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

$$\|\partial_x^{\gamma_1} U\|_{L^2(\Omega)} \leq \sup_{n\in\mathbb{Z}}\|n_\alpha\|_\infty^{\gamma_1}\|U\|_{L^2(\Omega)}. \qquad (19)$$

We can do exactly the same with the weak derivative of U with respect to y, but we need to take into account that $\sum_{n\in\mathbb{Z}}((c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{-i\beta_j^n|y-y_0|})/2i\beta_j^n) f^{(m_\alpha)}(y_0)$ is no longer continuous for $\gamma_2 > 1$ on each interface separating each medium S_j of constant wavenumber k_j. Therefore

$$\partial_y^{\gamma_2} U = \frac{1}{d} \sum_{m,n\in\mathbb{Z}} \int_0^d e^{in_\alpha(x-x_0)} e^{im_\alpha(x_0)} \times \int_{\mathbb{R}} \frac{(i\beta_j^n)^{\gamma_2} c_j^n e^{i\beta_j^n|y-y_0|} + (-i\beta_j^n)^{\gamma_2} d_j^n e^{-i\beta_j^n|y-y_0|}}{2i\beta_j^n} f^{(m_\alpha)}(y_0)\mathrm{d}y_0\mathrm{d}x_0 + \sum_{p=2}^{\gamma_2} \left[\sum_{n\in\mathbb{Z}} \frac{(c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{-i\beta_j^n|y-y_0|})}{2i\beta_j^n} f^{(n_\alpha)}(y_0) \right]_{S_j\cap S_l} \times f^{(n_\alpha)}(y_0), \tag{20}$$

such that $[\sum_{n\in\mathbb{Z}}((c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{-i\beta_j^n|y-y_0|})/2i\beta_j^n) f^{(n_\alpha)}(y_0)]_{S_j\cap S_l}$ denotes the jump at the interface separating S_j and S_l with j, $l \in 0, 1, 2, 3$,. Hence,

$$\left| \left[\sum_{n\in\mathbb{Z}} \left(\frac{c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{-i\beta_j^n|y-y_0|}}{2i\beta_j^n} \right) f^{(n_\alpha)}(y_0) \right]_{S_j\cap S_l} \right| \le \sup_{n\in\mathbb{Z}, l\in\{0,1,2,3\}} (|\beta_j^n|^{p-2} |c_j^n e^{i\beta_j^n|y-y_0|}|) \sum_{n\in\mathbb{Z}} |f^{(n_\alpha)}(y_0)| + \sup_{n\in\mathbb{Z}, l\in\{0,1,2,3\}} (|\beta_j^n|^{p-2} |d_j^n e^{-i\beta_j^n|y-y_0|}|) \sum_{n\in\mathbb{Z}} |f^{(n_\alpha)}(y_0)|.$$

First, we note that when $j \in \{1, 2\}$, $S_j \cap S_0$ is given by $y = \pm b$; therefore $y - y_0 = 0$ on the interface and so

$$|e^{\pm i\beta_j^n|y-y_0|}| = 1, \tag{21}$$

We also note that, for $y \ne y_0$,

$$|e^{i\beta_j^n|y-y_0|}| \le 1, \tag{22}$$

because $\Im(\beta_j^n) > 0$ and

$$|e^{-i\beta_j^n|y-y_0|}| \le e^{\sin(z_n/2)\|k_j\| \sup\left\{\frac{2\pi N_0}{d}, |\alpha|\right\} / k_{\mathrm{ref}} \sup_{\{y_0,y\}\in\partial\Omega_3} |y-y_0|}, \tag{23}$$

because $\Im(\bar{\beta}_j^n) = \sin(z_n/2)|k_j^2 - n_\alpha^2|^{1/2}$. From the radiation condition, there exists $\tilde{N}_0$ such that for $|n| > \tilde{N}_0$, d_j^n equals zero. Therefore, $|k_j^2 - n_\alpha^2| \le |k_j^2||1 - (n_\alpha^2/k_j^2)| \le 4|k_j^2|N_0^2/k_{\mathrm{ref}}^2$ such that $|k_j| > k_{\mathrm{ref}}$ for $j \in \{0, 1, 2, 3\}$ and $N_0 = \sup\{(2\pi\tilde{N}_0/d)^2, |\alpha|^2\}$. Combining Eqs. ([21]), ([22]), and ([23]), we have

$$\left| \left[\sum_{n\in\mathbb{Z}} \left(\frac{c_j^n e^{i\beta_j^n|y-y_0|} + d_j^n e^{-i\beta_j^n|y-y_0|}}{2i\beta_j^n} \right) f^{(n_\alpha)}(y_0) \right]_{S_j\cap S_l} \right| \le \sup_{n\in\mathbb{Z}, j\in\{0,1,2,3\}} |\beta_j^n|^{p-2} (|c_j^n|, |d_j^n|) \times \sup \left(e^{\sin(z_n/2)|k_j|\frac{N_0}{k_{\mathrm{ref}}} \sup_{\{y_0,y\}\in S_0\cap S_3} |y-y_0|}, 1 \right) \sum_{n\in\mathbb{Z}} |f^{(n_\alpha)}(y_0)| \le \sup_{n\in\mathbb{Z}, j\in\{0,1,2,3\}} C_{s0} \left(e^{\sin(z_n/2)|k_j|\frac{N_0}{k_{\mathrm{ref}}} \sup_{\{y_0,y\}\in S_0\cap S_3} |y-y_0|}, 1 \right) |\beta_j^n|^{p-2} \|f\|_{L^2(S)}.$$

with $C_{s0} = \sup_{n\in\mathbb{Z}, j\in\{0,1,2,3\}} (|c_j^n|, |d_j^n|)$. Hence, from Eqs. ([16]) and ([20]),

$$\|\partial_y^{\gamma_2} U\|_{L^2(S)} \le \sup_{n\in\mathbb{Z}} \|\beta_j^n\|_\infty^{\gamma_2} \|U\|_{L^2(S)} + \sup_{n\in Z, j\in\{0,1,2,3\}} C_{s0} \left(e^{\sin z_n/2|k_j|\frac{N_0}{k_{\mathrm{ref}}} \sup_{\{y_0,y\}\in S_0\cap S_3} |y-y_0|}, 1 \right) \times \left(\sum_{p=2}^{\gamma_2} |\beta_j^n|^{p-2} \right) \|f\|_{L^2(S)} \le \sup_{n\in\mathbb{Z}} \|\beta_j^n\|_\infty^{\gamma_2} \|U\|_{L^2(S)} + \sup_{n\in Z, j\in\{0,1,2,3\}} C_{s0} \left(e^{\sin z_n/2|k_j|\frac{N_0}{k_{\mathrm{ref}}} \sup_{\{y_0,y\}\in S_0\cap S_3} |y-y_0|}, 1 \right) \times (\gamma_2 - 1)|\beta_j^n|^{\gamma_2-2} \|f\|_{L^2(S)}. \tag{24}$$

We note that n_α satisfies $n_\alpha = (2\pi n/d) + k \sin\theta = k \sin\theta_n$ [[10], p. 56] and β_j^n satisfies $\beta_j^n = e^{iz_n/2}\sqrt{|k^2 - n_\alpha^2|} = |k|e^{iz_n/2}\cos\theta_n$, [[10], p. 57]; therefore using Eq. ([18]), inequality ([19]) becomes

$$\|\partial_x^{\gamma_1} U\|_{L^2(S)} \le \sup_{n\in\mathbb{Z}} \|n_\alpha\|_\infty^{\gamma_1} \sup_{j\in\{0,1,2,3\}, n\in\mathbb{Z}} \frac{\{|c_j^n|, |d_j^n|\}}{|\beta_j^n|} \|f(x,y)\|_{L^2(\Omega)} \le \sup_{j\in\{0,1,2,3\}, n\in\mathbb{Z}} \|k\|_\infty^{\gamma_1-1} \frac{|\sin\theta_n|^{\gamma_1}}{|\cos\theta_n|} \times \{|c_j^n|, |d_j^n|\} \|f(x,y)\|_{L^2(\Omega)}.$$

