COMMUN. MATH. SCI.
Vol. 23, No. 7, pp. 1881–1907

FINITE ELEMENT AND INTEGRAL EQUATION METHODS TO CONICAL DIFFRACTION BY IMPERFECTLY CONDUCTING GRATINGS*

GUANGHUI HU†, JIAYI ZHANG‡, AND LINLIN ZHU§

Abstract. This paper is concerned with the variational and integral equation methods for a conical diffraction problem (which is also called oblique incidence) for imperfectly conducting gratings modeled by the impedance boundary value problem of the Helmholtz equation in periodic structures. We prove new mapping properties of the Dirichlet-to-Neumann map, justify the strong ellipticity of the sesquilinear form corresponding to the variational formulation and obtain well-posedness of weak solutions if the wavenumber does not coincide with the third component of the wave vector. Convergence of the finite element method with the transparent boundary condition is verified. We also derive equivalent boundary integral equations based on the quasi-periodic Green's function and prove the unique solvability result for piecewise smooth grating profiles with a finite number of corner points.

Keywords. Diffraction gratings; conical diffraction; variational method; integral equation method; finite element analysis; well-posedness.

AMS subject classifications. 35Q61; 65N30; 78-10; 35J50.

1. Introduction

Optical gratings are widely used in semi-conductor industry and grating diffraction problems have been extensively studied over the last thirty years in the literature via variational and integral equation methods. We refer to [1,2,8,9,14,16–18,22–24,28] and references therein for mathematical analysis of time-harmonic Maxwell's and Navier equations in periodic structures and also under TE and TM polarization cases. Concerning the numerical treatment, we refer to [3,4,6,7,9–11,27] for finite element method and boundary element method. In the polarization cases, the diffraction grating material is supposed to be periodic in one direction (x_1-direction), invariant in another direction (x_3-direction) and the incident direction of a time-harmonic electromagnetic plane wave is supposed to be orthogonal to the x_3-axis. In the TE (resp. TM) polarization case, the electric (resp. magnetic) field is parallel to the x_3-direction. In this paper we suppose that a time-harmonic plane wave is incident obliquely onto an imperfectly conducting grating, leading to the so-called conical diffraction problems in periodic structures with the impedance boundary condition. Such kind of boundary value problem doesn't seem to be studied in the literature for diffraction gratings, but can be used efficiently to model the interface behavior of the wave fields between a highly conducting material and an isotropic and homogeneous background medium. We note that the impedance boundary condition can be also used to overcome numerical difficulties caused by a large ratio period over thickness of a thin coated layer.

To the best of our knowledge, conical diffraction problems in periodic structures have been studied only with transmission conditions, where the full time-harmonic Maxwell system can be reduced to two coupled Helmholtz equations. Elschner, Hinder,

*Received: December 21, 2023; Accepted (in revised form): March 28, 2025. Communicated by Gang Bao.

†School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, People's Republic of China (ghhu@nankai.edu.cn).

‡School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, People's Republic of China (zhangjy97@mail.nankai.edu.cn).

§Corresponding author. School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, People's Republic of China (llzhu@mail.nankai.edu.cn).

Penzel and Schmidit [15] proved the well-posedness and regularity of the conical diffraction problem via the variational method. Elschner and Schmidt [19] studied stability of the conical diffraction problem with respect to variation of the grating profile and obtained explicit formulas for the derivatives of reflection and transmission coefficients with respect to perturbations of interfaces. We refer to [29] for the analysis of the integral equation method for coated conical gratings modeled by transmission conditions. If the scattering object is an infinitely long cylinder, the conical diffraction is also referred to as oblique scattering. In [30], Wang and Nakamura applied the integral equation method to prove the well-posedness for impedance cylinders embedded in a homogeneous medium. In an inhomogeneous medium, the uniqueness and existence of the oblique problem were also proved through the Lax-Phillips method; see [26]. We note that our diffraction problem corresponds to the model of electromagnetic scattering problems in chiral media when the chirality admittance is zero. We refer to Feng et al. [20] and Bao & Zhang [12] for the analysis and numerics of electromagnetic scattering problems with oblique plane wave incidence in non-periodic chiral media where the classical Sommerfeld radiation is used in place of the Rayleigh expansion radiation condition. The contribution of this paper is to study the conical diffraction problem in periodic structures under the impedance boundary condition and investigate both finite element and boundary integral equation methods. We complement the work of [15] by deriving new mapping properties of the Dirichlet-to-Neumann map. Moreover, we justify the strong ellipticity of the sesquilinear form corresponding to the variational formulation and obtain well-posedness of weak solutions if the wavenumber does not coincide with the third component of the wave vector. Convergence of the finite element method with the transparent boundary condition is verified. We also derive equivalent boundary integral equations based on the quasi-periodic Green's function and prove the unique solvability result for piecewise smooth grating profiles with a finite number of corner points.

The outline of the remaining part of the paper is organized as follows. In Section 2, we formulate the conical diffraction problem by deriving a coupled Helmholtz system with the impedance boundary condition from Maxwell's system. In Section 3, we state the variational formulation in one periodic cell with the DtN operator imposed on the artificial boundary. An energy formula is verified to prove the uniqueness of the truncated boundary value problem. The strong ellipticity of the variational formulation is shown and the well-posedness of the diffraction problem follows from the Fredholm theory. In Section 4, we show the convergence of the finite element method based on the variational formulation. Finally, the solvability of an integral equation will be presented in Section 5.

2. Conical diffraction problem

Assume an incoming time-harmonic plane wave of the form

$$(\mathcal{E}^{in},\mathcal{H}^{in})=(\mathbf{p}e^{i\alpha x_1-i\beta x_2+i\gamma x_3},\mathbf{q}e^{i\alpha x_1-i\beta x_2+i\gamma x_3})e^{-i\omega t}=:(\mathbf{E}^{in},\mathbf{H}^{in})e^{-i\omega t}, \tag{2.1}$$

is incident onto an imperfectly conducting grating with a high conductivity embedded in an isotropic homogeneous medium in $\mathbb{R}^3$. Denote by $\tilde{\Gamma}$ the grating profile and $\tilde{\Omega}$ the unbounded domain above $\tilde{\Gamma}$. The diffraction problem can be modeled by the reduced Maxwell's system

$$\nabla\times\mathbf{E}=i\omega\mu\mathbf{H},\quad \nabla\times\mathbf{H}=-i\omega\epsilon\mathbf{E}\quad \text{in } \tilde{\Omega}, \tag{2.2}$$

where the total fields $(\mathbf{E},\mathbf{H})$ are the sum of the incident waves $(\mathbf{E}^{in},\mathbf{H}^{in})$ and the outgoing scattered waves $(\mathbf{E}^{sc},\mathbf{H}^{sc})$ in $\tilde{\Omega}$ and ω denotes the angular frequency. The

dielectric coefficient ϵ and the magnetic permeability μ of the homogeneous medium in $\tilde{\Omega}$ are both assumed to be positive constants. Set $k=\omega\sqrt{\epsilon\mu}$ as the wavenumber of the background medium. We enforce the impedance boundary condition on $\tilde{\Gamma}$:

$$\nu\times\mathbf{E}\times\nu=\lambda(\nu\times\mathbf{H}) \quad \text{on } \tilde{\Gamma}, \tag{2.3}$$

where $\nu=(\nu_1,\nu_2,\nu_3)\in\mathbb{S}^2:=\{x\in\mathbb{R}^3:|x|=1\}$ is normal to $\tilde{\Gamma}$ directed into the exterior of $\tilde{\Omega}$ and $\lambda<0$ is the impedance coefficient which is assumed to be a constant (If ν is normal to $\tilde{\Gamma}$ directed into the interior of $\tilde{\Omega}$, then the constant λ should be a positive real number). The problem (2.1)–(2.3) is called a conical diffraction problem if the incident direction $\mathbf{k}=:(\alpha,-\beta,\gamma)$ is not orthogonal to the x_3-direction, i.e., $\gamma\neq0$. For conical diffraction problems, the wave vectors of the reflected or transmitted propagating modes lie on the surface of a cone whose axis is parallel to the x_3-direction [29]. We refer to Figure 2.1 for an illustration of the grating conical diffraction problem.

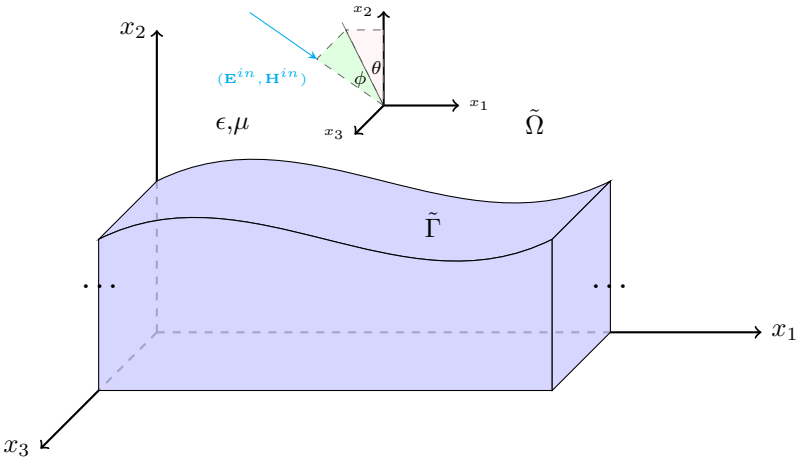

FIG. 2.1. *Geometry of the three-dimensional conical diffraction problem in one periodic cell. ϕ is the angle between incident direction $\boldsymbol{k}$ and (x_1,x_2)-plane. θ denotes the angle between $(\alpha,-\beta,0)$ and the x_2-axis.*

In order for $(\mathbf{E}^{in},\mathbf{H}^{in})$ given in (2.1) to satisfy (2.2), the constant amplitude vector $\mathbf{p}$ must be perpendicular to the wave vector $\mathbf{k}=(\alpha,-\beta,\gamma)$, that is $\mathbf{p}\cdot\mathbf{k}=0$. Furthermore $\mathbf{k}\cdot\mathbf{k}=k^2=\omega^2\epsilon\mu$ and $\mathbf{q}=(\omega\mu)^{-1}\mathbf{k}\times\mathbf{p}$. We can express the wave vector $\mathbf{k}$ as

$$\mathbf{k}=(\alpha,-\beta,\gamma):=k(\sin\theta\cos\phi,-\cos\theta\cos\phi,\sin\phi)\in\mathbb{R}^3,$$

in terms of the angles of incidence $\theta,\phi\in(-\pi/2,\pi/2)$.

Assume that $\tilde{\Gamma}$ remains invariant in x_3 and is 2π-periodic in x_1. If the incoming wave is of the form (2.1), we make an ansatz on the total field

$$(\mathbf{E},\mathbf{H})(x_1,x_2,x_3)=(E(x_1,x_2),H(x_1,x_2))\,e^{i\gamma x_3},$$

with $E=(E_1,E_2,E_3),H=(H_1,H_2,H_3):\mathbb{R}^2\to\mathbb{C}^3$. The Maxwell equations (2.2) can be reduced to two Helmholtz equations for the total fields $u=E_3$ and $v=H_3$ (see [15]):

$$\Delta u+\kappa^2u=0, \quad \Delta v+\kappa^2v=0 \quad \text{in } \Omega, \quad \kappa^2=k^2-\gamma^2.$$

Here Ω denotes the restriction of the cross-section of $\tilde{\Omega}$ by the (x_1,x_2)-plane to one periodic cell $(0,2\pi)$. Analogously, Γ denotes the counter part of $\tilde{\Gamma}$ in the periodic cell $(0,2\pi)$. The reduced geometry of the conical diffraction problem is shown in Figure 2.2. Next, we turn to the reduction of the boundary condition (2.3) in $\mathbb{R}^2$. Obviously, we

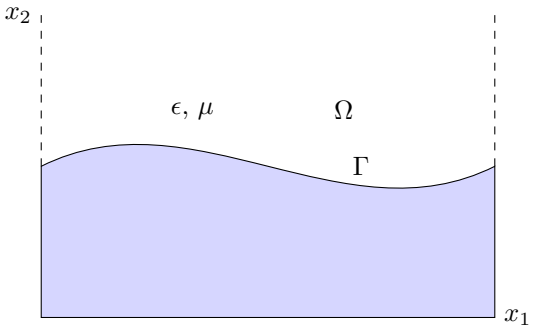

FIG. 2.2. *Geometry of the conical diffraction problem in* $\mathbb{R}^2$.

have $\nu_3=0$ and

$$\begin{aligned}&(\nu\times E)\times\nu=(-\nu_2(\nu_1E_2-\nu_2E_1),-\nu_1(\nu_1E_2-\nu_2E_1),E_3),\\&\nu\times H=(\nu_2H_3,-\nu_1H_3,\nu_1H_2-\nu_2H_1).\end{aligned}\tag{2.4}$$

Moreover, there holds (see [15])

$$\nu_1E_2-\nu_2E_1=\frac{i\gamma}{\kappa^2}\frac{\partial E_3}{\partial\tau}-\frac{i\omega\mu}{\kappa^2}\frac{\partial H_3}{\partial n},\quad \nu_1H_2-\nu_2H_1=\frac{i\gamma}{\kappa^2}\frac{\partial H_3}{\partial\tau}+\frac{i\omega\epsilon}{\kappa^2}\frac{\partial E_3}{\partial n},\tag{2.5}$$

with

$$n=(\nu_1,\nu_2),\ \tau=(-\nu_2,\nu_1),\ \partial_n=\nu_1\partial_1+\nu_2\partial_2,\ \partial_\tau=-\nu_2\partial_1+\nu_1\partial_2,\ \partial_j=\frac{\partial}{\partial x_j}.$$

Meanwhile, for the reduced Helmholtz equation, the incoming time-harmonic plane wave (2.1) takes the form

$$u^i=p_3e^{i\alpha x_1-i\beta x_2},\quad v^i=q_3e^{i\alpha x_1-i\beta x_2}.$$

Combining (2.4)-(2.5) and the impedance boundary condition (2.3), we get

$$-\frac{i\gamma}{\kappa^2}\frac{\partial E_3}{\partial\tau}+\frac{i\omega\mu}{\kappa^2}\frac{\partial H_3}{\partial n}=\lambda H_3,\quad E_3=\lambda\left(\frac{i\gamma}{\kappa^2}\frac{\partial H_3}{\partial\tau}+\frac{i\omega\epsilon}{\kappa^2}\frac{\partial E_3}{\partial n}\right),$$

which, for $u=E_3$ and $v=H_3$, is equivalent to the boundary conditions

$$\lambda\frac{\partial u}{\partial n}+\frac{i\kappa^2}{\omega\epsilon}u+\frac{\lambda\gamma}{\omega\epsilon}\frac{\partial v}{\partial\tau}=0,\quad \frac{\partial v}{\partial n}+\frac{i\lambda\kappa^2}{\omega\mu}v-\frac{\gamma}{\omega\mu}\frac{\partial u}{\partial\tau}=0\quad\text{on }\Gamma.\tag{2.6}$$

Using $\gamma=\omega\sqrt{\epsilon\mu}\sin\phi$ and $\kappa^2=k^2\cos^2\phi=\omega^2\mu\epsilon\cos^2\phi$, the previous boundary conditions can be written as

$$\begin{cases}\lambda\frac{\partial u}{\partial n}+i\omega\mu\cos^2\phi\,u+\lambda\sin\phi\sqrt{\frac{\mu}{\epsilon}}\frac{\partial v}{\partial\tau}=0,\\ \frac{\partial v}{\partial n}+i\lambda\omega\epsilon\cos^2\phi\,v-\sin\phi\sqrt{\frac{\epsilon}{\mu}}\frac{\partial u}{\partial\tau}=0,\end{cases}\quad\text{on }\Gamma.\tag{2.7}$$

REMARK 2.1. If $\lambda=0$, then the boundary conditions in (2.6) (or (2.7)) reduce to $\frac{\partial v}{\partial n}=u=0$, which correspond to the TE or TM polarization of the electromagnetic scattering by perfectly conducting gratings. If $\phi=0$, both u and v satisfy the standard impedance/Robin boundary condition for the Helmholtz equation, that is,

$$\frac{\partial u}{\partial n}+i\omega\mu/\lambda\, u=0, \qquad \frac{\partial v}{\partial n}+i\lambda\omega\epsilon v=0 \qquad \text{on } \Gamma.$$

3. Radiation condition and variational formulation

In this section we adapt the variational framework of [15] to the case of the impedance boundary condition and derive new mapping properties of the transparent boundary operator. Well-posedness of the grating diffraction problem will be justified for any $\kappa^2>0$ (that is, $k^2>\gamma^2$) by applying the Fredholm alternative. For $b>\Gamma_{\max}:=\max_{x\in\Gamma}\{x_2\}$, define

$$\Gamma_b:=\{(x_1,b):0<x_1<2\pi\}, \quad \Omega_b:=\{x\in\Omega:x_2<b\}.$$

A function $u(x_1,x_2)$ is called α-quasiperiodic if $e^{-i\alpha x_1}u(x_1,x_2)$ is 2π-periodic in x_1, or equivalently,

$$u(x_1+2\pi,x_2)=e^{2i\alpha\pi}u(x_1,x_2).$$

Since the incident field is α-quasiperiodic, the scattered fields u^s, v^s are also assumed to be α-quasiperiodic. Then the functions $u^s(x_1,x_2)e^{-i\alpha x_1}$, $v^s(x_1,x_2)e^{-i\alpha x_1}$ can be expanded as a Fourier series. Inserting these series into the Helmholtz equation, we can express u^s and v^s as a sum of plane waves. Physically, the scattered fields (u^s,v^s) remain bounded as $x_2\to+\infty$, leading to the well-known Rayleigh expansion condition:

$$u^s(x)=\sum_{n\in\mathbb{Z}}u_n e^{i\alpha_n x_1+i\beta_n x_2}, \quad v^s(x)=\sum_{n\in\mathbb{Z}}v_n e^{i\alpha_n x_1+i\beta_n x_2}, \quad x_2>\Gamma_{\max}, \tag{3.1}$$

with the Rayleigh coefficients $u_n,v_n\in\mathbb{C}$, where

$$\alpha_n:=n+\alpha, \quad \beta_n:=\begin{cases}\sqrt{\kappa^2-|\alpha_n|^2}, & |\alpha_n|\le\kappa,\\ i\sqrt{|\alpha_n|^2-\kappa^2}, & |\alpha_n|>\kappa,\end{cases}$$

with $i=\sqrt{-1}$. It is clear that (u^s,v^s) in (3.1) can be split into the finite sum $\sum_{|\alpha_n|\le k}$ of outgoing plane waves and the infinite sum $\sum_{|\alpha_n|>k}$ of exponentially decaying waves, which are called surface or evanescent waves. We summarize our conical diffraction problem as follows:

$$\begin{cases}\Delta u+\kappa^2 u=0,\ \Delta v+\kappa^2 v=0 & \text{in } \Omega,\\ \lambda\frac{\partial u}{\partial n}+\frac{i\kappa^2}{\omega\epsilon}u+\frac{\lambda\gamma}{\omega\epsilon}\frac{\partial v}{\partial\tau}=0,\ \frac{\partial v}{\partial n}+\frac{i\lambda\kappa^2}{\omega\mu}v-\frac{\gamma}{\omega\mu}\frac{\partial u}{\partial\tau}=0 & \text{on } \Gamma,\\ u^s \text{ and } v^s \text{ fulfill the Rayleigh expansion (3.1).}\end{cases} \tag{3.2}$$

Then we introduce the function spaces

$$H^k_\alpha(\Omega_b)=\{u\in H^k(\Omega_b):u \text{ is } \alpha\text{-quasiperiodic}\},$$

$$X=\{(u,v)\in H^1(\Omega_b)^2:u,v \text{ are } \alpha\text{-quasiperiodic}\}.$$

In order to derive the variational formula, we will need Green's formula for functions in $H^1_\alpha(\Omega_b)$.

LEMMA 3.1. *Assume* $f \in H^2_\alpha(\Omega_b)$ *and* $g \in H^1_\alpha(\Omega_b)$. *Then*

$$\int_{\Omega_b} \nabla f \cdot \nabla \overline{g} + \Delta f \overline{g}\, dx = \int_{\partial\Omega_b} \partial_n f \overline{g}\, ds, \quad \int_{\Omega_b} \nabla f \cdot \nabla^\perp \overline{g}\, dx = -\int_{\partial\Omega_b} \partial_\tau f \overline{g}\, ds,$$

where $\nabla = (\partial_1, \partial_2)$ *and* $\nabla^\perp = (-\partial_2, \partial_1)$.

Let $u, v \in H^1_\alpha(\Omega_b)$ solve the conical diffraction problem (3.2). Applying Green's formula to the Helmholtz equations yields

$$0 = \int_{\Omega_b} (\Delta u + \kappa^2 u)\overline{\varphi}\, dx = \int_{\Omega_b} -\nabla u \cdot \nabla \overline{\varphi} + \kappa^2 u \overline{\varphi}\, dx + \int_{\partial\Omega_b} \partial_n u \overline{\varphi}\, ds, \tag{3.3}$$

$$\int_{\Omega_b} \nabla v \cdot \nabla^\perp \overline{\varphi}\, dx = -\int_{\partial\Omega_b} \partial_\tau v \overline{\varphi}\, ds \tag{3.4}$$

for all $\varphi \in H^1_\alpha(\Omega_b)$. Multiplying the equations (3.3) and (3.4) by the constant factors $\frac{\omega\epsilon}{\kappa^2}$ and $\frac{\gamma}{\kappa^2}$, respectively, and taking the difference of the resulting formulas, we get

$$\int_{\partial\Omega_b} \frac{\omega\epsilon}{\kappa^2} \partial_n u \overline{\varphi} + \frac{\gamma}{\kappa^2} \partial_\tau v \overline{\varphi}\, ds = \int_{\Omega_b} \left[\frac{\omega\epsilon}{\kappa^2} \nabla u \cdot \nabla \overline{\varphi} - \frac{\gamma}{\kappa^2} \nabla v \cdot \nabla^\perp \overline{\varphi} - \omega\epsilon u \overline{\varphi} \right] dx. \tag{3.5}$$

Similarly, we get

$$0 = \int_{\Omega_b} (\Delta v + \kappa^2 v)\overline{\psi}\, dx = \int_{\Omega_b} -\nabla v \cdot \nabla \overline{\psi} + \kappa^2 v \overline{\psi}\, dx + \int_{\partial\Omega_b} \partial_n v \overline{\psi}\, ds, \tag{3.6}$$

$$\int_{\Omega_b} \nabla u \cdot \nabla^\perp \overline{\psi}\, dx = -\int_{\partial\Omega_b} \partial_\tau u \overline{\psi}\, ds, \tag{3.7}$$

for all $\psi \in H^1_\alpha(\Omega_b)$. Multiplying the equations (3.6) and (3.7) by the constant factors $\frac{\omega\mu}{\kappa^2}$ and $\frac{\gamma}{\kappa^2}$, respectively, and taking the sum of the two formulas, we get

$$\int_{\partial\Omega_b} \frac{\omega\mu}{\kappa^2} \partial_n v \overline{\psi} - \frac{\gamma}{\kappa^2} \partial_\tau u \overline{\psi}\, ds = \int_{\Omega_b} \left[\frac{\omega\mu}{\kappa^2} \nabla v \cdot \nabla \overline{\psi} + \frac{\gamma}{\kappa^2} \nabla u \cdot \nabla^\perp \overline{\psi} - \omega\mu v \overline{\psi} \right] dx. \tag{3.8}$$

Recalling the boundary conditions (2.6) on Γ, the left-hand terms of (3.5) and (3.8) over the integral Γ can be reformulated as

$$\int_\Gamma \frac{\omega\epsilon}{\kappa^2} \partial_n u \overline{\varphi} + \frac{\gamma}{\kappa^2} \partial_\tau v \overline{\varphi}\, ds = \int_\Gamma \frac{\omega\epsilon}{\lambda\kappa^2} \left(-\frac{i\kappa^2}{\omega\epsilon} u \right) \overline{\varphi}\, ds = -\frac{i}{\lambda} \int_\Gamma u \overline{\varphi}\, ds,$$
$$\int_\Gamma \frac{\omega\mu}{\kappa^2} \partial_n v \overline{\psi} - \frac{\gamma}{\kappa^2} \partial_\tau u \overline{\psi}\, ds = \int_\Gamma \frac{\omega\mu}{\kappa^2} \left(-\frac{i\lambda\kappa^2}{\omega\mu} v \right) \overline{\psi}\, ds = -i\lambda \int_\Gamma v \overline{\psi}\, ds.$$

Therefore, we need to find $(u, v) \in X$ such that for all $(\varphi, \psi) \in X$,

$$\begin{aligned} 0 = {} & \frac{i}{\lambda} \int_\Gamma u \overline{\varphi}\, ds + \int_{\Omega_b} \left[\frac{\omega\epsilon}{\kappa^2} \nabla u \cdot \nabla \overline{\varphi} - \frac{\gamma}{\kappa^2} \nabla v \cdot \nabla^\perp \overline{\varphi} - \omega\epsilon u \overline{\varphi} \right] dx \\ & - \int_{\Gamma_b} \left[\frac{\omega\epsilon}{\kappa^2} \partial_n u \overline{\varphi} + \frac{\gamma}{\kappa^2} \partial_\tau v \overline{\varphi} \right] ds, \end{aligned} \tag{3.9}$$

$$0 = i\lambda \int_\Gamma v\overline{\psi} ds + \int_{\Omega_b} \left[\frac{\omega\mu}{\kappa^2} \nabla v \cdot \nabla \overline{\psi} + \frac{\gamma}{\kappa^2} \nabla u \cdot \nabla^\perp \overline{\psi} - \omega\mu v \overline{\psi} \right] dx$$
$$- \int_{\Gamma_b} \left[\frac{\omega\mu}{\kappa^2} \partial_n v \overline{\psi} - \frac{\gamma}{\kappa^2} \partial_\tau u \overline{\psi} \right] ds. \tag{3.10}$$

Combining (3.9) and (3.10), we get

$$\int_\Gamma \frac{i}{\lambda} u\overline{\varphi} + i\lambda v\overline{\psi} ds + \int_{\Omega_b} \left[\frac{\omega\epsilon}{\kappa^2} \nabla u \cdot \nabla \overline{\varphi} - \frac{\gamma}{\kappa^2} \nabla v \cdot \nabla^\perp \overline{\varphi} - \omega\epsilon u \overline{\varphi} + \frac{\omega\mu}{\kappa^2} \nabla v \cdot \nabla \overline{\psi} \right.$$
$$\left. + \frac{\gamma}{\kappa^2} \nabla u \cdot \nabla^\perp \overline{\psi} - \omega\mu v \overline{\psi} \right] dx - \int_{\Gamma_b} \frac{1}{\kappa^2} \begin{pmatrix} \omega\epsilon\partial_n u + \gamma\partial_\tau v \\ \omega\mu\partial_n v - \gamma\partial_\tau u \end{pmatrix} \cdot \begin{pmatrix} \overline{\varphi} \\ \overline{\psi} \end{pmatrix} ds = 0. \tag{3.11}$$

Definition 3.1 (DtN map). *The Dirichlet-to-Neumann (DtN) map T is defined by*

$$T : (g_1, g_2)^\top \to -\left(\frac{\omega\epsilon}{\kappa^2} \partial_n w_1 + \frac{\gamma}{\kappa^2} \partial_\tau w_2, \frac{\omega\mu}{\kappa^2} \partial_n w_2 - \frac{\gamma}{\kappa^2} \partial_\tau w_1 \right)^\top \quad \text{on } \Gamma_b,$$

where $w_j (j = 1, 2)$ is the unique radiation solution to the Helmholtz equation $\Delta w_j + \kappa^2 w_j = 0$ in $x_2 > b$ with the Dirichlet boundary condition $w_j = g_j$ on Γ_b.