In a similar fashion to Eq. ([18]), inequality ([24]) becomes

$$\|\partial_y^{\gamma_2} U\|_{L^2(S)} \le \sup_{j\in\{0,1,2,3\}, n\in\mathbb{Z}} \{|c_j^n|, |d_j^n|\} (\|k\|_\infty |e^{iz_n/2}\cos\theta_n|)^{\gamma_2-1} \| \times f(x,y)\|_{L^2(\Omega)} + C(k_0,k_3)C_s \sup_{n\in\mathbb{Z}, j} |\beta_j^n|^{\gamma_2-2} \|f\|_{L^2(S)} \le \sup_{n\in\mathbb{Z}} \|k\|_\infty^{\gamma_2-1} (|\cos\theta_n|)^{\gamma_2-1} \|f\|_{L^2(S)} + C(k_0,k_3)C_s \sup_{n\in\mathbb{Z}, j} (\|k\|_\infty |e^{iz_n/2}\cos\theta_n|)^{\gamma_2-1} \|f\|_{L^2(S)}$$

with $C_s = C_{s0}(\gamma_2 - 1)$ and $C(k_0, k_3) = \sup_{n\in Z, j\in\{0,3\}} (e^{\sin z_n/2|k_j|(N_0/k_{\mathrm{ref}})\sup_{\{y_0,y\}\in S_0\cap S_3}|y-y_0|}, 1)$. Denoting

$$C_{\mathrm{reg}} = \sup_{n\in\mathbb{Z}} \left\{ \frac{|\sin\theta_n|^{\gamma_1}}{|\cos\theta_n|}, |\cos\theta_n|^{\gamma_2-1} \right\},$$

which is well defined because $\beta_j^n \ne 0$, therefore $\cos\theta_n \ne 0$. Hence,

$$\|\partial_y^{\gamma_2} U\|_{L^2(S)} \le (C_{\mathrm{reg}}\|k\|_\infty^{\gamma_2-1} + C_{\mathrm{reg}}\|k\|_\infty^{\gamma_2-1} C(k_0,k_3)C_s)\|f\|_{L^2(S)} \le C_{\mathrm{reg}}(1 + C_s C(k_0,k_3))\|k\|_\infty^{\gamma_2-1}\|f\|_{L^2(S)}.$$

Since $\mathrm{supp} f \subset \Omega$,

$$\|\partial_y^{\gamma_2} U\|_{L^2(\Omega)} \le C_{\mathrm{reg}}(1 + C_s C(k_0,k_3))\|k\|_\infty^{\gamma_2-1}\|f\|_{L^2(\Omega)}. \tag{25}$$

Combining Eqs. ([18]), ([19]), and ([25]), we finish the proof.

3. WELL-POSEDNESS OF THE PERIODIC PROBLEM

Although it is easier to study the scattering problem analytically using U, since it lends itself more readily to a variational formulation, it will transpire that it is more efficient to implement periodic rather than quasi-periodic boundary conditions within the finite element method. This is essentially due to the quasi-periodic boundary condition containing the extra term $e^{i\alpha x}$, which is oscillatory and can lead to computational errors. From Eq. (1), the Helmholtz problem (6) is transformed, and the following lemma holds.

Lemma 8. *Let $U_\alpha(x,y) \in C^2(\Omega)$ satisfy Eq. (1); then $U_\alpha(x,y)$ is the solution of the following problem:*

$$\nabla_\alpha \cdot \left(\frac{1}{k^2(x,y)} \nabla_\alpha U_\alpha(x,y)\right) + U_\alpha(x,y) = 0, \qquad (x,y) \in \Omega \tag{26}$$

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

$$\left(T_+^\alpha - \frac{\partial}{\partial n}\right) U_\alpha = 2i\beta_1^0 e^{-i\beta_1^0 B}, \qquad \text{on } \Gamma_+,$$
$$\left(T_-^\alpha - \frac{\partial}{\partial n}\right) U_\alpha = 0, \qquad \text{on } \Gamma_-.$$

The periodic condition

$$U_\alpha(d,y) = U_\alpha(0,y), \qquad y \in [-B,B],$$

holds, where $U(x,y)$ is the solution of the original Helmholtz problem given by Eq. (6), where $T_\pm^\alpha$ is given by Eq. (9) and $\nabla_\alpha = \nabla + i(\alpha,0)$.

Proof. Since $U = e^{i\alpha x} U_\alpha$ and U satisfies Eq. (6), we have

$$\nabla U = \nabla(e^{i\alpha x} U_\alpha) = \nabla(e^{i\alpha x}) U_\alpha + e^{i\alpha x} \nabla U_\alpha$$
$$= \begin{bmatrix} i\alpha e^{i\alpha x} \\ 0 \end{bmatrix} U_\alpha + e^{i\alpha x} \nabla U_\alpha.$$

Denoting $\nabla_\alpha = \nabla + i(\alpha,0)$ and with some straightforward algebraic manipulation, it can be shown that U_α satisfies the above equations [10, p. 222–223].

To show the well-posedness of the variational formulation, one needs to show that the solution to the problem exists, that it is unique, and that it depends continuously on the data [22]. Since the variational form associated with U is easier to study analytically, the α-quasi-periodic problem is shown to be well-posed, before deriving the well-posedness of the periodic problem. It is therefore necessary to show the equivalence of these two variational formulations.

Let $v \in H_{\alpha\#}^1(\Omega)$; then Eq. (6) gives

$$\int_\Omega \nabla . \left(\frac{1}{k^2} \nabla U\right) \bar{v} + \int_\Omega U \bar{v} = 0.$$

Integrating by parts, using Eqs. (7) and (8) and denoting

$$a(U,v) = \left(\frac{1}{k^2} \nabla U, \nabla v\right)_\Omega - (U,v)_\Omega - \left(\frac{1}{k^2} T_\pm U, v\right)_{\Gamma_\pm}, \tag{27}$$

and

$$(f,v)_{\Gamma_+} = -\int_{\Gamma_+} \frac{2i\beta_1^0}{k_1^2} e^{i(\alpha x - \beta_1^0 B)} \bar{v}, \tag{28}$$

it can be shown that solving Eq. (6) is equivalent to the variational problem of finding $U \in H_{\alpha\#}^1(\Omega)$ for all $v \in H_{\alpha\#}^1(\Omega)$ such that

$$a(U,v) = (f,v)_{\Gamma_+}. \tag{29}$$

To establish an upper bound on the error arising when the scattering problem is solved numerically, the equivalence of the variational form for the periodic and α-quasi-periodic problems is required. For the periodic function U_α, let

$$a(U_\alpha, v_\alpha) = \left(\frac{1}{k^2} \nabla U_\alpha, \nabla v_\alpha\right)_\Omega - \left(\left(\frac{1-\alpha^2}{k^2}\right) U_\alpha, v_\alpha\right)_\Omega$$
$$- i\alpha \left(\frac{1}{k^2} \partial_x U_\alpha, v_\alpha\right)_\Omega + i\alpha \left(\frac{1}{k^2} U_\alpha, \partial_x v_\alpha\right)_\Omega$$
$$- \left(\frac{1}{k^2} T_\pm^\alpha U_\alpha, v_\alpha\right)_{\Gamma_\pm},$$
$$(f_\alpha, v_\alpha)_{\Gamma_+} = -\int_{\Gamma_+} \frac{2i\beta_1^0}{k_1^2} e^{-i\beta_1^0 B} \bar{v}_\alpha.$$

From Eq. (26), it can be shown that the corresponding variational problem is to find $U_\alpha \in H_\#^1(\Omega)$ for all $v_\alpha \in H_\#^1(\Omega)$ such that

$$a(U_\alpha, v_\alpha) = (f_\alpha, v_\alpha)_{\Gamma_+}. \tag{30}$$

Note that to ease the notation, we have noted a in the bilinear form associated to functions in both $H_\#^1(\Omega)$ and $H_{\alpha\#}^1(\Omega)$.

Lemma 9. *Finding $U_\alpha \in H_\#^1(\Omega)$ for all $v_\alpha \in H_\#^1(\Omega)$ such that $a(U_\alpha, v_\alpha) = (f_\alpha, v_\alpha)_{\Gamma_+}$ as given in Eq. (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 Eq. (29).*

Proof. Use Eq. (1); $f = e^{i\alpha x} f_\alpha$ and $v = e^{i\alpha x} v_\alpha$ in Eq. (30) to get Eq. (29).

Lemma 10. *Let $|k| > k_{\text{ref}} > 0$ such that $k_{\text{ref}} < |k_j|$. For all $v \in H_\#^1(\Omega)$, the solution $U_\alpha \in H_\#^1(\Omega)$ that satisfies equation (30) exists and is unique at all but a discrete set of frequencies, for each fixed horizontal wavenumber $k_j = n_\alpha$ [7]*

Proof. To start with, the sesquilinear form a in Eq. (27) is shown to be continuous. From the Cauchy–Schwarz inequality [15],

$$\left|\left(\frac{1}{k^2} \nabla U, \nabla v\right)_\Omega\right| \le \frac{1}{k_{\text{ref}}^2} \int_\Omega |\nabla U . \nabla \bar{v}| \mathrm{d}x\mathrm{d}y,$$
$$\le \frac{1}{k_{\text{ref}}^2} \|\nabla U\|_{L_{\alpha\#}^2(\Omega)} \|\nabla v\|_{L_{\alpha\#}^2(\Omega)} \tag{31}$$

and similarly it can be shown that

$$|(U,v)_\Omega| \le \|U\|_{L_{\alpha\#}^2(\Omega)} \|v\|_{L_{\alpha\#}^2(\Omega)}. \tag{32}$$

Since

$$\left|\int_{\Gamma_\pm} T_\pm U\bar{v}\mathrm{d}x\right|^2 \le C^2 d^2 (|k_j^2| \|U\|^2_{L^2_{\alpha\#}(\Omega)} + \|U\|^2_{H^1_{\alpha\#}(\Omega)}) \|v\|^2_{H^1_{\alpha\#}(\Omega)}, \tag{33}$$

using the trace theorem [15,18,23,24], it follows that

$$\left|\int_{\Gamma_\pm} \frac{1}{k^2} T_\pm U\bar{v}\mathrm{d}x\right| \le Cd\frac{1}{k^2_{\mathrm{ref}}} (|k_j^2| \|U\|^2_{L^2_{\alpha\#}(\Omega)} + \|U\|^2_{H^1_{\alpha\#}(\Omega)})^{1/2} \|v\|_{H^1_{\alpha\#}(\Omega)}. \tag{34}$$

Hence, Eq. (27) leads to

$$\begin{aligned} |a(U,v)| \le{}& \frac{1}{k^2_{\mathrm{ref}}} |U|_{H^1_{\alpha\#}(\Omega)} |v|_{H^1_{\alpha\#}(\Omega)} + \|U\|_{L^2_{\alpha\#}(\Omega)} \|v\|_{L^2_{\alpha\#}(\Omega)} \\ &+ Cd\frac{1}{k^2_{\mathrm{ref}}} (|k_j^2| \|U\|^2_{L^2_{\alpha\#}(\Omega)} + \|U\|^2_{H^1_{\alpha\#}(\Omega)}) \|v\|^2_{H^1_{\alpha\#}(\Omega)}, \end{aligned}$$

and so

$$|a(U,v)| \le C_0 \sup\left(1, \frac{1}{k^2_{\mathrm{ref}}}, \frac{|k^2|}{k^2_{\mathrm{ref}}}\right) \|U\|_{H^1_{\alpha\#}(\Omega)} \|v\|_{H^1_{\alpha\#}(\Omega)}.$$

Hence, $a(U,U)$ is continuous [15,18,25]. From Eq. (27),

$$\begin{aligned} |a(U,U)| + \int_\Omega |U|^2 &= \left|\int_\Omega \frac{1}{k^2} |\nabla U|^2 - \int_{\Gamma_\pm} \frac{1}{k^2} TU\bar{U}\right| \\ &\ge \left| |\int_\Omega \frac{1}{k^2} |\nabla U|^2| - |\int_{\Gamma_\pm} \frac{1}{k^2} TU\bar{U}| \right| \\ &\ge | \|k\|^{-2}_\infty - Cd\sup(|k_j|, 1)/k^2_{\mathrm{ref}} | \|U\|^2_{H^1_{\alpha\#}(\Omega)} \end{aligned}$$

using Eq. (10) and the equivalence of the norm in $H^l_{\alpha\#}(\Omega)$ for $l \ge 0$. Hence,

$$|a(U,U) + \|U\|^2_{L^2_{\alpha\#}(\Omega)}| \ge M_1 \|U\|^2_{H^1_{\alpha\#}(\Omega)}.$$

$a(U,U)$ is then $H^1_{\alpha\#}(\Omega)$ coercive and the existence of a solution can be shown from its uniqueness [26, p. 51]. Suppose that there are two solutions U_1 and U_2 and let $w = U_1 - U_2$. Taking the imaginary part and using Eqs. (5) and (27), it can be shown that $w = 0$, and so $U_1 = U_2$. From Lemma 9 and since $U_1 = e^{i\alpha x} U_{\alpha 1}$ and $U_2 = e^{i\alpha x} U_{\alpha 2}$, then $U_{\alpha 1} = U_{\alpha 2}$, 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. To this end an explicit dependency on the wavenumber k is derived in a regularity bound in the following theorem.

Theorem 11. *Let $f_\alpha \in H^l_\#(\Omega)$ be a general forcing function and let $U_\alpha \in H^l_\#(\Omega)$ be the solution of the inhomogeneous Helmholtz equation*

$$\begin{aligned} \nabla_\alpha \cdot \left(\frac{1}{k^2} \nabla_\alpha U_\alpha\right) + U_\alpha &= f_\alpha, && \text{in } \Omega, \\ \left(T^\alpha_+ - \frac{\partial}{\partial n}\right) U_\alpha &= 0, && \text{on } \Gamma_+, \\ \left(T^\alpha_- - \frac{\partial}{\partial n}\right) U_\alpha &= 0, && \text{on } \Gamma_-. \end{aligned} \tag{35}$$

Then there exists a constant C_{stab} that is dependent on the wavenumbers k_0 and k_3 such that

$$\|U_\alpha\|_{\mathcal{H}} \le C_{\mathrm{stab}} \|f_\alpha\|_{L^2_\#(\Omega)},$$

where $C_{\mathrm{stab}} = C_{\mathrm{reg}}(1 + C_s C(k_0, k_3))$ and $C_s C(k_0, k_3)$ is defined in Theorem 7.

Proof. Let $U_\alpha \in H^l_\#(\Omega)$ be the solution of Eq. (35); then from Definition 1 and Eq. (3),

$$\|U_\alpha\|^2_{\mathcal{H}} = |U_\alpha|^2_{H^1_\#(\Omega)} + \|k\|^2_\infty \|U_\alpha\|^2_{L^2_{\alpha\#}(\Omega)}.$$

From Eqs. (2) and (3),

$$\|U_\alpha\|^2_{\mathcal{H}} \le 2^2 \|U\|^2_{H^1_{\alpha\#}(\Omega)} + \|k\|^2_\infty \|U\|^2_{L^2_{\alpha\#}(\Omega)}.$$

Since $U = e^{i\alpha x} U_\alpha$, the proof is finished by using the regularity estimate of U as given in Theorem 7.

Hence, the problem given by Eq. (30) is well-posed since its solution exists, is unique (Lemma 10), and satisfies a regularity result (Theorem 11).