Now we want to derive an analytical expression of the DtN map T. For the α-quasiperiodic vector function $g = (g_1, g_2)^\top \in H_\alpha^{1/2}(\Gamma_b)^2$, its Fourier expansion takes the form $g(x_1) = \sum_{n\in\mathbb{Z}} \hat{g}_n e^{i\alpha_n x_1}$, where $\hat{g}_n = (\hat{g}_{n,1}, \hat{g}_{n,2})^\top$. It is easy to deduce that

$$w_j(x) = \sum_{n\in\mathbb{Z}} \hat{g}_{n,j} e^{i\alpha_n x_1 + i\beta_n (x_2 - b)}, \quad x_2 > b, \ j = 1, 2,$$

where w_j is the function specified in the Definition 3.1. Direct calculations show

$$\begin{aligned} & -\left(\frac{\omega\epsilon}{\kappa^2} \partial_n w_1 + \frac{\gamma}{\kappa^2} \partial_\tau w_2, \ \frac{\omega\mu}{\kappa^2} \partial_n w_2 - \frac{\gamma}{\kappa^2} \partial_\tau w_1 \right)^\top \Big|_{\Gamma_b} \\ = & -\frac{1}{\kappa^2} \left(\omega\epsilon \sum_{n\in\mathbb{Z}} i\beta_n \hat{g}_{n,1} e^{i\alpha_n x_1} + \gamma \sum_{n\in\mathbb{Z}} (-i\alpha_n) \hat{g}_{n,2} e^{i\alpha_n x_1}, \right. \\ & \left. \omega\mu \sum_{n\in\mathbb{Z}} i\beta_n \hat{g}_{n,2} e^{i\alpha_n x_1} - \gamma \sum_{n\in\mathbb{Z}} (-i\alpha_n) \hat{g}_{n,1} e^{i\alpha_n x_1} \right)^\top \\ = & \sum_{n\in\mathbb{Z}} M_n \hat{g}_n e^{i\alpha_n x_1}, \end{aligned}$$

where

$$M_n = \frac{1}{\kappa^2} \begin{pmatrix} -i\omega\epsilon\beta_n & i\gamma\alpha_n \\ -i\gamma\alpha_n & -i\omega\mu\beta_n \end{pmatrix}. \tag{3.12}$$

Hence, the operator T acting on the α-quasiperiodic vector function $w \in H_\alpha^{1/2}(\Gamma_b)^2$ can be expressed as

$$(Tw)(x) = \sum_{n\in\mathbb{Z}} M_n \hat{w}_n e^{i\alpha_n x}, \quad \hat{w}_n = \frac{1}{2\pi} \int_0^{2\pi} w(x) e^{-i\alpha_n x} dx \in \mathbb{C}^2.$$

LEMMA 3.2 (see [15]). *The DtN operator $T: H_\alpha^{1/2}(\Gamma_b)^2 \to H_\alpha^{-1/2}(\Gamma_b)^2$ is continuous, i.e., there exists a positive constant C such that*

$$\|Tw\|_{H_\alpha^{-1/2}(\Gamma_b)^2} \le C\|w\|_{H_\alpha^{1/2}(\Gamma_b)^2} \qquad \text{for all } w \in H_\alpha^{1/2}(\Gamma_b)^2.$$

Then we return back to the last term of the left-hand side of (3.11). Direct calculations show

$$T\begin{pmatrix} u^s|_{\Gamma_b} \\ v^s|_{\Gamma_b} \end{pmatrix} = \sum_{n\in\mathbb{Z}} M_n \begin{pmatrix} u_n \\ v_n \end{pmatrix} e^{i\alpha_n x_1 + i\beta_n b} = \sum_{n\in\mathbb{Z}} \frac{1}{\kappa^2} \begin{pmatrix} -i\omega\epsilon\beta_n u_n + i\gamma\alpha_n v_n \\ -i\gamma\alpha_n u_n - i\omega\mu\beta_n v_n \end{pmatrix} e^{i\alpha_n x_1 + i\beta_n b},$$
$$T\begin{pmatrix} u^i|_{\Gamma_b} \\ v^i|_{\Gamma_b} \end{pmatrix} = M_0 \begin{pmatrix} p_3 \\ q_3 \end{pmatrix} e^{i\alpha x_1 - i\beta b} = -\frac{1}{\kappa^2} \begin{pmatrix} i\omega\epsilon\beta p_3 - i\gamma\alpha q_3 \\ i\gamma\alpha p_3 + i\omega\mu\beta q_3 \end{pmatrix} e^{i\alpha x_1 - i\beta b}.$$

For $x_2 > \Gamma_{\max}$, (u,v) are given by

$$u = p_3 e^{i\alpha x_1 - i\beta x_2} + \sum_{n\in\mathbb{Z}} u_n e^{i\alpha_n x_1 + i\beta_n x_2}, \quad v = q_3 e^{i\alpha x_1 - i\beta x_2} + \sum_{n\in\mathbb{Z}} v_n e^{i\alpha_n x_1 + i\beta_n x_2}.$$

Then we have

$$\frac{1}{\kappa^2}\begin{pmatrix} \omega\epsilon\partial_\nu u + \gamma\partial_\tau v \\ \omega\mu\partial_\nu v - \gamma\partial_\tau u \end{pmatrix}\Big|_{\Gamma_b} = -T\begin{pmatrix} u|_{\Gamma_b} \\ v|_{\Gamma_b} \end{pmatrix} - \frac{2}{\kappa^2}\begin{pmatrix} i\omega\epsilon\beta p_3 \\ i\omega\mu\beta q_3 \end{pmatrix} e^{i\alpha x_1 - i\beta b}. \tag{3.13}$$

Inserting (3.13) into (3.11), we get the variational formulation

$$B(u,v;\varphi,\psi) = F(\varphi,\psi) \qquad \text{for all } (\varphi,\psi) \in X, \tag{3.14}$$

where

$$B(u,v;\varphi,\psi) := \int_\Gamma \frac{i}{\lambda} u\overline{\varphi} + i\lambda v\overline{\psi}\, ds + \int_{\Omega_b} \Big[\frac{\omega\epsilon}{\kappa^2}\nabla u\cdot\nabla\overline{\varphi} - \frac{\gamma}{\kappa^2}\nabla v\cdot\nabla^\perp\overline{\varphi} + \frac{\omega\mu}{\kappa^2}\nabla v\cdot\nabla\overline{\psi}$$
$$+ \frac{\gamma}{\kappa^2}\nabla u\cdot\nabla^\perp\overline{\psi} - \omega\epsilon u\overline{\varphi} - \omega\mu v\overline{\psi}\Big]\, dx + \int_{\Gamma_b} T\begin{pmatrix} u \\ v \end{pmatrix}\cdot\begin{pmatrix} \overline{\varphi} \\ \overline{\psi} \end{pmatrix} ds, \tag{3.15}$$

$$F(\varphi,\psi) := -\frac{2i\omega\beta e^{-i\beta b}}{\kappa^2}\int_{\Gamma_b} (\epsilon p_3\overline{\varphi} + \mu q_3\overline{\psi}) e^{i\alpha x_1}\, ds. \tag{3.16}$$

Below we prove an energy formula under the impedance boundary condition.

LEMMA 3.3. *Let $u, v \in H_\alpha^1(\Omega_b)$ be the total fields to our conical diffraction problem. We have the energy formula*

$$\frac{2\pi\omega}{\kappa^2}\sum_{|\alpha_n|\le\kappa} \beta_n(\epsilon|u_n|^2 + \mu|v_n|^2) = \int_\Gamma \frac{1}{\lambda}|u|^2 + \lambda|v|^2\, ds + \frac{2\pi\omega\beta}{\kappa^2}(\epsilon|p_3|^2 + \mu|q_3|^2). \tag{3.17}$$

Proof. Choosing $\varphi = u$ and $\psi = v$ in (3.11), we have

$$0 = \int_\Gamma \frac{i}{\lambda}|u|^2 + i\lambda|v|^2\, ds - \frac{1}{\kappa^2}\int_{\Gamma_b}\begin{pmatrix} \omega\epsilon\partial_n u + \gamma\partial_\tau v \\ \omega\mu\partial_n v - \gamma\partial_\tau u \end{pmatrix}\cdot\begin{pmatrix} \overline{u} \\ \overline{v} \end{pmatrix} ds$$
$$+ \int_{\Omega_b}\Big[\frac{\omega\epsilon}{\kappa^2}|\nabla u|^2 - \frac{\gamma}{\kappa^2}\nabla v\cdot\nabla^\perp\overline{u} - \omega\epsilon|u|^2 + \frac{\omega\mu}{\kappa^2}|\nabla v|^2 + \frac{\gamma}{\kappa^2}\nabla u\cdot\nabla^\perp\overline{v} - \omega\mu|v|^2\Big]\, dx. \tag{3.18}$$

For the total fields u and v, we can easily get the imaginary part of the integral over Γ_b,

$$\begin{aligned}
&\mathrm{Im}\int_{\Gamma_b}\begin{pmatrix}\omega\epsilon\partial_n u+\gamma\partial_\tau v\\ \omega\mu\partial_n v-\gamma\partial_\tau u\end{pmatrix}\cdot\begin{pmatrix}\overline{u}\\ \overline{v}\end{pmatrix}ds\\
=&\mathrm{Im}\int_0^{2\pi}\begin{pmatrix}\omega\epsilon\partial_2 u-\gamma\partial_1 v\\ \omega\mu\partial_2 v+\gamma\partial_1 u\end{pmatrix}\Big|_{\Gamma_b}\cdot\begin{pmatrix}\overline{u}\\ \overline{v}\end{pmatrix}\Big|_{\Gamma_b}dx_1\\
=&\mathrm{Im}\,2\pi\left[\omega\epsilon\left(-i\beta|p_3|^2+\sum_{n\in\mathbb{Z}}i\beta_n|u_n|^2\right)-\gamma\left(i\alpha q_3\bar{p}_3+\sum_{n\in\mathbb{Z}}i\alpha_n v_n\bar{u}_n\right)\right.\\
&\left.+\omega\mu\left(-i\beta|q_3|^2+\sum_{n\in\mathbb{Z}}i\beta_n|v_n|^2\right)+\gamma\left(i\alpha p_3\bar{q}_3+\sum_{n\in\mathbb{Z}}i\alpha_n u_n\bar{v}_n\right)\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathrm{Im}\int_{\Gamma_b}\begin{pmatrix}\omega\epsilon\partial_n u+\gamma\partial_\tau v\\ \omega\mu\partial_n v-\gamma\partial_\tau u\end{pmatrix}\cdot\begin{pmatrix}\overline{u}\\ \overline{v}\end{pmatrix}ds\\
=&-2\pi\omega\beta\left(\epsilon|p_3|^2+\mu|q_3|^2\right)+2\pi\sum_{|\alpha_n|\le\kappa}\omega\beta_n\left(\epsilon|u_n|^2+\mu|v_n|^2\right).
\end{aligned}\tag{3.19}$$

In addition,

$$\begin{aligned}
&\mathrm{Im}\int_{\Gamma_b}\nabla v\cdot\nabla^\perp\overline{u}-\nabla u\cdot\nabla^\perp\overline{v}\,dx\\
=&\mathrm{Im}\int_{\Gamma_b}-\partial_1 v\partial_2\bar{u}+\partial_2 v\partial_1\bar{u}-(-\partial_1 u\partial_2\bar{v}+\partial_2 u\partial_1\bar{v})\,dx\\
=&\mathrm{Im}\int_{\Gamma_b}-(\partial_1 v\partial_2\bar{u}+\partial_2 u\partial_1\bar{v})+(\partial_2 v\partial_1\bar{u}+\partial_1 u\partial_2\bar{v})\,dx=0.
\end{aligned}\tag{3.20}$$

Taking the imaginary part of (3.18) and using (3.19) and (3.20), we obtain

$$\begin{aligned}
0=&\int_\Gamma\frac{1}{\lambda}|u|^2+\lambda|v|^2\,ds-\mathrm{Im}\,\frac{1}{\kappa^2}\int_0^{2\pi}\begin{pmatrix}\omega\epsilon\partial_2 u-\gamma\partial_1 v\\ \omega\mu\partial_2 v+\gamma\partial_1 u\end{pmatrix}\Big|_{\Gamma_b}\cdot\begin{pmatrix}\overline{u}\\ \overline{v}\end{pmatrix}\Big|_{\Gamma_b}dx_1\\
=&\int_\Gamma\frac{1}{\lambda}|u|^2+\lambda|v|^2\,ds+\frac{2\pi\omega\beta}{\kappa^2}\left(\epsilon|p_3|^2+\mu|q_3|^2\right)-\frac{2\pi\omega}{\kappa^2}\sum_{|\alpha_n|\le\kappa}\beta_n\left(\epsilon|u_n|^2+\mu|v_n|^2\right),
\end{aligned}$$

which completes the proof. ∎

THEOREM 3.1. *Suppose that Γ is a Lipschitz curve, $k^2\neq\gamma^2$ and the impedance coefficient $\lambda<0$. Then, the variational problem (3.14) has at most one solution $(u,v)\in X$.*

Proof. To prove uniqueness, we assume $u^i=v^i=0$, i.e. $p_3=q_3=0$. Choosing $\varphi=u,\psi=v$ in (3.14) and taking the imaginary part, we have

$$\int_\Gamma\frac{1}{\lambda}|u|^2+\lambda|v|^2\,ds+\mathrm{Im}\int_{\Gamma_b}T\begin{pmatrix}u\\ v\end{pmatrix}\cdot\begin{pmatrix}\overline{u}\\ \overline{v}\end{pmatrix}ds=0.\tag{3.21}$$