4. *A PRIORI* ERROR ESTIMATE FOR THE EXACT SOLUTION

To derive an *a priori* error estimate for the periodic solution U_α, the following three key results are needed:

- an estimate of the error arising from the discretization of the problem
- an estimate of the error arising from the truncation of the DtN operator
- an estimate of the total error

A. *A Priori* Error Estimate for the Discretized Problem

Let $X \subset H^l_{\alpha\#}(\Omega)$ be a finite element subspace of order p with $l \ge 1$, and let ζ_h be any regular partition of Ω [15,18,25,27]. Denote by h the maximum mesh size after partitioning Ω using ζ_h. The following standard assumption on the subspace 15] X] is used:

$$\begin{aligned} \inf_{\psi\in X} \Bigg\{ & \|v - \psi\|_{L^2_{\alpha\#}(\Omega)} + \frac{h}{p} \|\nabla v - \nabla\psi\|_{L^2_{\alpha\#}(\Omega)} \\ & + \left(\frac{h}{p}\right)^{\frac{1}{2}} \|v - \psi\|_{L^2_{\alpha\#}(\Gamma_\pm)} + \frac{h}{p} \|v - \psi\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_\pm)} \Bigg\} \le C\left(\frac{h}{p}\right)^l \|v\|_{H^l_{\alpha\#}(\Omega)}. \end{aligned} \tag{36}$$

Similarly, let X^α be a finite element subspace of order p of $H^l_\#(\Omega)$. The discretized problem corresponding to Eq. (29) is to find $U_h \in X$ such that

$$a(U_h, \phi) = (f, \phi)_{\Gamma_+} \tag{37}$$

with $a(U_h, \phi)$ and $(f, \phi)_{\Gamma_+}$ given by Eqs. (27) and (28) for all $\phi \in X$, where $T_\pm$ is given by Definition 3.

Lemma 12. *Let $U \in H^1_{\alpha\#}(\Omega)$; then for all $v \in H^1_{\alpha\#}(\Omega)$,*

$$|a(U, v)| \le C_c \|U\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$$

such that $C_c = (Cd + 1)/k^2_{\text{ref}}$ depends only on the period of the diffraction grating d and a lower bound on the wavenumber k_{ref}.

Proof. From the triangle inequality, putting Eqs. (31), (32), and (33) in Eq. (27) and using Definition 1,

$$|a(U, v)| \le \frac{1}{k^2_{\text{ref}}} |U|_{H^1_{\alpha\#}(\Omega)} |v|_{H^1_{\alpha\#}(\Omega)} + \|U\|_{L^2_{\alpha\#}(\Omega)} \|v\|_{L^2_{\alpha\#}(\Omega)} + \frac{Cd}{k^2_{\text{ref}}} \|U\|_{\mathcal{H}} \|v\|_{\mathcal{H}}.$$

Since $(\|k^2\|_\infty / k^2_{\text{ref}}) \ge 1$,

$$|a(U, v)| \le \frac{1}{k^2_{\text{ref}}} (|U|_{H^1_{\alpha\#}(\Omega)} |v|_{H^1_{\alpha\#}(\Omega)} + \|k^2\|_\infty \|U\|_{L^2_{\alpha\#}(\Omega)} \|v\|_{L^2_{\alpha\#}(\Omega)} + Cd \|U\|_{\mathcal{H}} \|v\|_{\mathcal{H}}).$$

Noting that

$$|U|_{H^1_{\alpha\#}(\Omega)} |v|_{H^1_{\alpha\#}(\Omega)} + \|k\|^2_\infty \|U\|_{L^2_{\alpha\#}(\Omega)} \|v\|_{L^2_{\alpha\#}(\Omega)} \le \|U\|_{\mathcal{H}} \|v\|_{\mathcal{H}} \tag{38}$$

using Definition 1, we have

$$|a(U, v)| \le C_c \|U\|_{\mathcal{H}} \|v\|_{\mathcal{H}},$$

where $C_c = (1/k^2_{\text{ref}})(Cd + 1)$.

Lemma 13. *For $U \in H^1_{\alpha\#}(\Omega)$,*

$$\frac{1}{\|\Re(k^2)\|_\infty} |U|^2_{H^1_{\alpha\#}(\Omega)} - \|U\|^2_{L^2_{\alpha\#}(\Omega)} \le |a(U, U)| + \left| \left(\frac{1}{k^2} T_\pm U, U \right)_{\Gamma_\pm} \right|$$

such that $a(u, v)$ is given by Eq. (27).

Proof. From Eq. (27),

$$\left(\frac{1}{k^2} \nabla U, \nabla U \right)_\Omega - (U, U)_\Omega = a(U, U) + \left(\frac{1}{k^2} T_\pm U, U \right)_{\Gamma_\pm}.$$

Since $|\Re(c)| \le |c|$ for any $c \in \mathbb{C}$,

$$\left| \Re \left(\frac{1}{k^2} \nabla U, \nabla U \right)_\Omega - (U, U)_\Omega \right| \le \left| a(U, U) + \left(\frac{1}{k^2} T_\pm U, U \right)_{\Gamma_\pm} \right|.$$

By noting that $|b - c| \ge |b| - |c|$ and with the triangle inequality, it can be shown that

$$\left| \Re \left(\frac{1}{k^2} \nabla U, \nabla U \right)_\Omega \right| - (U, U)_\Omega \le |a(U, U)| + \left| \left(\frac{1}{k^2} T_\pm U, U \right)_{\Gamma_\pm} \right|.$$

The proof is finished by noting that

$$\left| \Re \left(\frac{1}{k^2} \nabla U, \nabla U \right)_\Omega \right| \ge \frac{1}{\|\Re(k^2)\|_\infty} |U|^2_{H^1_{\alpha\#}(\Omega)}.$$

The following lemma is needed to express the norm of the error in $L^2_{\alpha\#}$ in terms of the norm of the error in $\mathcal{H}$.

Lemma 14. *Let $U \in H^1_{\alpha\#}(\Omega)$ be the solution of Eq. (29), and let U_h be the corresponding discretized solution of Eq. (37). By denoting $e_h = U - U_h$, there exists a constant $C_1 = CC_{\text{stab}}(h/p)(\|k\|_\infty / k^2_{\text{ref}})(Cd + 1)$, where $C_{\text{stab}} = (1 + C_s C(k_0, k_3)) C_{\text{reg}}$ is as defined in Theorem 7, such that*

$$\|e_h\|_{L^2_{\alpha\#}(\Omega)} \le C_1 \|e_h\|_{\mathcal{H}}.$$

Proof. Let w to be the dual solution of

$$\nabla . \left(\frac{1}{k^2} \nabla w \right) + w = \phi \qquad (x, y) \in \Omega,$$
$$(T^*_\pm - \partial_n w) = 0 \qquad \text{on } \Gamma_\pm, \tag{39}$$

for all ϕ, $w \in H^1_{\alpha\#}(\Omega)$, where $T^*_\pm$ are the dual operators of $T_\pm$ [28, p. 476]. Using the duality argument [18, p. 137],

$$\|e_h\|_{L^2_{\alpha\#}(\Omega)} = \sup_{\phi \in C_\infty(\Omega)} \frac{|a(e_h, w - \psi)|}{\|\phi\|_{L^2_{\alpha\#}(\Omega)}} \tag{40}$$

such that $\psi \in X$, and so from Eq. (27),

$$\begin{aligned} |a(e_h, w - \psi)| &= \left| \left(\frac{1}{k^2} \nabla e_h, \nabla (w - \psi) \right)_\Omega - (e_h, w - \psi)_\Omega - \left(\frac{1}{k^2} T_\pm e_h, w - \psi \right)_{\Gamma_\pm} \right| \\ &\le \frac{1}{k^2_{\text{ref}}} (|e_h|_{H^1_{\alpha\#}(\Omega)} |w - \psi|_{H^1_{\alpha\#}(\Omega)} + \|k\|^2_\infty \|e_h\|_{L^2_{\alpha\#}(\Omega)} \|w - \psi\|_{L^2_{\alpha\#}(\Omega)} + Cd \|e_h\|_{\mathcal{H}} \|w - \psi\|_{H^1_{\alpha\#}(\Omega)}) \end{aligned}$$

using Eq. (33) with Cauchy's inequality [18, p. 50]. Hence, Eq. (38) gives

$$|a(e_h, w - \psi)| \le (Cd + 1) \frac{1}{k^2_{\text{ref}}} \|e_h\|_{\mathcal{H}} \|w - \psi\|_{H^1_{\alpha\#}(\Omega)}.$$

From the standard approximation estimate in a finite element space given by Eq. (36), $(h/p) \|w - \psi\|_{H^1_{\alpha\#}(\Omega)} \le C(h/p)^2 \|w\|_{H^2_{\alpha\#}(\Omega)}$, and so

$$|a(e_h, w - \psi)| \le C(Cd + 1) \frac{1}{k^2_{\text{ref}}} \frac{h}{p} \|e_h\|_{\mathcal{H}} \|w\|_{H^2_{\alpha\#}(\Omega)}. \tag{41}$$

From Theorem 7 the regularity estimate is derived as follows from Eq. (39):

$$\frac{\|w\|_{H^2_{\alpha\#}(\Omega)}}{\|\phi\|_{L^2_{\alpha\#}(\Omega)}} \leq (1 + C_s C(k_0, k_3))C_{\mathrm{reg}}\|k\|_\infty = C_{\mathrm{stab}}\|k\|_\infty.$$

Hence, from Eqs. (40) and (41),

$$\|e_h\|_{L^2_{\alpha\#}(\Omega)} \leq CC_{\mathrm{stab}}\frac{h\|k\|_\infty}{pk^2_{\mathrm{ref}}}(Cd + 1)\|e_h\|_{\mathcal{H}} = C_1\|e_h\|_{\mathcal{H}}.$$

The previous three lemmas are used to derive the following *a priori* error estimate for the periodic solution U_α.

Theorem 15. *Let the wavenumber $|k| \geq k_{\mathrm{ref}} > 0$, let the maximum mesh size $h \in [0, h_0]$, and let the order of the polynomial basis $p \in [p_0, \infty]$ such that $k(h_0/p_0) < 1$, and $C_4 = 1–2\|k\|_\infty C_1 > 0$ with C_1 as given in Lemma 14. Let U_α be the continuous solution of Eq. (35); then the corresponding discretized solution, $U_{\alpha_h} \in X^\alpha$, exists and is unique. If $e_{\alpha_h} = U_\alpha - U_{\alpha_h}$, then*

$$\|e_{\alpha_h}\|_{\mathcal{H}} \leq 4\frac{c_k}{C_4}(2Cd + 1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}},$$

and

$$\|e_{\alpha_h}\|_{L^2_\#(\Omega)} \leq 2\frac{c_k}{C_4}C_1(2Cd + 1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}},$$

for all test functions $\psi_\alpha \in X^\alpha$, where $C \in \mathbb{R}$, $c_k = \|\Re(k^2)\|_\infty/k^2_{\mathrm{ref}}$, and d is the period of the grating.

Proof. Let $e_h = U - U_h$ and $\psi = e^{i\alpha x}\psi_\alpha$. From Lemma 13 and by using Galerkin orthogonality [18, p. 58],

$$\frac{1}{\|\Re(k^2)\|_\infty}(|e_h|^2_{H^1_{\alpha\#}(\Omega)} - \|\Re(k^2)\|_\infty\|e_h\|^2_{L^2_{\alpha\#}(\Omega)}) \leq |a(e_h, U - \psi)| + \left|\left(\frac{1}{k^2}T_\pm e_h, e_h\right)_{\Gamma_\pm}\right|.$$

Since $\|k\|^2_\infty \geq \|\Re(k^2)\|_\infty$, Lemma 12 and Eq. (34) lead to

$$\frac{1}{\|\Re(k^2)\|_\infty}(|e_h|^2_{H^1_{\alpha\#}(\Omega)} - \|k\|^2_\infty\|e_h\|^2_{L^2_{\alpha\#}(\Omega)}) \leq C_c\|e_h\|_{\mathcal{H}}\|U - \psi\|_{\mathcal{H}} + \frac{Cd}{k^2_{\mathrm{ref}}}\|e_h\|_{\mathcal{H}}\|U - \psi\|_{\mathcal{H}}$$

using Céa's theorem [18, p. 64]. Let $C_c = (Cd + 1)/k^2_{\mathrm{ref}}$ as given in Lemma 12, and so

$$\frac{1}{\|\Re(k^2)\|_\infty}(|e_h|^2_{H^1_{\alpha\#}(\Omega)} - \|k\|^2_\infty\|e_h\|^2_{L^2_{\alpha\#}(\Omega)}) \leq \frac{2Cd + 1}{k^2_{\mathrm{ref}}}\|e_h\|_{\mathcal{H}}\|U - \psi\|_{\mathcal{H}}.$$

By letting $c_k = \|\Re(k^2)\|_\infty/k^2_{\mathrm{ref}}$, and noting that $\|k\|_\infty\|e_h\|_{L^2_{\alpha\#}(\Omega)} \leq \|e_h\|_{\mathcal{H}}$, we have

$$|e_h|^2_{H^1_{\alpha\#}(\Omega)} - \|k\|_\infty\|e_h\|_{L^2_{\alpha\#}(\Omega)}\|e_h\|_{\mathcal{H}} \leq c_k(2Cd + 1)\|e_h\|_{\mathcal{H}}\|U - \psi\|_{\mathcal{H}}.$$

Using Lemma 14 and Definition 1,

$$\|e_h\|_{\mathcal{H}} - 2\|k\|_\infty C_1\|e_h\|_{\mathcal{H}} \leq c_k(2Cd + 1)\|U - \psi\|_{\mathcal{H}}. \qquad (42)$$

Suppose that $2\|k\|_\infty C_1 < 1$ and so $C_4 = 1 - 2\|k\|_\infty C_1 > 0$:

$$\|e_h\|_{\mathcal{H}} \leq \frac{c_k}{C_4}(2Cd + 1)\|U - \psi\|_{\mathcal{H}}. \qquad (43)$$

From Theorem 2 and Definition 1, $\|e_{\alpha_h}\|_{\mathcal{H}} \leq 2\|e_h\|_{\mathcal{H}}$, and $\|U - \psi\|_{\mathcal{H}} \leq 2\|U_\alpha - \psi_\alpha\|_{\mathcal{H}}$, and then equation (43) is used to get

$$\|e_{\alpha_h}\|_{\mathcal{H}} \leq 4\frac{c_k}{C_4}(2Cd + 1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}}.$$

For the second result, Lemma 14 and Eq. (43) are used to get

$$\|e_h\|_{L^2_{\alpha\#}(\Omega)} \leq \frac{c_k}{C_4}C_1(2Cd + 1)\|U - \psi\|_{\mathcal{H}}.$$

From Theorem 2 and from Definition 1,

$$\|e_{\alpha_h}\|_{L^2_\#(\Omega)} \leq 2\frac{c_k}{C_4}C_1(2Cd + 1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}}.$$

By assuming that there exists two solutions and letting h/p tend to zero, it can be shown that U_{α_h} is unique.

B. *A Priori* Error Estimate of the Continuous Problem Arising from Truncating the DtN Operator

For computational purposes, the infinite sum inside the DtN map that is used as transparent boundary conditions must be truncated. Let $M \in \mathbb{N}$ and $M < \infty$; then from Definition 9, $T^\alpha_\pm$ is approximated by $T^{\alpha^M}_\pm$, which is given by

$$T^{\alpha^M}_\pm U_{\alpha_h}(x) = \sum_{n=-M}^{M} i\beta^n_j U^{(n)}_{\alpha h}(\pm B)e^{i\frac{2\pi n}{d}x}.$$

Let U^M_α be the approximated solution of the continuous problem where $T^{\alpha^M}_\pm$ is used instead of $T^\alpha_\pm$ in Eq. (35). Then, the error estimate by truncating $T^\alpha_\pm$ is given in the following theorem.