By the definition of T, we have that for $w=(u,v)^\top$,

$$\mathrm{Im}\int_{\Gamma_b}Tw\cdot\overline{w}\,ds=\mathrm{Im}\int_{\Gamma_b}\sum_{n\in\mathbb{Z}}M_n\hat{w}_n e^{i\alpha_n x}\cdot\overline{\sum_{m\in\mathbb{Z}}\hat{w}_m e^{i\alpha_m x}}\,ds$$

$$
\begin{aligned}
&= \operatorname{Im} 2\pi \sum_{n\in\mathbb{Z}} M_n \hat{w}_n \cdot \overline{\hat{w}_n} \\
&= 2\pi \sum_{n\in\mathbb{Z}} (\operatorname{Im} M_n) \hat{w}_n \cdot \overline{\hat{w}_n},
\end{aligned}
$$

where $\hat{w}_n = (\hat{w}_{n,1}, \hat{w}_{n,2}) = (u_n, v_n)$. Recalling the expression of M_n, we have

$$
\begin{aligned}
\operatorname{Im} M_n &= \frac{1}{2i}(M_n - M_n^*) \\
&= \frac{1}{2i}\frac{1}{\kappa^2}\left[\begin{pmatrix} -i\omega\epsilon\beta_n & i\gamma\alpha_n \\ -i\gamma\alpha_n & -i\omega\mu\beta_n \end{pmatrix} - \begin{pmatrix} i\omega\epsilon\overline{\beta_n} & i\gamma\alpha_n \\ -i\gamma\alpha_n & i\omega\mu\overline{\beta_n} \end{pmatrix}\right] \\
&= \frac{1}{\kappa^2}\begin{pmatrix} \operatorname{Im}(-i\omega\epsilon\beta_n) & 0 \\ 0 & \operatorname{Im}(-i\omega\mu\beta_n) \end{pmatrix} \\
&= \begin{cases} \frac{1}{\kappa^2}\begin{pmatrix} -\omega\epsilon\beta_n & 0 \\ 0 & -\omega\mu\beta_n \end{pmatrix}, & |\alpha_n| \le \kappa, \\ 0, & |\alpha_n| > \kappa. \end{cases}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\operatorname{Im} \int_{\Gamma_b} Tw \cdot \overline{w}\, ds &= 2\pi \sum_{|\alpha_n|\le\kappa} \frac{1}{\kappa^2} \begin{pmatrix} -\omega\epsilon\beta_n & 0 \\ 0 & -\omega\mu\beta_n \end{pmatrix} \begin{pmatrix} \hat{w}_{n1} \\ \hat{w}_{n2} \end{pmatrix} \cdot \begin{pmatrix} \overline{\hat{w}_{n1}} \\ \overline{\hat{w}_{n2}} \end{pmatrix} \\
&= -\frac{2\pi\omega}{\kappa^2} \sum_{|\alpha_n|\le\kappa} \beta_n (\epsilon |\hat{w}_{n1}|^2 + \mu |\hat{w}_{n2}|^2) \le 0.
\end{aligned}
$$

Inserting these results into (3.21), we have

$$
\int_\Gamma \frac{1}{\lambda}|u|^2 + \lambda |v|^2\, ds - \frac{2\pi\omega}{\kappa^2} \sum_{|\alpha_n|\le\kappa} \beta_n (\epsilon |u_n|^2 + \mu |v_n|^2) = 0.
$$

Noting that $\lambda < 0$, we have $u = v = 0$ on Γ. By the impedance boundary condition (2.6), we have $\partial_n u = \partial_n v = 0$ on Γ. We get $u = v = 0$ in Ω as the consequence of the Holmgren theorem. ∎

REMARK 3.1. The proof of Theorem 3.1 provides an alternative approach to the proof of the energy formula (3.17) via matrix operations. One can also prove the uniqueness result by taking the imaginary part of the energy formula (3.17) with $p_3 = q_3 = 0$.

DEFINITION 3.2 (Strong ellipticity). *We call a bounded sesquilinear form $B(\cdot,\cdot)$ given on some Hilbert space X strongly elliptic if there exists a complex number θ, $|\theta| = 1$ and a compact form $q(\cdot,\cdot)$ such that*

$$
\operatorname{Re}(\theta B(u,u)) \ge c\|u\|_X^2 - q(u,u) \quad \text{for all } u \in X,
$$

for some constant $c > 0$.

The following theorem establishes the strong ellipticity of the form (3.15) and leads, together with Theorem 3.1 or Remark 3.1, to the solvability results for the conical diffraction problem.

THEOREM 3.2. *The sesquilinear form B defined in (3.14) is strongly elliptic over X.*

Before proving this theorem, we need to make some preparations. It is convenient to reformulate the variational form (3.15) as follows (see [15]):

$$B(u,v;\varphi,\psi)=A(u,v;\varphi,\psi)+B_1(u,v;\varphi,\psi)+C(u,v;\varphi,\psi)+D(u,v;\varphi,\psi),$$

where

$$\begin{aligned}
A(u,v;\varphi,\psi)&:=\int_\Gamma \frac{i}{\lambda}u\overline{\varphi}+i\lambda v\overline{\psi}\,ds,\\
C(u,v;\varphi,\psi)&:=\int_{\Omega_b}\omega\epsilon u\,\overline{\varphi}+\omega\mu v\,\overline{\psi}\,dx,\\
D(u,v;\varphi,\psi)&:=\int_{\Gamma_b} T\begin{pmatrix}u\\ v\end{pmatrix}\cdot\begin{pmatrix}\overline{\varphi}\\ \overline{\psi}\end{pmatrix}ds,
\end{aligned}$$

and

$$\begin{aligned}
B_1(u,v;\varphi,\psi)&:=\int_{\Omega_b}\left[\frac{\omega\epsilon}{\kappa^2}\nabla u\cdot\nabla\overline{\varphi}-\frac{\gamma}{\kappa^2}\nabla v\cdot\nabla^\perp\overline{\varphi}+\frac{\omega\mu}{\kappa^2}\nabla v\cdot\nabla\overline{\psi}+\frac{\gamma}{\kappa^2}\nabla u\cdot\nabla^\perp\overline{\psi}\right]dx\\
&=\int_{\Omega_b}\mathcal{D}(\partial_1 u,\partial_1 v,\partial_2 u,\partial_2 v)^{\mathsf{T}}\cdot\overline{(\partial_1\varphi,\partial_1\psi,\partial_2\varphi,\partial_2\psi)^{\mathsf{T}}}\,dx,
\end{aligned}$$

with the matrix $\mathcal{D}\in\mathbb{R}^{4\times4}$ given by (see [15])

$$\mathcal{D}=\frac{1}{\kappa^2}\begin{pmatrix}\omega\epsilon & 0 & 0 & -\gamma\\ 0 & \omega\mu & \gamma & 0\\ 0 & \gamma & \omega\epsilon & 0\\ -\gamma & 0 & 0 & \omega\mu\end{pmatrix}.$$

We can further write B_1 into the matrix form

$$B_1(u,v;\varphi,\psi)=\int_{\Omega_b}N^+\partial^+\begin{pmatrix}u\\ v\end{pmatrix}\cdot\overline{\partial^+\begin{pmatrix}\varphi\\ \psi\end{pmatrix}}+N^-\partial^-\begin{pmatrix}u\\ v\end{pmatrix}\cdot\overline{\partial^-\begin{pmatrix}\varphi\\ \psi\end{pmatrix}}\,ds \tag{3.22}$$

where

$$N^\pm=\frac{1}{\kappa^2}\begin{pmatrix}\omega\epsilon & \pm i\gamma\\ \mp i\gamma & \omega\mu\end{pmatrix},\quad \partial^+:=\frac{1}{\sqrt{2}}(-i\partial_1+\partial_2),\quad \partial^-:=\frac{1}{\sqrt{2}}(\partial_1-i\partial_2).$$

To study the form B, we need the following lemma.

LEMMA 3.4. *Choose $\theta=\frac{i+\delta}{|i+\delta|}$ with $\delta>0$ sufficiently small.*

(i) *For any $\xi\in\mathbb{C}^2$, we have $Re(\theta N^\pm\xi\cdot\bar{\xi})\geq C_N|\xi|^2$, where*

$$C_N=\frac{1}{2\omega\epsilon\mu\cos^2\phi}\mathrm{Re}\,\theta\left[(\epsilon+\mu)-\sqrt{(\epsilon-\mu)^2+4\epsilon\mu\sin^2\phi}\right]\geqslant 0. \tag{3.23}$$

(ii) *Let $M_n\in\mathbb{C}^{2\times2}$ be defined by (3.12). It holds that $Re(\theta M_n)\geq0$ for all $n\in\mathbb{Z}\backslash\mathcal{A}$, where the index set $\mathcal{A}$ is defined by*

$$\begin{aligned}
\mathcal{A}=\{n\in\mathbb{Z}:&-k(1+\sin\theta\cos\phi)<n\leq-k\cos\phi(1+\sin\theta)\\
&\text{or } k\cos\phi(1-\sin\theta)\leq n<k(1-\sin\theta\cos\phi)\}.
\end{aligned} \tag{3.24}$$

Proof.

(i) By the definition of $N^{\pm}$, we have

$$\mathrm{Re}(\theta N^{\pm}) = \frac{\theta N^{\pm} + (\theta N^{\pm})^*}{2} = \frac{1}{\kappa^2}\begin{pmatrix} \omega\epsilon \mathrm{Re}\theta & \pm i\gamma \mathrm{Re}\theta \\ \mp i\gamma \mathrm{Re}\theta & \omega\mu \mathrm{Re}\theta \end{pmatrix},$$

which is a Hermitian matrix. Recalling that $\gamma = k\sin\phi = \omega\sqrt{\epsilon\mu}\sin\phi$ and $\kappa^2 = k^2\cos^2\phi$, we compute the eigenvalues of $\mathrm{Re}(\theta N^{\pm})$ as following

$$\lambda_1 = \frac{1}{2\omega\epsilon\mu\cos^2\phi}\mathrm{Re}\theta\left[(\epsilon+\mu) + \sqrt{(\epsilon-\mu)^2 + 4\epsilon\mu\sin^2\phi}\right] > 0,$$
$$\lambda_2 = \frac{1}{2\omega\epsilon\mu\cos^2\phi}\mathrm{Re}\theta\left[(\epsilon+\mu) - \sqrt{(\epsilon-\mu)^2 + 4\epsilon\mu\sin^2\phi}\right] \geq 0.$$

Defining $C_N = \lambda_2 < \lambda_1$ and applying [21, Theorem 4.2.2], we complete the first part of the proof.

(ii) Recalling the definition of M_n, we have

$$\theta M_n = \frac{1}{\kappa^2}\begin{pmatrix} -i\omega\epsilon\beta_n\theta & i\gamma\alpha_n\theta \\ -i\gamma\alpha_n\theta & -i\omega\mu\beta_n\theta \end{pmatrix} = \frac{1}{\kappa^2}\begin{pmatrix} -i(\omega\mu)^{-1}k^2\beta_n\theta & i\gamma\alpha_n\theta \\ -i\gamma\alpha_n\theta & -i\omega\mu\beta_n\theta \end{pmatrix}.$$

Case 1. $|\alpha_n| < \kappa$, i.e., $\beta_n \in \mathbb{R}$ is a real number. We have

$$(\theta M_n)^* = \frac{1}{\kappa^2}\begin{pmatrix} i(\omega\mu)^{-1}k^2\beta_n\bar{\theta} & i\gamma\alpha_n\bar{\theta} \\ -i\gamma\alpha_n\bar{\theta} & i\omega\mu\beta_n\bar{\theta} \end{pmatrix}.$$

Therefore,

$$\mathrm{Re}(\theta M_n) = \frac{\theta M_n + (\theta M_n)^*}{2} = \frac{1}{\kappa^2}\begin{pmatrix} (\omega\mu)^{-1}k^2\beta_n \mathrm{Im}\theta & i\gamma\alpha_n \mathrm{Re}\theta \\ -i\gamma\alpha_n \mathrm{Re}\theta & \omega\mu\beta_n \mathrm{Im}\theta \end{pmatrix}.$$

In this case, $\mathrm{Re}(\theta M_n) \geq 0$ if and only if the following two conditions are satisfied:

$$\begin{aligned} &\mathrm{Im}\theta \geq 0, \\ &\det(\mathrm{Re}(\theta M_n)) = \frac{1}{\kappa^4}\left[(\omega\mu)^{-1}k^2\beta_n \mathrm{Im}\theta\omega\mu\beta_n \mathrm{Im}\theta - \gamma^2\alpha_n^2(\mathrm{Re}\theta)^2\right] \\ &\qquad = \frac{1}{\kappa^4}\left[k^2\beta_n^2 - \gamma^2\alpha_n^2\delta^2\right](\mathrm{Im}\theta)^2 \geq 0. \end{aligned} \tag{3.25}$$

The conditions in (3.25) obviously hold due to the definition of θ with a small $\delta > 0$.

Case 2. $|\alpha_n| \geq \kappa$, i.e., β_n is a pure imaginary number. We have

$$(\theta M_n)^* = \frac{1}{\kappa^2}\begin{pmatrix} (\omega\mu)^{-1}k^2|\beta_n|\bar{\theta} & i\gamma\alpha_n\bar{\theta} \\ -i\gamma\alpha_n\bar{\theta} & \omega\mu|\beta_n|\bar{\theta} \end{pmatrix}.$$

Therefore,

$$\mathrm{Re}(\theta M_n) = \frac{\theta M_n + (\theta M_n)^*}{2} = \frac{1}{\kappa^2}\begin{pmatrix} (\omega\mu)^{-1}k^2|\beta_n|\mathrm{Re}\theta & i\gamma\alpha_n \mathrm{Re}\theta \\ -i\gamma\alpha_n \mathrm{Re}\theta & \omega\mu|\beta_n|\mathrm{Re}\theta \end{pmatrix}.$$

In this case, $\mathrm{Re}(\theta M_n) \geq 0$ if and only if the following two conditions are satisfied:

$$\mathrm{Re}\theta \geq 0,$$

$$\det(\mathrm{Re}(\theta M_n)) = \frac{1}{\kappa^4}\left[(\omega\mu)^{-1}k^2|\beta_n|(\mathrm{Re}\theta)\omega\mu|\beta_n|\mathrm{Re}\theta - \gamma^2\alpha_n^2(\mathrm{Re}\theta)^2\right]$$
$$= \frac{1}{\kappa^4}\left[k^2|\beta_n|^2 - \gamma^2\alpha_n^2\right](\mathrm{Re}\theta)^2 \geq 0.$$

The first condition is obvious. The second condition can be fulfilled if $k^2|\beta_n|^2 - \gamma^2\alpha_n^2 \geq 0$. Recalling that $|\beta_n|^2 = \alpha_n^2 - (k^2 - \gamma^2)$, we have $\alpha_n^2 \geq k^2$. Combining the above two cases, we find that the matrix $\mathrm{Re}(\theta M_n)$ is not positive definite only when

$$n \in \mathcal{B} := \{n \in \mathbb{Z} : k^2 - \gamma^2 \leq \alpha_n^2 < k^2\}.$$

We should point out that $n \in \mathcal{B}$ if $\beta_n = 0$ (i.e. $|\alpha_n| = \kappa$). Recalling that $\alpha = k\sin\theta\cos\phi$ and $\alpha_n = n + \alpha$, the set $\mathcal{B}$ coincides with the set $\mathcal{A}$ given by (3.24). ∎

REMARK 3.2. The set $\mathcal{A}$ consists of a finite number of indexes only.