Theorem 16. *Let $M \in \mathbb{N}$ such that $M > M_0 = |k| + |\alpha|$ and $2\pi|n|/d > M$. Let $\|k\|_\infty \geq k_{\mathrm{ref}}$ and denote by*

$$e^M_\alpha = U_\alpha - U^M_\alpha.$$

If $2\|k\|_\infty C_1 < 1$, such that C_1 is as given in Lemma 14, and C is as defined in Lemma 6, then U^M_α exists and is unique. In addition, if $C_4 = 1 - 2\|k\|_\infty C_1 > 0$, then

$$\|e^M_\alpha\|_{\mathcal{H}} \leq 4c_k\frac{d}{C_4}\left(C\|U_\alpha - \psi_\alpha\|_{\mathcal{H}} + e^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2-k_j^2}}\|U_\alpha\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_{1,\pm})}\right),$$

$$\|e^M_\alpha\|_{L^2_\#(\Omega)} \leq 2c_k d\frac{C_1}{C_4}\left(C\|U_\alpha - \psi_\alpha\|_{\mathcal{H}} + e^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2-k_j^2}}\|U_\alpha\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_{1,\pm})}\right),$$

for all test functions $\psi_\alpha \in X^\alpha$ *with* $c_k = (\|\Re(k^2)\|_\infty / k_{\mathrm{ref}}^2)$, $c_{\min} = \inf_{|n|>(Md/2\pi)} \sin(z_n/2)$, z_n *given by Eq. (4), and* $\Gamma_{1,\pm} = \{(x, \pm b) \in \Omega\}$.

Proof. Let $U \in H^1_{\alpha\#}(\Omega)$ satisfy Eq. (29) for all $v \in H^1_{\alpha\#}(\Omega)$. Approximate the continuous problem by finding $U^M \in H^1_{\alpha\#}(\Omega)$ such that

$$a^M(U^M, v) = (f, v)_{\Gamma_+} \tag{44}$$

with

$$a^M(U^M, v) = \left(\frac{1}{k^2}\nabla U^M, \nabla v\right)_\Omega - (U^M, v)_\Omega - \left(\frac{1}{k^2}T^M_\pm U^M, v\right)_{\Gamma_\pm},$$

$$(f, v)_{\Gamma_+} = \left(-\frac{2i\beta_1^0}{k_1^2}e^{i(\alpha x - \beta_1^0 B)}, v\right)_{\Gamma_+},$$

and $T^M_\pm v = \sum_{n=-M}^{M} i\beta_j^n v^{(n_\alpha)}(\pm B)e^{in_\alpha x}$, for all $v \in \Omega$. From Eqs. (29) and (44), note that

$$a(U, v) - a^M(U^M, v) = 0. \tag{45}$$

By letting $e^M = U - U^M$ and noting that $T_\pm = T^M_\pm + (T_\pm - T^M_\pm)$, Eq. (45) leads to

$$\left(\frac{1}{k^2}\nabla e^M, \nabla v\right)_\Omega - (e^M, v)_\Omega - \left(\frac{1}{k^2}T^M_\pm e^M, v\right)_{\Gamma_\pm} = \left(\frac{1}{k^2}(T_\pm - T^M_\pm)U, v\right)_{\Gamma_\pm}$$

and so

$$a^M(e^M, v) = \left(\frac{1}{k^2}(T_\pm - T^M_\pm)U, v\right)_{\Gamma_\pm}. \tag{46}$$

Let M such that $M > |k| + |\alpha|$ so that $n_\alpha^2 > k^2$ for $|n| \geq (Md/2\pi)$; then it can be shown that [10, p. 256]

$$|((T_\pm - T^M_\pm)U, v)_{\Gamma_\pm}| \leq d e^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2 - k_j^2}} \|U\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_{1,\pm})} \|v\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_\pm)}$$

with $c_{\min} = \inf_{|n|>(Md/2\pi)} \sin(z_n/2)$. Since truncating $T_\pm$ does not affect the validity of Lemma 13,

$$\frac{1}{\|\Re(k^2)\|_\infty}(|e^M|^2_{H^1_{\alpha\#}(\Omega)} - \|k\|^2_\infty \|e^M\|^2_{L^2_{\alpha\#}(\Omega)}) \leq \left|\left(\frac{1}{k^2}T^M_\pm e^M, e^M\right)_{\Gamma_\pm}\right| + |a^M(e^M, U - \psi)|$$

by using Galerkin orthogonality [18, p. 58], where $\psi = e^{i\alpha x}\psi_\alpha$ with $\psi_\alpha \in X^\alpha$. By noting that $|(T^M_\pm v, v)_{\Gamma_\pm}| \leq |(T_\pm v, v)_{\Gamma_\pm}|$, and using Eqs. (34) and (46) with U_h^M minimizing a^M, the last inequality is derived using Céa's lemma [18, p. 64] and the trace theorem [15,18,23,24] with the equivalence of the norms in $\mathcal{H}$ and in $H^1_{\alpha\#}(\Omega)$. The duality argument [18, p. 137] can be used to approximate $\|\cdot\|_{L^2_{\alpha\#}(\Omega)}$, with the dual problem given by Eq. (39) and Lemma 14 to write

$$\frac{1}{\|\Re(k^2)\|_\infty}(\|e^M\|_{\mathcal{H}} - 2\|k\|_\infty C_1\|e^M\|_{\mathcal{H}}) \leq \frac{Cd}{k_{\mathrm{ref}}^2}\|U - \psi\|_{\mathcal{H}} + \frac{d}{k_{\mathrm{ref}}^2}e^{-(B-b)\sin(z_n/2)\sqrt{(M-|\alpha|)^2 - k_j^2}}\|U\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_{1,\pm})}.$$

Letting $c_k = \|\Re(k^2)\|_\infty / k_{\mathrm{ref}}^2$ and supposing $2\|k\|_\infty C_1 < 1$, then $C_4 = 1 - 2\|k\|_\infty C_1 > 0$ and

$$\|e^M\|_{\mathcal{H}} \leq \frac{c_k}{C_4}d\left(C\|U - \psi\|_{\mathcal{H}} + e^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2 - k_j^2}}\|U\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_{1,\pm})}\right)$$

with $c_{\min} = \inf_{|n|>(Md/2\pi)} \sin(z_n/2)$. From Definition 1 and Theorem 2,

$$\|e^M_\alpha\|_{\mathcal{H}} \leq 4\frac{c_k}{C_4}d\left(C\|U_\alpha - \psi_\alpha\|_{\mathcal{H}} + e^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2 - k_j^2}}\|U_\alpha\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_{1,\pm})}\right). \tag{47}$$

Lemma 14, Eq. (47), and Definition 1 together with Theorem 2 lead to

$$\|e^M_\alpha\|_{L^2_{\alpha\#}(\Omega)} \leq 2\frac{d}{C_4}c_k C_1\left(C\|U_\alpha - \psi_\alpha\|_{\mathcal{H}} + e^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2 - k_j^2}}\|U_\alpha\|_{H^{\frac{1}{2}}_{\alpha\#}(\Gamma_{1,\pm})}\right).$$

It can be shown that U^M_α is unique by considering M tending to infinity.

C. Estimate of the Total Error

By denoting the total error by $e_\alpha = U_\alpha - U^M_{\alpha_h}$, it can be estimated as follows.

Theorem 17. *Let* $|k| \geq k_{\mathrm{ref}} > 0$, *the maximum mesh size* $h \in [0, h_0]$, *and the degree of the polynomial basis* $p \in [p_0, \infty]$ *be such that* $kh_0/p_0 < 1$ *with* $2C_1\|k\|_\infty < 1$, *where* C_1 *is defined in Lemma 14 so that* $C_4 = 1 - 2C_1\|k\|_\infty > 0$. *Let* $M \in \mathbb{N}$ *be such that* $M \geq M_0$, *let* U_α *be the continuous solution of Eq. (35), let* $U^M_{\alpha_h}$ *be the corresponding discretized solution with the truncated DtN operator, and let the total error be* $e_\alpha = U_\alpha - U^M_{\alpha_h}$. *Then the total error satisfies*

$$\|e_\alpha\|_{\mathcal{H}} \leq 4\frac{c_k}{C_4}(3Cd + 1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}} + 4\frac{c_k}{C_4}de^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2 - k_j^2}}\|U_\alpha\|_{H^{\frac{1}{2}}_{\#}(\Gamma_{1,\pm})},$$

and

$$\|e_\alpha\|_{L^2_{\#}(\Omega)} \leq 2\frac{c_k}{C_4}C_1(3Cd + 1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}} + 2d\frac{c_k}{C_4}C_1 e^{-(B-b)c_{\min}\sqrt{(M-|\alpha|)^2 - k_j^2}}\|U_\alpha\|_{H^{\frac{1}{2}}_{\#}(\Gamma_{1,\pm})},$$

for all $\psi_\alpha \in X^\alpha$, *where* $c_k = \|k\|_\infty / k_{\mathrm{ref}}^2$, C_c *is given in Lemma 12,* C *is as in Lemma 6, and* $c_{\min} = \inf_{|n|>(Md/2\pi)} \sin(z_n/2)$ *with* z_n *as defined in Eq. (4).*

Proof. Note that $\|e_\alpha\|_{\mathcal{H}} \le \|U_\alpha - U_\alpha^M\|_{\mathcal{H}} + \|U_\alpha^M - U_{\alpha_h}^M\|_{\mathcal{H}}$, where $\|U_\alpha - U_\alpha^M\|_{\mathcal{H}}$ has already been shown in Theorem 16. An *a priori* error estimate for the second term can be derived in a similar way to that performed in Theorem 16. By denoting $e_h^M = U^M - U_h^M$, where $U^M = e^{i\alpha x}U_\alpha^M$ and $U_h^M = e^{i\alpha x}U_{\alpha_h}^M$, it can be shown similarly as in Lemma 13 that

$$\frac{1}{\|\Re(k^2)\|_\infty}|e_h^M|^2_{H^1_{\alpha\#}(\Omega)} - \|e_h^M\|^2_{L^2_{\alpha\#}(\Omega)} \le |a^M(e_h^M, e_h^M)| + \left|\left(\frac{1}{k^2}T_\pm e_h^M, e_h^M\right)_{\Gamma_\pm}\right|.$$

Hence, from Definition 1, and letting $|k| \ge |k_{\mathrm{ref}} > 0$,

$$\frac{1}{\|\Re(k^2)\|_\infty}\|e_h^M\|^2_{\mathcal{H}} - 2\|k\|^2_\infty\|e_h^M\|^2_{L^2_{\alpha\#}(\Omega)} \le |a^M(e_h^M, e_h^M)| + \frac{1}{k^2_{\mathrm{ref}}}|(T_\pm^M e_h^M, e_h^M)|.$$

Similarly to Eq. (42), it can be shown that

$$\|e_h^M\|_{\mathcal{H}} - 2C_1\|k\|_\infty\|e_h^M\|_{\mathcal{H}} \le c_k(2Cd+1)\|U^M - \psi\|_{\mathcal{H}},$$

where $\psi = e^{i\alpha x}\psi_\alpha$ such that $\psi_\alpha \in X^\alpha$. Since $2C_1\|k\|_\infty < 1$, $C_4 = 1 - 2C_1\|k\|_\infty > 0$ and so

$$\|e_h^M\|_{\mathcal{H}} \le \frac{c_k}{C_4}(2Cd+1)\|U^M - \psi\|_{\mathcal{H}}.$$

For $M \ge M_0$, U^M tends to U; therefore

$$\|e_h^M\|_{\mathcal{H}} \le \frac{c_k}{C_4}(2Cd+1)\|U - \psi\|_{\mathcal{H}}.$$

From Definition 1,

$$\|e_{\alpha_h}^M\|_{\mathcal{H}} \le 4\frac{c_k}{C_4}(2Cd+1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}}. \quad (48)$$

From Lemma 14 and Theorem 2,

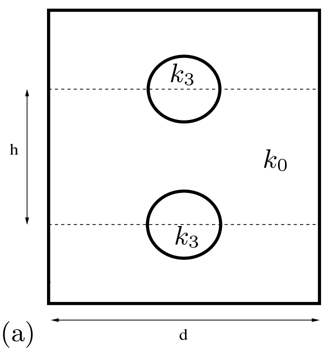
$$\|e_{\alpha_h}^M\|_{L^2_\#(\Omega)} \le 2\frac{c_k}{C_4}C_1(2Cd+1)\|U_\alpha - \psi_\alpha\|_{\mathcal{H}}. \quad (49)$$

Theorem 16 together with Eqs. (48) and (49) is used to finish the proof.

5. NUMERICAL RESULTS

The α-quasi-periodic method can be applied straightforwardly to any geometry, and its implementation is independent of the number of scatterers inside the scattered region. In this section, a grating composed of two dielectric transmitting cylinders is considered as shown in Fig. 2. The reflection efficiency of order zero (R_0) (the ratio of the reflected field to the incident field [11, p. 65]), for the TM case (Case 2B), where the wavelength-spatial periodicity ratio λ/d varies from 0.7 to 1, is computed numerically. These structures are of interest in the field of 2D photonic bandgap structures that are used as tunable filters. A cylindrical harmonic expansion approach, called the lattice sum technique, has been used to study this problem previously [29]. However, the lattice sum technique is limited to scattering with polygonal geometry and is dependent on the number of cylinders inside the scattered region. This limitation does not apply to the finite element method presented in this paper. The α-quasi-periodic method is used below to solve the problem and a comparison with the lattice sum technique is presented. To illustrate the accuracy of the α-quasi-periodic method, a polynomial basis of degree 4 with 21 633 degrees of freedom is used and R_0 is plotted as a function of λ/d in Fig. 2(b). It can be concluded from this figure that the numerical results from the α-quasi-periodic method are in good agreement with those from the lattice sum technique. The deployment of the α-quasi-periodic method to more complex geometries is the subject of ongoing work.

In the following example, the transmitting dielectric lamellar grating as shown in Fig. 3 is considered. This type of grating is used in modeling multiscale phenomena grating problems, and has been studied in [9] using a hybrid approach that combines a perfectly matching layer technique and an adaptive finite element method. Here the wavenumbers are fixed as $k_1 = 2\pi$ and $k_2 = (0.22 + 6.71i)2\pi$, the angle of incidence is $\theta = \pi/6$, and the period is $d = 1$. The α-quasi-periodic method is once again used below to solve the problem with uniform mesh using polynomial degree 6, and a comparison with [9] is presented. To provide a basis for a relative error, the global method in [9] is used with 201205 *dof*, which gives $R_0^{\mathrm{Adapt}} = 0.8484815$. The relative error using the adaptive finite



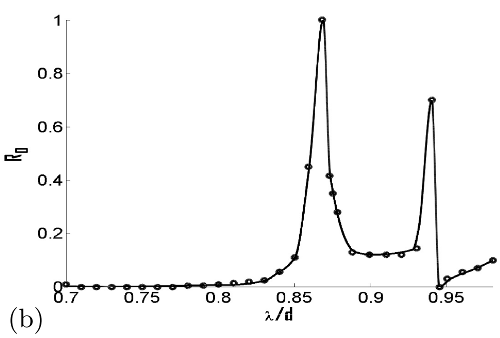


Fig. 2. (a) Double layered dielectric transmitting cylinders. (b) Comparison between the reflection efficiency of order 0 from the α-quasi-periodic method (full line) and the lattice sum technique (dots) [29] for dielectric transmitting cylinders for the TM case [see (a)]. The reflection efficiency is shown as a function of the ratio of the wavelength of the incident field (λ) to the lattice period (d).