From the Rayleigh expansion, we can rewrite the restriction of (u, v) to Γ_b as

$$u|_{\Gamma_b} = \sum_{n\in\mathbb{Z}} \tilde{u}_n e^{i\alpha_n x_1}, \quad v|_{\Gamma_b} = \sum_{n\in\mathbb{Z}} \tilde{v}_n e^{i\alpha_n x_1},$$

where

$$\tilde{u}_n := \begin{cases} u_n e^{i\beta_n b}, & n \neq 0, \\ u_0 e^{i\beta b} + p_3 e^{-i\beta b}, & n = 0, \end{cases} \quad \tilde{v}_n := \begin{cases} v_n e^{i\beta_n b}, & n \neq 0, \\ v_0 e^{i\beta b} + q_3 e^{-i\beta b}, & n = 0. \end{cases}$$

We define the sesquilinear form

$$q(u, v; u, v) = 2\pi\mathrm{Re}\sum_{n\in\mathcal{A}} \theta M_n \begin{pmatrix} \tilde{u}_n \\ \tilde{v}_n \end{pmatrix} \cdot \begin{pmatrix} \overline{\tilde{u}_n} \\ \overline{\tilde{v}_n} \end{pmatrix}, \tag{3.26}$$

where the set $\mathcal{A}$ is defined by (3.24).

Proof. **(Proof of Theorem 3.2).** Choose $\theta = \frac{i+\delta}{|i+\delta|}$. By the definition of $\mathcal{A}$,

$$\mathrm{Re}(\theta A(u, v; u, v)) = \mathrm{Re}\int_\Gamma \frac{i+\delta}{|i+\delta|}\left(\frac{i}{\lambda}|u|^2 + i\lambda|v|^2\right) ds$$
$$= -\int_\Gamma \frac{1}{|i+\delta|}\left(\frac{1}{\lambda}|u|^2 + \lambda|v|^2\right) ds \geq 0.$$

Before calculating $\mathrm{Re}(\theta B_1(u, v; u, v))$, we compute the following relation:

$$\partial^+\begin{pmatrix} u \\ v \end{pmatrix} \cdot \overline{\partial^+\begin{pmatrix} u \\ v \end{pmatrix}}$$
$$= \frac{1}{2}\begin{pmatrix} -i\partial_1 u + \partial_2 u \\ -i\partial_1 v + \partial_2 v \end{pmatrix} \cdot \begin{pmatrix} i\partial_1\bar{u} + \partial_2\bar{u} \\ i\partial_1\bar{v} + \partial_2\bar{v} \end{pmatrix}$$
$$= \frac{1}{2}\left[|\nabla u|^2 + |\nabla v|^2 + i(\partial_1\bar{u}\partial_2 u + \partial_1\bar{v}\partial_2 v) - i(\partial_1 u\partial_2\bar{u} + \partial_1 v\partial_2\bar{v})\right].$$

Similarly,

$$\partial^-\begin{pmatrix} u \\ v \end{pmatrix} \cdot \overline{\partial^-\begin{pmatrix} u \\ v \end{pmatrix}} = \frac{1}{2}\left[|\nabla u|^2 + |\nabla v|^2 + i(\partial_1 u\partial_2\bar{u} + \partial_1 v\partial_2\bar{v}) - i(\partial_1\bar{u}\partial_2 u + \partial_1\bar{v}\partial_2 v)\right].$$

Therefore,

$$\partial^+\begin{pmatrix}u\\v\end{pmatrix}\cdot\overline{\partial^+\begin{pmatrix}u\\v\end{pmatrix}}+\partial^-\begin{pmatrix}u\\v\end{pmatrix}\cdot\overline{\partial^-\begin{pmatrix}u\\v\end{pmatrix}}=|\nabla u|^2+|\nabla v|^2.$$

Then, by (3.22) and Lemma 3.4, we have

$$\begin{aligned}
&\mathrm{Re}(\theta B_1(u,v;u,v))\\
&=\mathrm{Re}\left(\theta\int_{\Omega_b}N^+\partial^+\begin{pmatrix}u\\v\end{pmatrix}\cdot\overline{\partial^+\begin{pmatrix}u\\v\end{pmatrix}}+N^-\partial^-\begin{pmatrix}u\\v\end{pmatrix}\cdot\overline{\partial^-\begin{pmatrix}u\\v\end{pmatrix}}\,dx\right)\\
&\geq C_N\int_{\Omega_b}\partial^+\begin{pmatrix}u\\v\end{pmatrix}\cdot\overline{\partial^+\begin{pmatrix}u\\v\end{pmatrix}}+\partial^-\begin{pmatrix}u\\v\end{pmatrix}\cdot\overline{\partial^-\begin{pmatrix}u\\v\end{pmatrix}}\,dx\\
&=C_N\int_{\Omega_b}|\nabla u|^2+|\nabla v|^2\,dx\\
&=C_N\left(\|u\|^2_{H^1(\Omega_b)}+\|v\|^2_{H^1(\Omega_b)}\right)-C_N\int_{\Omega_b}|u|^2+|v|^2\,dx,
\end{aligned}$$

where $C_N\geq 0$ is defined by (3.23). It is obvious that

$$\begin{aligned}
\mathrm{Re}(\theta C(u,v;u,v))&=\mathrm{Re}\int_{\Omega_b}\frac{i+\delta}{|i+\delta|}\left(\omega\epsilon|u|^2+\omega\mu|v|^2\right)dx\\
&=\frac{\delta}{|i+\delta|}\int_{\Omega_b}\omega\epsilon|u|^2+\omega\mu|v|^2\,dx\geq 0.
\end{aligned}$$

For $u|_{\Gamma_b}=\sum_{n\in\mathbb{Z}}\tilde{u}_n e^{i\alpha_n x_1}$, $v|_{\Gamma_b}=\sum_{n\in\mathbb{Z}}\tilde{v}_n e^{i\alpha_n x_1}$, we get by using Lemma 3.4 (ii) that

$$\begin{aligned}
\mathrm{Re}(\theta D(u,v;u,v))&=\mathrm{Re}\left(\theta\int_{\Gamma_b}T\begin{pmatrix}u\\v\end{pmatrix}\cdot\begin{pmatrix}\overline{u}\\\overline{v}\end{pmatrix}ds\right)\\
&=2\pi\mathrm{Re}\left(\sum_{n\in\mathbb{Z}}\theta M_n\begin{pmatrix}\tilde{u}_n\\\tilde{v}_n\end{pmatrix}\cdot\begin{pmatrix}\overline{\tilde{u}_n}\\\overline{\tilde{v}_n}\end{pmatrix}\right)\\
&=2\pi\mathrm{Re}\left(\sum_{n\in\mathbb{Z}/\mathcal{A}}\theta M_n\begin{pmatrix}\tilde{u}_n\\\tilde{v}_n\end{pmatrix}\cdot\begin{pmatrix}\overline{\tilde{u}_n}\\\overline{\tilde{v}_n}\end{pmatrix}\right)+q(u,v;u,v)\\
&\geq q(u,v;u,v).
\end{aligned}$$

Note that $q(u,v;u,v)$ is a compact form, because $\mathcal{A}$ is a finite set. Therefore, by Lemma 3.4, we have

$$\begin{aligned}
&\mathrm{Re}(\theta B(u,v;u,v))\\
&=\mathrm{Re}(\theta A(u,v;u,v))+\mathrm{Re}(\theta B_1(u,v;u,v))+\mathrm{Re}(\theta C(u,v;u,v))+\mathrm{Re}(\theta D(u,v;u,v))\\
&\geq C_N\left(\|u\|^2_{H^1(\Omega_b)}+\|v\|^2_{H^1(\Omega_b)}\right)-Q(u,v;u,v),
\end{aligned}$$

where $C_N\geq 0$ is defined by (3.23) and

$$Q(u,v;u,v):=C_N\int_{\Omega_b}|u|^2+|v|^2\,dx-\frac{\delta}{|i+\delta|}\int_{\Omega_b}\omega\epsilon|u|^2+\omega\mu|v|^2\,dx-q(u,v;u,v)$$

is a compact form over $X\times X$. By Definition 3.2, we finish the proof. ∎

THEOREM 3.3. *Suppose that Γ is a Lipschitz curve, $k^2\neq\gamma^2$ and that the impedance coefficient $\lambda<0$. Then, the variational problem (3.9)–(3.10) admits a unique solution $(u,v)\in X$.*

Proof. Under the assumption of Theorem 3.2, the operator defined in (3.15) is a Fredholm operator with index zero. Using Theorem 3.1, we obtain the existence and uniqueness result as a consequence of the Fredholm alternative. ∎

4. Finite element analysis and error estimate

In this section we study the finite element approximation of the variational problem (3.14), following the framework of [6] and [9, Chapter 4]. Let $\{X_h : h\in(0,1)\}$ be a family of finite dimensional subspaces of $H^1_\alpha(\Omega_b)^2$, where h stands for the maximum mesh size after partitioning Ω_b into simple domains, for example, a regular triangulation of Ω_b. We make the following assumption on the subspace X_h (see [13]): for $(\varphi,\psi)\in H^\rho_\alpha(\Omega_b)^2$ with $\rho\geq 2$ it holds that

$$\begin{aligned}\inf_{(\xi,\eta)\in X_h}&\Big(\|(\varphi,\psi)-(\xi,\eta)\|_{L^2(\Omega_b)^2}+h\|(\nabla\varphi,\nabla\psi)-(\nabla\xi,\nabla\eta)\|_{L^2(\Omega_b)^2}\\&+h^{1/2}\|(\varphi,\psi)-(\xi,\eta)\|_{L^2(\Gamma_b)^2}+h\|(\varphi,\psi)-(\xi,\eta)\|_{H^{1/2}(\Gamma_b)^2}\\&+h^{1/2}\|(\varphi,\psi)-(\xi,\eta)\|_{L^2(\Gamma_b)^2}+h\|(\varphi,\psi)-(\xi,\eta)\|_{H^{1/2}(\Gamma)^2}\Big)\\\leq& Ch^l\|(\varphi,\psi)\|_{H^l(\Omega)^2},\quad l\in[2,\rho]\end{aligned}\tag{4.1}$$

where the positive constant C is independent of h and (φ,ψ). The finite element approximation to the variational formulation (3.14) is to find $(u_h,v_h)\in X_h$ such that

$$B(u_h,v_h;\varphi_h,\psi_h)=F(\varphi_h,\psi_h)\quad\text{for all }(\varphi_h,\psi_h)\in X_h,\tag{4.2}$$

where B is defined by (3.15) and F is defined by (3.16). The finite element method consists of the following steps to solve (4.2):

(1) Choose a finite set of basis functions $\{\phi_1,\phi_2,\cdots,\phi_m\}$ of the finite dimensional space of $H^1_\alpha(\Omega)$;

(2) Let $u_h=c_1\phi_1+c_2\phi_2+\cdots+c_m\phi_m$, $v_h=d_1\phi_1+d_2\phi_2+\cdots+d_m\phi_m$. Substitute the expression into (4.2) and choose $(\varphi_h,\psi_h)=(\phi_i,0),(0,\phi_i)$, $i=1,2,\cdots,m$ to get a system of linear equations;

(3) Solve the linear system for the coefficients $c_1,c_2,\cdots,c_m,d_1,d_2,\cdots,d_m$ and get the approximation of (u,v) in X_h.

More precisely, we have

$$\begin{aligned}&B(u_h,v_h;\phi_i,0)\\&=\int_{\Omega_b}\left[\frac{\omega\epsilon}{\kappa^2}\left(\sum_{j=1}^m c_j\nabla\phi_j\right)\cdot\nabla\overline{\phi}_i-\frac{\gamma}{\kappa^2}\left(\sum_{j=1}^m d_j\nabla\phi_j\right)\cdot\nabla^\perp\overline{\phi}_i-\omega\epsilon\left(\sum_{j=1}^m c_j\phi_j\right)\overline{\phi}_i\right]dx\\&\quad+\int_\Gamma\frac{i}{\lambda}\left(\sum_{j=1}^m c_j\phi_j\right)\overline{\phi}_i ds+\int_{\Gamma_b}T\begin{pmatrix}\sum_{j=1}^m c_j\phi_j\\\sum_{j=1}^m d_j\phi_j\end{pmatrix}\cdot\begin{pmatrix}\overline{\phi}_i\\0\end{pmatrix}ds,\\&B(u_h,v_h;0,\phi_i)\\&=\int_{\Omega_b}\left[\frac{\omega\mu}{\kappa^2}\left(\sum_{j=1}^m d_j\nabla\phi_j\right)\cdot\nabla\overline{\phi}_i+\frac{\gamma}{\kappa^2}\left(\sum_{j=1}^m c_j\nabla\phi_j\right)\cdot\nabla^\perp\overline{\phi}_i-\omega\mu\left(\sum_{j=1}^m d_j\phi_j\right)\overline{\phi}_i\right]dx\end{aligned}$$

$$+\int_\Gamma i\lambda\left(\sum_{j=1}^m d_j\phi_j\right)\overline{\phi}_i\,ds+\int_{\Gamma_b} T\begin{pmatrix}\sum_{j=1}^m c_j\phi_j\\ \sum_{j=1}^m d_j\phi_j\end{pmatrix}\cdot\begin{pmatrix}0\\ \overline{\phi}_i\end{pmatrix}ds.$$

In order to deduce the stiffness matrix, we need to define the following inner product

$$\langle f,g\rangle_{\Omega_b}=\int_{\Omega_b} f\bar{g}\,dx,\quad \langle f,g\rangle_\Gamma=\int_\Gamma f\bar{g}\,ds,\quad \langle f,g\rangle_{\Gamma_b}=\int_{\Gamma_b} f\bar{g}\,ds.$$