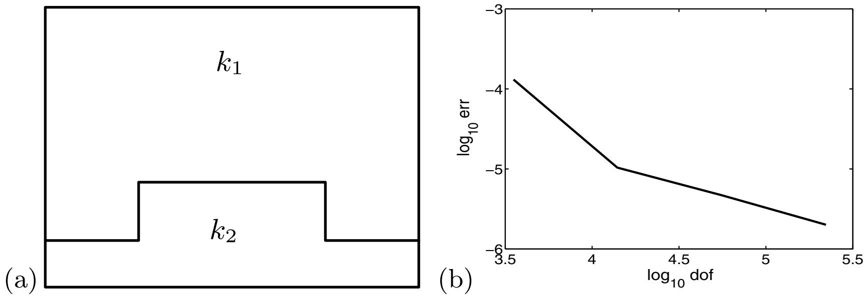


Fig. 3. (a) Transmitting dielectric lamellar grating. (b) Error in computing R_0 using the α-quasi-periodic method with respect to the adaptive method in [9].

element method in [9] and our method can be compared by defining err $= |R_0(\alpha - \text{quasiperiodic}) - R_0^{\text{Adapt}}|/R_0^{\text{Adapt}}$.

6. CONCLUSION

Diffraction gratings have been used, for example, in crystalline silicon solar cells [1], gas sensors [2], high-intensity color displays [30], and medical x-ray imaging [3,4]. 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 [31]. 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. In this paper, a rigorous *a priori* error analysis for the α-quasi-periodic transformation method has been derived.

To start with, the physical and mathematical aspects of the problem of diffraction when an electromagnetic wave interacts with a periodic grating have been described. From Maxwell's equations, it can be shown that the problem can be decomposed into two elementary mathematical problems, which are TM and the TE Helmholtz problems. For each problem, the grating can be perfectly conducting or transmitting, and so there are four cases. To keep this paper at a reasonable length, the results presented here were restricted to Case 2B (TM case for the transmitting dielectric grating). The domain was truncated with respect to the y direction, and appropriate DtN maps were introduced. The boundary value problem corresponding to the truncated domain was formulated, where the incident wave was included via the boundary conditions. An equivalent but alternative formulation that incorporated the incident wave via an inhomogeneous forcing term (with compact support) was considered so that a regularity result could be derived. This regularity result showed an explicit dependence on the wavenumber k and the forcing term f, and it was then used to prove the well-posedness of the variational formulation. It also gave a hold on the convergence and the stability of the solution when the scattering problem was approximated using finite elements. In fact, if h denotes the maximum mesh size of the elements, and p the highest order of the finite element basis, since the dependence of the regularity results on the wavenumber k is known explicitly, the *a priori* error estimate presented a power factor of kh/p. Hence, the mesh size h and the order of the polynomial basis p for a given wavenumber k can be chosen to balance the computational time and the accuracy of the approximate solution.

The α-quasi-periodic method is used to solve numerically the problem of two dielectric transmitting cylinders studied in [29], using the lattice sum technique and the lamellar grating studied in [9] using the adaptive finite element technique. The application is straightforward regardless of the shape and the number of scatterers inside the scattered region as opposed to the lattice sum technique. Good agreement of the numerical results with those in [29] and in [9], using the α-quasi-periodic method, is obtained.

REFERENCES

1. M. Peters, M. Rüdiger, B. Bläsi, and W. Platzer, "Electro-optical simulation of diffraction in solar cells," Opt. Express **18**, A584–A593 (2010).
2. R. A. Potyrailo, H. Ghiradella, A. Vertiatchikh, J. R. Cournoyer, K. Dovidenko, and E. Olson, "Morpho butterfly wing scales demonstrate highly selective vapour response," Nat. Photonics **1**, 123–128 (2007).
3. 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**, 3317–3323 (2010).
4. 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," Proc. Natl. Acad. Sci. USA **107**, 13576–13581 (2010).
5. G. Bao, L. Cowsar, and W. Masters, *Mathematical Modeling in Optical Science*, Vol. **22** of SIAM Frontiers in Applied Mathematics (Springer, 2001).
6. G. Bao and A. Zhou, "Analysis of finite dimensional approximations to a class of partial differential equations," Math. Methods Appl. Sci. **27**, 2055–2066 (2004).
7. R. Petit, *Electromagnetic Theory of Gratings* (Springer, 1980).
8. G. Bao, "Numerical analysis of diffraction by periodic structures: TM polarization," Numer. Math. **75**, 1–16 (1996).
9. G. Bao, Z. Chen, and H. Wu, "Adaptive finite element method for diffraction gratings," J. Opt. Soc. Am. A **22**, 1106–1114 (2005).
10. N. Lord, "Analysis of electromagnetic waves in a periodic diffraction grating using a priori error estimates and a dual weighted residual method," Ph.D. thesis (University of Strathclyde, 2012) (http://www.mathstat.strath.ac.uk/research/phd_mphil_theses).
11. N. Lord and A. J. Mulholland, "Analysis of the α,0-quasi periodic transformation for a periodic diffraction grating," Research Report, No. 19 (Department of Mathematics and Statistics, University of Strathclyde, 2011) (http://www.mathstat.strath.ac.uk/research/reports/2011).
12. S. Chandler-Wilde, "Boundary value problems for the Helmholtz equation in a half-plane," in *Mathematical and Numerical Aspects of Wave Propagation* (SIAM, 1995), pp. 188–197.
13. G. Bao, Y. Cao, and H. Yang, "Numerical solution of diffraction problems by a least squares FEM," Math. Methods Appl. Sci. **23**, 1073–1092 (2000).
14. E. Wolf, ed., *Rigorous Vector Theories of Diffraction Gratings*, Vol. **21** of Progress in Optics (North-Holland, 1984), pp. 1–67.
15. P. G. Ciarlet, *The Finite Element Method for Elliptic Equations* (North-Holland, 1978).

16. J. M. Melenk and S. Sauter, "Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions," Math. Comput. **79**, 1871–1914 (2010).
17. A. Lechleiter and D. Nguyen, "Volume integral equations for scattering from anisotropic diffraction gratings," Math. Methods Appl. Sci. **36**, 262–274 (2012).
18. S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Vol. **15** of Texts in Applied Mathematics (Springer, 2002).
19. E. Kreyszig, *Introductory Functional Analysis with Applications* (Wiley, 1989).
20. D. Maystre, *Electromagnetic Theory of Gratings* (Springer, 1980).
21. L. F. Richardson, *Measure and Integration. A Concise Introduction to Real Analysis* (Wiley, 2009).
22. D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Vol. **93** of Applied Mathematical Sciences (Springer-Verlag, 1992).
23. A. Buffa, "Trace theorems for functional spaces related to Maxwell equations: an overview, ," in *Computational Electromagnetics*, Vol. **28** of Lecture Notes in Computational Science and Engineering (Springer, 2002), pp. 23–34.
24. Z. Ding, "A proof of the trace theorem of Sobolev spaces on Lipschitz domains," Proc. Am. Math. Soc. **124**, 591–601 (1996).
25. D. Braess, *Finite Elements* (Cambridge University, 1997).
26. F. Ihlenburgh, *Finite Element Analysis of Acoustic Scattering* (Springer, 1998), Vol. **132**.
27. J. T. Oden and M. Ainsworth, *A Posteriori Error Estimation in Finite Element Analysis* (Wiley, 2000).
28. K. Ito, *Encyclopedic Dictionary of Mathematics*, 2nd ed. (Massachusetts Institute of Technology, 1987).
29. K. Yasumoto, T. Kushta, and H. Toyama, "Accurate analysis of two-dimensional electromagnetic scattering from multilayered periodic arrays of circular cylinders using lattice sums technique," IEEE Trans. Antennas Propag. **52**, 2603–2611 (2004).
30. R. E. Coath, "Investigating the use of replica Morpho butterfly scales for colour displays," Society **5**, 1–9 (2007) (http://printfu.org/blue+morpho+didius+butterfly).
31. G. Berger, K. Müller, C. Denz, I. Földvári, and A. Péter, "Digital data storage in a phase-encoded holographic memory system: data quality and security," Proc. SPIE **4988**, 104–111 (2003).