Let $\mathcal{B}\in\mathbb{C}^{2m\times 2m}$ be the stiffness matrix with the entries

$$B_{ij}=\begin{cases}\frac{\omega\epsilon}{\kappa^2}\langle\nabla\phi_j,\nabla\phi_i\rangle_{\Omega_b}-\omega\epsilon\langle\phi_j,\phi_i\rangle_{\Omega_b}\\ \quad+\frac{i}{\lambda}\langle\phi_j,\phi_i\rangle_\Gamma+\left\langle T\begin{pmatrix}\phi_j\\0\end{pmatrix},\begin{pmatrix}\phi_i\\0\end{pmatrix}\right\rangle_{\Gamma_b}, & 1\le i,j\le m,\\ \frac{\gamma}{\kappa^2}\langle\nabla\phi_{j-m},\nabla^\perp\phi_i\rangle_{\Omega_b}+\left\langle T\begin{pmatrix}0\\\phi_{j-m}\end{pmatrix},\begin{pmatrix}\phi_i\\0\end{pmatrix}\right\rangle_{\Gamma_b}, & 1\le i\le m,m+1\le j\le 2m,\\ \frac{\gamma}{\kappa^2}\langle\nabla\phi_j,\nabla^\perp\phi_{i-m}\rangle_{\Omega_b}+\left\langle T\begin{pmatrix}\phi_j\\0\end{pmatrix},\begin{pmatrix}0\\\phi_{i-m}\end{pmatrix}\right\rangle_{\Gamma_b}, & m+1\le i\le 2m,1\le j\le m,\\ \frac{\omega\mu}{\kappa^2}\langle\nabla\phi_{j-m},\nabla\phi_{i-m}\rangle_{\Omega_b}-\omega\mu\langle\phi_{j-m},\phi_{i-m}\rangle_{\Omega_b}\\ \quad+\frac{i}{\lambda}\langle\phi_{j-m},\phi_{i-m}\rangle_\Gamma+\left\langle T\begin{pmatrix}0\\\phi_{j-m}\end{pmatrix},\begin{pmatrix}0\\\phi_{i-m}\end{pmatrix}\right\rangle_{\Gamma_b}, & m+1\le i,j\le 2m,\end{cases}$$

and let $F\in\mathbb{C}^{2m}$ be a vector whose components are given by

$$F_i=\begin{cases}-\frac{2i\omega\epsilon\beta e^{-i\beta b}}{\kappa^2}\int_{\Gamma_b}\epsilon p_3\overline{\phi}_i e^{i\alpha x_1}\,ds, & 1\le i\le m,\\ -\frac{2i\omega\epsilon\beta e^{-i\beta b}}{\kappa^2}\int_{\Gamma_b}\mu q_3\overline{\phi}_{i-m} e^{i\alpha x_1}\,ds, & m+1\le i\le 2m.\end{cases}$$

Then we get the system of linear equations

$$\sum_{j=1}^{2m} B_{ij}a_j=F_i,\quad 1\le i\le 2m. \tag{4.3}$$

Having obtained $\{a_j\}_{j=1}^{2m}$ from (4.3), we can get u_h and v_h by setting $c_j=a_j$, $d_j=a_{j+m}$ for $1\le j\le m$.

Below we prove the well-posedness of the finite element approximation problem (4.2) and derive an error estimate of the finite element solution. We shall follow the approach outlined in [9, Chapter 4] for the scalar Helmholtz equation. It is worthy noting that, the real part of the DtN map for the conical diffraction problem is not positively definite (due to the presence of the index set $\mathcal{A}$; see Lemma 3.4 (ii)), which is in big contrast to the scalar case and brings essential difficulties in estimating the error estimate.

Denote $e_h:=(u-u_h,v-v_h)^\top$. It is obvious that e_h is α-quasiperiodic. Define the projection operator $P:L^2(\Gamma_b)^2\to L^2(\Gamma_b)^2$ by

$$(Pf)(x_1)=\sum_{n\in\mathcal{A}} f_n e^{i\alpha_n x_1},\quad f=\sum_{n\in\mathbb{Z}} f_n e^{i\alpha_n x_1}\in L^2(\Gamma_b)^2,$$

where the set $\mathcal{A}$ is defined by (3.24).

LEMMA 4.1. *There exists a constant $h_1 \in (0,1)$ such that for $h \in (0,h_1)$ the following estimate holds:*

$$\|e_h\|^2_{H^1(\Omega_b)^2} \le C\left(h^{2\rho-2}\|(u,v)\|^2_{H^\rho(\Omega_b)^2} + \|e_h\|^2_{L^2(\Omega_b)^2} + \|Pe_h\|^2_{L^2(\Gamma_b)^2}\right),$$

where the constant C depends on ρ but is independent of h and (u,v).

Proof. It follows from the sesquilinear form (4.2) that

$$\begin{aligned} B(e_h;e_h) := \int_{\Omega_b} \Big[& \frac{\omega\epsilon}{\kappa^2}|\nabla(u-u_h)|^2 - \frac{\gamma}{\kappa^2}\nabla(v-v_h)\cdot\nabla^\perp\overline{(u-u_h)} - \omega\epsilon|u-u_h|^2 \\ & + \frac{\omega\mu}{\kappa^2}|\nabla(v-v_h)|^2 + \frac{\gamma}{\kappa^2}\nabla(u-u_h)\cdot\nabla^\perp\overline{(v-v_h)} - \omega\mu|v-v_h|^2\Big]\,dx \\ & + \int_\Gamma \frac{i}{\lambda}|u-u_h|^2 + i\lambda|v-v_h|^2\,ds + \int_{\Gamma_b} Te_h\cdot\overline{e_h}\,ds. \end{aligned} \tag{4.4}$$

Multiplying both sides of (4.4) by $\theta = \frac{i+\delta}{|i+\delta|}$ and taking the real part, we get

$$\begin{aligned} \mathrm{Re}[\theta B(e_h;e_h)] = \mathrm{Re}\Big\{ \theta \int_{\Omega_b} \Big[& \frac{\omega\epsilon}{\kappa^2}|\nabla(u-u_h)|^2 - \frac{\gamma}{\kappa^2}\nabla(v-v_h)\cdot\nabla^\perp\overline{(u-u_h)} - \omega\epsilon|u-u_h|^2 \\ & + \frac{\omega\mu}{\kappa^2}|\nabla(v-v_h)|^2 + \frac{\gamma}{\kappa^2}\nabla(u-u_h)\cdot\nabla^\perp\overline{(v-v_h)} - \omega\mu|v-v_h|^2\Big]\,dx\Big\} \\ & + \mathrm{Re}\left\{\theta\int_\Gamma \frac{i}{\lambda}|u-u_h|^2 + i\lambda|v-v_h|^2\,ds\right\} + \left\{\mathrm{Re}\,\theta\int_{\Gamma_b} Te_h\cdot\overline{e_h}\,ds\right\}. \end{aligned} \tag{4.5}$$

From the strongly elliptic analysis (see the proof of Theorem 3.2), we can easily get

$$\mathrm{Re}\left\{\theta\int_\Gamma \frac{i}{\lambda}|u-u_h|^2 + i\lambda|v-v_h|^2\,ds\right\} \ge 0,$$

and

$$\mathrm{Re}\left\{\theta\int_{\Gamma_b} Te_h\cdot\overline{e_h}\,ds + q(e_h;e_h)\right\} \ge 0,$$

where the compact form q is defined by (3.26), that is, for $e_h = \sum_{n\in\mathbb{Z}} A_n e^{i\alpha_n x_1}$, we have

$$q(e_h;e_h) = 2\pi\mathrm{Re}\sum_{n\in\mathcal{A}} \theta M_n A_n\cdot\overline{A_n} \le C\|Pe_h\|^2_{L^2(\Gamma_b)^2}.$$

Therefore, by (4.5),

$$\begin{aligned} & \mathrm{Re}\left\{\theta\int_{\Omega_b}\Big[\frac{\omega\epsilon}{\kappa^2}|\nabla(u-u_h)|^2 - \frac{\gamma}{\kappa^2}\nabla(v-v_h)\cdot\nabla^\perp\overline{(u-u_h)} \right. \\ & \quad \left. + \frac{\omega\mu}{\kappa^2}|\nabla(v-v_h)|^2 + \frac{\gamma}{\kappa^2}\nabla(u-u_h)\cdot\nabla^\perp\overline{(v-v_h)}\Big]\,dx\right\} \\ & = \mathrm{Re}(\theta B(e_h;e_h)) - \mathrm{Re}\left\{\theta\int_\Gamma \frac{i}{\lambda}|u-u_h|^2 + i\lambda|v-v_h|^2\,ds\right\} - \mathrm{Re}\left\{\theta\int_{\Gamma_b} Te_h\cdot\overline{e_h}\,ds\right\} \\ & \quad + \mathrm{Re}\left\{\theta\int_{\Omega_b}\omega\epsilon|u-u_h|^2 + \omega\mu|v-v_h|^2\,dx\right\} \\ & \le \mathrm{Re}(\theta B(e_h;e_h)) + \mathrm{Re}\left\{\theta\int_{\Omega_b}\omega\epsilon|u-u_h|^2 + \omega\mu|v-v_h|^2\,dx\right\} + q(e_h;e_h). \end{aligned}$$

Using Lemma 3.4 (i) we get

$$C_1\|e_h\|^2_{H^1(\Omega_b)^2} \le |B(e_h;e_h)| + C_2\|e_h\|^2_{L^2(\Omega_b)^2} + C\|Pe_h\|^2_{L^2(\Gamma_b)^2}. \tag{4.6}$$

Observing for any $(\xi,\eta)\in X_h$ that

$$B(u,v;\xi-u_h,\eta-v_h)=F(\xi-u_h,\eta-v_h), \quad B(u_h,v_h;\xi-u_h,\eta-v_h)=F(\xi-u_h,\eta-v_h),$$

we obtain

$$B(u-u_h,v-v_h;\xi-u_h,\eta-v_h)=0. \tag{4.7}$$

Therefore for any $(\xi,\eta)\in X_h$, we have

$$B(u-u_h,v-v_h;u-u_h,v-v_h)=B(u-u_h,v-v_h;u-\xi,v-\eta). \tag{4.8}$$

Since X_h is finite-dimensional, it is complete and therefore closed. Hence, the infimum in (4.1) is actually attained for $(\varphi,\psi)=(u,v)$ in (4.1). For any small positive constants $\epsilon_i\,(i=1,2,3,4)$, it follows from (4.8) and Young's inequality that

$$\begin{aligned}
|B(e_h;e_h)| &= |B(u-u_h,v-v_h;u-\xi,v-\eta)| \\
&= \Big| \int_\Gamma \frac{i}{\lambda}(u-u_h)\overline{(u-\xi)} + i\lambda(v-v_h)\overline{(v-\eta)}\,ds + \int_{\Omega_b} \frac{\omega\epsilon}{\kappa^2}\nabla(u-u_h)\cdot\nabla\overline{(u-\xi)} \\
&\quad -\frac{\gamma}{\kappa^2}\nabla(v-v_h)\cdot\nabla^\perp\overline{(v-\eta)} - \omega\epsilon(u-u_h)\overline{(u-\xi)} + \frac{\omega\mu}{\kappa^2}\nabla(v-v_h)\cdot\nabla\overline{(v-\eta)} \\
&\quad +\frac{\gamma}{\kappa^2}\nabla(u-u_h)\cdot\nabla^\perp\overline{(v-\eta)} - \omega\mu(v-v_h)\overline{(v-\eta)}\,dx \\
&\quad + \int_{\Gamma_b} T\begin{pmatrix} u-u_h \\ v-v_h\end{pmatrix}\cdot\begin{pmatrix}\overline{u-\xi} \\ \overline{v-\eta}\end{pmatrix} ds\Big| \\
&\le \frac{1}{|\lambda|}\left(h\|u-u_h\|^2_{L^2(\Gamma)} + \frac{1}{h}\|u-\xi\|^2_{L^2(\Gamma)}\right) + |\lambda|\left(h\|v-v_h\|^2_{L^2(\Gamma)} + \frac{1}{h}\|v-\eta\|^2_{L^2(\Gamma)}\right) \\
&\quad +\frac{\omega\epsilon}{\kappa^2}\left(\epsilon_1\|\nabla u-\nabla u_h\|^2_{L^2(\Omega_b)} + \frac{1}{4\epsilon_1}\|\nabla u-\nabla\xi\|^2_{L^2(\Omega_b)}\right) \\
&\quad +\frac{|\gamma|}{\kappa^2}\left(\epsilon_2\|\nabla v-\nabla v_h\|^2_{L^2(\Omega_b)} + \frac{1}{4\epsilon_2}\|\nabla u-\nabla\xi\|^2_{L^2(\Omega_b)}\right) \\
&\quad +\omega\epsilon\left(h^2\|u-u_h\|^2_{L^2(\Omega_b)} + h^{-2}\|u-\xi\|^2_{L^2(\Omega_b)}\right) \\
&\quad +\frac{\omega\mu}{\kappa^2}\left(\epsilon_3\|\nabla v-\nabla v_h\|^2_{L^2(\Omega_b)} + \frac{1}{4\epsilon_3}\|\nabla v-\nabla\eta\|^2_{L^2(\Omega_b)}\right) \\
&\quad +\frac{|\gamma|}{\kappa^2}\left(\epsilon_4\|\nabla u-\nabla u_h\|^2_{L^2(\Omega_b)} + \frac{1}{4\epsilon_4}\|\nabla v-\nabla\eta\|^2_{L^2(\Omega_b)}\right) \\
&\quad +\omega\mu\left(h^2\|v-v_h\|^2_{L^2(\Omega_b)} + h^{-2}\|v-\eta\|^2_{L^2(\Omega_b)}\right) \\
&\quad +\left|\int_{\Gamma_b} T\begin{pmatrix} u-u_h \\ v-v_h\end{pmatrix}\cdot\begin{pmatrix}\overline{u-\xi} \\ \overline{v-\eta}\end{pmatrix} ds\right|.
\end{aligned} \tag{4.9}$$

Using the continuity of the DtN map T (see Lemma 3.2), trace theorem and Young's inequality, we have

$$\left|\int_{\Gamma_b} T\begin{pmatrix} u-u_h \\ v-v_h\end{pmatrix}\cdot\begin{pmatrix}\overline{u-\xi} \\ \overline{v-\eta}\end{pmatrix} ds\right| \le \left\|T\begin{pmatrix} u-u_h \\ v-v_h\end{pmatrix}\right\|_{H^{-1/2}(\Gamma_b)^2}\left\|\begin{pmatrix} u-\xi \\ v-\eta\end{pmatrix}\right\|_{H^{1/2}(\Gamma_b)^2}$$

$$\leq C\|e_h\|_{H^{1/2}(\Gamma_b)^2}\left\|\begin{pmatrix}u-\xi\\ v-\eta\end{pmatrix}\right\|_{H^{1/2}(\Gamma_b)^2}$$
$$\leq C\left(\epsilon_5\|e_h\|^2_{H^1(\Omega_b)^2}+\frac{1}{4\epsilon_5}\left\|\begin{pmatrix}u-\xi\\ v-\eta\end{pmatrix}\right\|_{H^{1/2}(\Gamma_b)^2}\right). \quad (4.10)$$

One deduces from (4.1) and (4.9) - (4.10) that

$$|B(e_h;e_h)|\leq Ch\|e_h\|^2_{L^2(\Gamma)^2}+\sigma\|e_h\|^2_{H^1(\Omega_b)^2}$$
$$+C_1h^2\|e_h\|^2_{L^2(\Omega_b)^2}+C(\sigma)h^{2\rho-2}\|(u,v)\|^2_{H^\rho(\Omega_b)^2}, \quad (4.11)$$

where $\sigma=\sigma(\epsilon_1,\cdots,\epsilon_5)>0$. Combining (4.6) and (4.11) leads to

$$C_1\|e_h\|^2_{H^1(\Omega_b)^2}\leq|B(e_h;e_h)|+C_2\|e_h\|^2_{L^2(\Omega_b)^2}+|q(e_h;e_h)|$$
$$\leq C_3h\|e_h\|^2_{L^2(\Gamma)^2}+\sigma\|e_h\|^2_{H^1(\Omega_b)^2}+C_4h^2\|e_h\|^2_{L^2(\Omega_b)^2}$$
$$+C(\sigma)h^{2\rho-2}\|(u,v)\|^2_{H^\rho(\Omega_b)^2}+C_2\|e_h\|^2_{L^2(\Omega_b)^2}+C_5\|Pe_h\|^2_{L^2(\Gamma_b)^2}. \quad (4.12)$$

Using the estimate

$$\|e_h\|_{L^2(\Gamma)^2}\leq\|e_h\|_{H^{1/2}(\Gamma)^2}\leq C\|e_h\|_{H^1(\Omega_b)^2},\quad C>0,$$

we get from (4.12) that

$$C_1\|e_h\|^2_{H^1(\Omega_b)^2}\leq(CC_3h+\sigma+C_4h^2)\|e_h\|^2_{H^1(\Omega_b)^2}+C(\sigma)h^{2\rho-2}\|(u,v)\|^2_{H^\rho(\Omega_b)^2}$$
$$+C_2\|e_h\|^2_{L^2(\Omega_b)^2}+C_5\|Pe_h\|^2_{L^2(\Gamma_b)^2}.$$

Now choose σ sufficienlty small and let h_1 be a constant such that $\sigma+CC_3h_1+C_4h_1^2<C_1$. Then we obtain the desired estimate of this lemma for all $h\in(0,h_1)$. ∎

We next estimate the L^2-norm of e_h in Ω_b.

LEMMA 4.2. *There exists a constant $h_2\in(0,1)$ such that*

$$\|e_h\|_{L^2(\Omega_b)^2}\leq C(h+C_1h^{3/2})\|e_h\|_{H^1(\Omega_b)^2}\quad \textit{for all } h\in(0,h_2),$$

where the constants C,C_1 depend on ρ but are independent of h and (u,v).

Proof. We use the duality argument. By the definition

$$\|e_h\|_{L^2(\Omega_b)^2}=\sup_{(\phi,\zeta)\in C_0^\infty(\Omega_b)^2}\frac{(e_h,(\phi,\zeta))_{L^2(\Omega_b)^2}}{\|(\phi,\zeta)\|_{L^2(\Omega_b)^2}}, \quad (4.13)$$

where

$$(e_h,(\phi,\zeta))_{L^2(\Omega_b)^2}:=\int_{\Omega_b}\frac{\omega\epsilon}{\kappa^2}(u-u_h)\bar\phi+\frac{\omega\mu}{\kappa^2}(v-v_h)\bar\zeta\,dx. \quad (4.14)$$

Consider a quasi-periodic solution (w,z) of the following boundary value problem:

$$\begin{cases}\Delta w+k^2w=-\bar\phi & \text{in}\quad \Omega_b,\\ \Delta z+k^2z=-\bar\zeta & \text{in}\quad \Omega_b,\\ \lambda\partial_n w-\frac{i\kappa^2}{\omega\epsilon}w+\frac{\lambda\gamma}{\omega\epsilon}\partial_\tau z=0 & \text{on}\quad \Gamma,\\ \partial_n z-\frac{i\lambda\kappa^2}{\omega\mu}z-\frac{\lambda}{\omega\mu}\partial_\tau w=0 & \text{on}\quad \Gamma,\\ T^*\begin{pmatrix}w\\ z\end{pmatrix}=\sum_{n\in\mathbb{Z}}M_n^*\begin{pmatrix}\hat w_n\\ \hat z_n\end{pmatrix}e^{i\alpha_nx_1} & \text{on}\quad \Gamma_b,\end{cases} \quad (4.15)$$

where T^* is the adjoint operator of T. We can easily get the variational formulation of (4.15) that for all $(\varphi,\psi)\in X$,

$$\int_{\Omega_b}\Big[\frac{\omega\epsilon}{\kappa^2}\nabla w\cdot\nabla\varphi-\frac{\gamma}{\kappa^2}\nabla z\cdot\nabla^\perp\varphi-\omega\epsilon w\varphi+\frac{\omega\mu}{\kappa^2}\nabla z\cdot\nabla\psi+\frac{\gamma}{\kappa^2}\nabla w\cdot\nabla^\perp\psi-\omega\mu z\psi\Big]\,dx$$
$$+\int_\Gamma\frac{i}{\lambda}w\varphi+i\lambda z\psi\,ds-\int_{\Gamma_b}T^*\begin{pmatrix}w\\z\end{pmatrix}\cdot\begin{pmatrix}\varphi\\\psi\end{pmatrix}ds=\int_{\Omega_b}\frac{\omega\epsilon}{\kappa^2}\bar\phi\varphi+\frac{\omega\mu}{\kappa^2}\bar\zeta\psi. \tag{4.16}$$

Taking $\varphi=u-u_h,\psi=v-v_h$ in (4.16), using the definition of $(e_h,(\phi,\zeta))_{L^2(\Omega_b)^2}$ in (4.14) and recalling the form $B(u,v;\varphi,\psi)$ in (3.15), we get

$$B(u-u_h,v-v_h;\bar w,\bar z)=(e_h,(\phi,\zeta))_{L^2(\Omega_b)^2}. \tag{4.17}$$

The well-posedness of the problem (4.15) can be established by the same argument as the proof for the variational problem (3.14). Moreover, we have

$$\|(w,z)\|_{H^2(\Omega_b)^2}\le C\|(\phi,\zeta)\|_{L^2(\Omega_b)^2}. \tag{4.18}$$

Using the orthogonal formula (4.7), we have

$$\begin{aligned}B(u-u_h,v-v_h;\bar w-\xi,\bar z-\eta)&=B(u-u_h,v-v_h;\bar w,\bar z)-B(u-u_h,v-v_h;\xi,\eta)\\&=B(u-u_h,v-v_h;\bar w,\bar z).\end{aligned} \tag{4.19}$$

Combining (4.17) and (4.19) gives

$$|(e_h,(\phi,\zeta))_{L^2(\Omega_b)^2}|=|B(e_h,(\bar w,\bar z))|=|B(e_h,(\bar w,\bar z)-(\xi,\eta))|\ \text{for all}\ (\xi,\eta)\in X_h. \tag{4.20}$$

In particular, (ξ,η) can be chosen in such a way that the infimum is attained for $(\varphi,\psi)=(w,z)$ in (4.1). By arguing analogously to the proof of Lemma 4.1, we deduce from (4.9)–(4.10) that

$$\begin{aligned}&|B(e_h,(\bar w,\bar z)-(\xi,\eta))|\\&=\Big|\int_\Gamma\frac{i}{\lambda}(u-u_h)\overline{(\bar w-\xi)}+i\lambda(v-v_h)\overline{(\bar z-\eta)}\,ds+\int_{\Omega_b}\frac{\omega\epsilon}{\kappa^2}\nabla(u-u_h)\cdot\nabla\overline{(\bar w-\xi)}\\&\quad-\frac{\gamma}{\kappa^2}\nabla(v-v_h)\cdot\nabla^\perp\overline{(\bar w-\xi)}-\omega\epsilon(u-u_h)\overline{(\bar w-\xi)}+\frac{\omega\mu}{\kappa^2}\nabla(v-v_h)\cdot\nabla\overline{(\bar z-\eta)}\\&\quad+\frac{\gamma}{\kappa^2}\nabla(u-u_h)\cdot\nabla^\perp\overline{(\bar z-\eta)}-\omega\mu(v-v_h)\overline{(\bar z-\eta)}\,dx\\&\quad+\int_{\Gamma_b}T\begin{pmatrix}u-u_h\\v-v_h\end{pmatrix}\cdot\begin{pmatrix}\overline{\bar w-\xi}\\\overline{\bar z-\eta}\end{pmatrix}ds\Big|.\end{aligned}$$

Using the Cauchy-Schwarz inequality, we continue to estimate the above equation by

$$\begin{aligned}&|B(e_h,(\bar w,\bar z)-(\xi,\eta))|\\&\le\tfrac{1}{|\lambda|}\big(h^{-1/2}\|u-u_h\|_{L^2(\Gamma)}h^{1/2}\|\bar w-\xi\|_{L^2(\Gamma)}\big)+|\lambda|\big(h^{-1/2}\|v-v_h\|_{L^2(\Gamma)}h^{1/2}\|\bar z-\eta\|_{L^2(\Gamma)}\big)\\&\quad+\tfrac{\omega\epsilon}{\kappa^2}\big(\tfrac{1}{h}\|\nabla u-\nabla u_h\|_{L^2(\Omega_b)}h|\bar w-\xi|_{H^1(\Omega_b)}\big)+\tfrac{|\gamma|}{\kappa^2}\big(\tfrac{1}{h}\|\nabla v-\nabla v_h\|_{L^2(\Omega_b)}h|\bar w-\xi|_{H^1(\Omega_b)}\big)\\&\quad+\omega\epsilon\big(\|u-u_h\|_{L^2(\Omega_b)}\|\bar w-\xi\|_{L^2(\Omega_b)}\big)+\frac{\omega\mu}{\kappa^2}\left(\frac{1}{h}\|\nabla v-\nabla v_h\|_{L^2(\Omega_b)}h|\bar z-\eta|_{H^1(\Omega_b)}\right)\\&\quad+\frac{|\gamma|}{\kappa^2}\left(\frac{1}{h}\|\nabla u-\nabla u_h\|_{L^2(\Omega_b)}h|\bar z-\eta|_{H^1(\Omega_b)}\right)+\omega\mu\big(\|v-v_h\|_{L^2(\Omega_b)}\|\bar z-\eta\|_{L^2(\Omega_b)}\big)\end{aligned}$$

$$+\frac{1}{h}\|e_h\|_{H^1(\Omega_b)^2}h\left\|\begin{pmatrix}u-\xi\\ v-\eta\end{pmatrix}\right\|_{H^{1/2}(\Gamma_b)^2}$$
$$\le C_1h\|(w,z)\|_{H^2(\Omega_b)^2}\left[h^{1/2}\|e_h\|_{H^1(\Omega_b)^2}+C_2\|\nabla e_h\|_{L^2(\Omega_b)^2}\right.$$
$$\left.+C_3h\|e_h\|_{L^2(\Omega_b)^2}+C_4\|e_h\|_{H^1(\Omega_b)^2}\right], \tag{4.21}$$

where the constants $C_j\,(j=1,2,3,4)$ depend on k but are independent of h. Combining (4.13) and (4.18), (4.20) and (4.21), we can find a positive constant $h_2\le 1$ such that for all $h\in(0,h_2)$,

$$\|e_h\|_{L^2(\Omega_b)^2}=\sup_{(\phi,\zeta)\in C_0^\infty(\Omega_b)^2}\frac{(e_h,(\phi,\zeta))_{L^2(\Omega_b)^2}}{\|(\phi,\zeta)\|_{L^2(\Omega_b)^2}}$$
$$\le\sup_{(\phi,\zeta)\in C_0^\infty(\Omega_b)^2}\frac{C_1h\|(w,s)\|_{H^2(\Omega_b)^2}\left[h^{1/2}\|e_h\|_{H^1(\Omega_b)^2}+C_2\|e_h\|_{H^1(\Omega_b)^2}+C_3h\|e_h\|_{L^2(\Omega_b)^2}\right]}{\|(\phi,\zeta)\|_{L^2(\Omega_b)^2}}$$
$$\le\sup_{(\phi,\zeta)\in C_0^\infty(\Omega_b)^2}\frac{C_1h\|(\phi,\zeta)\|_{L^2(\Omega_b)^2}\left[h^{1/2}\|e_h\|_{H^1(\Omega_b)^2}+C_2\|e_h\|_{H^1(\Omega_b)^2}+C_3h\|e_h\|_{L^2(\Omega_b)^2}\right]}{\|(\phi,\zeta)\|_{L^2(\Omega_b)^2}}$$
$$=Ch\|e_h\|_{H^1(\Omega_b)^2}+C_1h^2\|e_h\|_{L^2(\Omega_b)^2}+C_2h^{3/2}\|e_h\|_{H^1(\Omega_b)^2}.$$

Now let h_2 be a positive constant such that $1-C_1h_2^2>0$. We then obtain the estimate of this lemma for all $h\in(0,h_2)$. ∎

We proceed with the estimate of the L^2-norm of Pe_h on Γ_b.

LEMMA 4.3. *There exists a constant $C>0$ such that*

$$\|Pe_h\|_{L^2(\Gamma_b)^2}\le C\|e_h\|_{L^2(\Omega_b)^2},$$

where the positive constant C is independent of h and (u,v).

Proof. Define

$$D=\{(x_1,x_2)\in\mathbb{R}^2:0<x_1<2\pi,b-\epsilon<x_2<b\},$$

where the constant $\epsilon>0$ is chosen to satisfy $b-\epsilon>\max_{x\in\Gamma}\{x_2\}$. Suppose

$$e_h(x)=\sum_{n\in\mathbb{Z}}A_ne^{i(\alpha_nx_1+\beta_nx_2)},\quad x\in D.$$

Then $P(e_h|_{\Gamma_b})=\sum_{n\in\mathcal{A}}A_ne^{i(\alpha_nx_1+\beta_nb)}$. Direct calculations show that

$$\|e_h\|^2_{L^2(\Omega_b)^2}\ge\|e_h\|^2_{L^2(D)^2}=\int_0^{2\pi}\int_{b-\epsilon}^{b}|e_h|^2\,dx_2dx_1$$
$$=\int_0^{2\pi}\int_{b-\epsilon}^{b}\sum_{n\in\mathbb{Z}}A_ne^{i(\alpha_nx_1+\beta_nx_2)}\cdot\sum_{m\in\mathbb{Z}}\overline{A_m}e^{-i\alpha_mx_1-i\overline{\beta_m}x_2}\,dx_2dx_1$$
$$=2\pi\sum_{n\in\mathbb{Z}}|A_n|^2\int_{b-\epsilon}^{b}e^{i(\beta_n-\overline{\beta_n})x_2}\,dx_2$$
$$=2\pi\left(\sum_{|\alpha_n|\le\kappa}\epsilon|A_n|^2+\sum_{|\alpha_n|>\kappa}|A_n|^2C_n\right),$$

where $C_n = -\frac{\epsilon}{2|\beta_n|}\left(e^{-2|\beta_n|b} - e^{-2|\beta_n|(b-1)}\right) > 0$. By (3.24) and the proof of Theorem 3.4, we can easily get that $\mathcal{A}$ coincides with the set $\{n \in \mathbb{Z} : \kappa \le |\alpha_n| < k\}$. Then we have

$$\mathcal{A} \subset \mathcal{C} := \{n \in \mathbb{Z} : |\alpha_n| \ge \kappa\}.$$

Therefore,

$$2\pi\left(\sum_{|\alpha_n|\le\kappa} \epsilon|A_n|^2 + \sum_{|\alpha_n|>\kappa} |A_n|^2 C_n\right) \ge 2\pi C\left(\sum_{n\in\mathcal{A}} |A_n|^2\right) = C\|Pe_h\|^2_{L^2(\Gamma_b)^2}$$

where $C = \min\{\epsilon, \min_{n\in\mathcal{A}} C_n\}$. ∎

The main result of this section is stated below.

THEOREM 4.1. *Suppose that $(u,v) \in H^\rho(\Omega_b)^2, \rho \ge 2$, satisfies the variational problem (3.14). Suppose also that the family of finite element spaces $\{X_h\}$ satisfies the assumption (4.1). Then there exists $h_0 \in (0,1)$ such that for $h \in (0,h_0)$, the problem (4.2) admits a unique solution (u_h, v_h) with the estimates*

$$\|(u,v) - (u_h,v_h)\|_{L^2(\Omega_b)^2} \le C\left(h^\rho + C_1 h^{\rho+1/2}\right)\|(u,v)\|_{H^\rho(\Omega_b)^2},$$
$$\|(u,v) - (u_h,v_h)\|_{H^1(\Omega_b)^2} \le Ch^{\rho-1}\|(u,v)\|_{H^\rho(\Omega_b)^2},$$

where the positive constant C depends on ρ but is independent of h and (u,v).

Proof. Let h_1 and h_2 be specified as in Lemmas 4.1 and 4.2 and set $h_0 = \min\{h_1,h_2\}$. For $h \in (0,h_0)$, we deduce from Lemmas 4.1–4.3 that

$$\begin{aligned}\|e_h\|^2_{H^1(\Omega_b)^2} &\le C\left(h^{2\rho-2}\|(u,v)\|^2_{H^\rho(\Omega_b)^2} + C\|e_h\|^2_{L^2(\Omega_b)^2}\right)\\ &\le C\left(h^{2\rho-2}\|(u,v)\|^2_{H^\rho(\Omega_b)^2} + C_1(h + C_2h^{3/2})^2\|e_h\|^2_{H^1(\Omega_b)^2}\right).\end{aligned}$$

Now letting h_0 be a constant such that $1 - C_1\left(h_0 + C_2 h_0^{3/2}\right)^2 > 0$, we obtain

$$\|e_h\|_{H^1(\Omega_b)^2} \le Ch^{\rho-1}\|(u,v)\|_{H^\rho(\Omega_b)^2} \quad \text{for all } h \in (0,h_0).$$

Therefore, using Lemma 4.2,

$$\|e_h\|_{L^2(\Omega_b)^2} \le C(h + C_1h^{3/2})\|e_h\|_{H^1(\Omega_b)^2} \le C\left(h^\rho + C_1h^{\rho+1/2}\right)\|(u,v)\|_{H^\rho(\Omega_b)^2},$$

which completes the proof. ∎

5. Integral equation methods

The aim of this section is to develop an integral equation method for the conical diffraction problem (3.2). The case of transmission conditions has been investigated in [29] for coated gratings. Note that the analysis performed here differs from those in [26,30] for infinitely long cylinders. We make the following assumption.

Assumption (A): The grating profile Γ is the graph of some 2π-periodic function $x_2 = f(x_1)$, $x_1 \in \mathbb{R}$, where f is either C^2-smooth or piecewise linear with a finite number of corner points in one periodic cell.

Introduce the α-quasiperiodic fundamental solution to the Helmholtz equation $(\Delta+\kappa^2)u=0$ by

$$\begin{aligned}G(x,y)&=\frac{i}{4}\sum_{n\in\mathbb{Z}}\exp(-i\alpha 2\pi n)H_0^{(1)}\left(k\sqrt{(x_1+2n\pi-y_1)^2+(x_2-y_2)^2}\right)\\&=\frac{i}{4\pi}\sum_{n\in\mathbb{Z}}\frac{1}{\beta_n}\exp\left(i\alpha_n(x_1-y_1)+i\beta_n|x_2-y_2|\right)\end{aligned}$$

for $x-y\neq n(2\pi,0)$, with $H_0^{(1)}(t)$ being the first kind Hankel function of order zero. Define the single-layer potential by

$$(\mathcal{S}g)(x)=2\int_\Gamma G(x,y)g(y)ds(y),\quad x\in\Omega,$$

with the density g. We make the ansatz for the solution (u^s,v^s) in the form

$$u^s=\mathcal{S}g_1,\quad v^s=\mathcal{S}g_2.$$

Further, define the single- and double-layer operators S and K by

$$\begin{aligned}(S\rho)(x)&:=2\int_\Gamma G(x,y)\rho(y)ds(y),\quad x\in\Gamma,\\(K\rho)(x)&:=2\int_\Gamma\frac{\partial G(x,y)}{\partial\nu(y)}\rho(y)ds(y),\quad x\in\Gamma,\end{aligned}$$

and the normal and tangential derivative operators K' and H' by

$$\begin{aligned}(K'\rho)(x)&:=2\int_\Gamma\frac{\partial G(x,y)}{\partial\nu(x)}\rho(y)ds(y),\quad x\in\Gamma,\\(H'\rho)(x)&:=2\int_\Gamma\frac{\partial G(x,y)}{\partial\tau(x)}\rho(y)ds(y),\quad x\in\Gamma,\end{aligned}$$

where ν denotes the unit normal vector to the boundary Γ directed into the exterior of Ω and τ denotes the unit tangential vector to Γ.

LEMMA 5.1. *Let g_1,g_2 be the density functions of u^s,v^s, respectively. Then the following jump relations hold*

$$\begin{aligned}&u^s(x)=2\int_\Gamma G(x,y)g_1(y)ds(y)=Sg_1,\quad x\in\Gamma,\\&v^s(x)=2\int_\Gamma G(x,y)g_2(y)ds(y)=Sg_2,\quad x\in\Gamma,\\&\frac{\partial u^s_\pm}{\partial\nu}(x)=2\int_\Gamma\frac{\partial G(x,y)}{\partial\nu(x)}g_1(y)ds(y)\pm g_1(x)=K'g_1(x)\pm g_1(x),\quad x\in\Gamma,\\&\frac{\partial v^s_\pm}{\partial\nu}(x)=2\int_\Gamma\frac{\partial G(x,y)}{\partial\nu(x)}g_2(y)ds(y)\pm g_2(x)=K'g_2(x)\pm g_2(x),\quad x\in\Gamma,\\&\frac{\partial u^s}{\partial\tau}(x)=2\int_\Gamma\frac{\partial G(x,y)}{\partial\tau(x)}g_1(y)ds(y)=H'g_1(x),\quad x\in\Gamma,\\&\frac{\partial v^s}{\partial\tau}(x)=2\int_\Gamma\frac{\partial G(x,y)}{\partial\tau(x)}g_2(y)ds(y)=H'g_2(x),\quad x\in\Gamma,\end{aligned}$$

where

$$\frac{\partial u^s_\pm}{\partial \nu}(x) := \lim_{h\to+0} \nu(x)\cdot\nabla u^s(x\pm h\nu(x)), \qquad \frac{\partial v^s_\pm}{\partial \nu}(x) := \lim_{h\to+0} \nu(x)\cdot\nabla v^s(x\pm h\nu(x)),$$
$$\frac{\partial u^s}{\partial \tau}(x) := \tau(x)\cdot\nabla u^s(x), \qquad \frac{\partial v^s}{\partial \tau}(x) := \tau(x)\cdot\nabla v^s(x).$$

Proof. Since the difference of the quasi-periodic fundamental function G and the free-space fundamental solution is of C^∞-smooth (see e.g. [24]), mapping properties of the single- and double-layer operators follow from standard approach; we refer to e.g., [3–5, 14, 24, 27, 29] for the details. Note that the continuity of $\frac{\partial u^s}{\partial \tau}$ (thus also for $\frac{\partial v^s}{\partial \tau}$) on Γ follows from the relation (see e.g., [25])

$$\lim_{x\to\Gamma^\pm} \nabla u^s(x) = 2\int_\Gamma \nabla_x G(x,y) g_1(y) ds(y) \pm g_1(y)\nu(y),$$

which implies that

$$\begin{aligned}\lim_{x\to\Gamma^\pm} \frac{\partial u^s}{\partial \tau}(x) &= \lim_{x\to\Gamma^\pm} \tau(x)\cdot\nabla u^s(x)\\ &= \tau(x)\cdot 2\int_\Gamma \nabla_x G(x,y) g_1(y) ds(y) \pm \tau(x)\cdot g_1(y)\nu(y)\\ &= 2\int_\Gamma \frac{\partial G(x,y)}{\partial \tau(x)} g_1(y) ds(y).\end{aligned}$$

□

From the above jump relations together with the boundary condition (2.7), one can derive on Γ that

$$\begin{aligned}&\lambda\frac{\partial u^s}{\partial \nu} + i\omega\mu\cos^2\phi\, u^s + \lambda\sin\phi\sqrt{\frac{\mu}{\epsilon}}\frac{\partial v^s}{\partial \tau}\\ &= \lambda(K' g_1 + g_1) + i\omega\mu\cos^2\phi S g_1 + \lambda\sin\phi\sqrt{\frac{\mu}{\epsilon}} H' g_2 = h_1, \qquad (5.1)\\ &\frac{\partial v^s}{\partial n} + i\lambda\omega\epsilon\cos^2\phi\, v^s - \sin\phi\sqrt{\frac{\epsilon}{\mu}}\frac{\partial u^s}{\partial \tau}\\ &= (K' g_2 + g_2) + i\lambda\omega\epsilon\cos^2\phi S g_2 - \sin\phi\sqrt{\frac{\epsilon}{\mu}} H' g_1 = h_2, \qquad (5.2)\end{aligned}$$

where

$$\begin{aligned}h_1 &:= -\left(\lambda\frac{\partial u^i}{\partial \nu} + i\omega\mu\cos^2\phi\, u^i + \lambda\sin\phi\sqrt{\frac{\mu}{\epsilon}}\frac{\partial v^i}{\partial \tau}\right),\\ h_2 &:= -\left(\frac{\partial v^i}{\partial \nu} + i\lambda\omega\epsilon\cos^2\phi\, v^i - \sin\phi\sqrt{\frac{\epsilon}{\mu}}\frac{\partial u^i}{\partial \tau}\right).\end{aligned}$$

Combining (5.1) with (5.2), we obtain the integral equations

$$\begin{pmatrix} \lambda(K'+I) & \lambda\sin\phi\sqrt{\mu/\epsilon}H' \\ -\sin\phi\sqrt{\epsilon/\mu}H' & K'+I \end{pmatrix}\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} + \begin{pmatrix} i\omega\mu\cos^2\phi S & 0 \\ 0 & i\lambda\omega\epsilon\cos^2\phi S \end{pmatrix}\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}.$$

Therefore, an equivalent system to our conical diffraction problem is

$$Ag + Bg := \begin{pmatrix} \lambda(K'+I) & dH' \\ -cH' & K'+I \end{pmatrix} g + \begin{pmatrix} iaS & 0 \\ 0 & ibS \end{pmatrix} g = h, \qquad (5.3)$$

with $g=(g_1,g_2)^\top, h=(h_1,h_2)^\top \in H^{-1/2}(\Gamma)^2$ and

$$d=\lambda\sin\phi\sqrt{\mu/\epsilon},\ c=\sin\phi\sqrt{\epsilon/\mu},\ a=\omega\mu\cos^2\phi,\ b=\lambda\omega\epsilon\cos^2\phi.$$

Note that under the Assumption (A) the single-layer operator S is invertible from $H^{-1/2}(\Gamma)\to H^{1/2}(\Gamma)$.

THEOREM 5.1. *Suppose that Assumption (A) holds. Then the operator $A+B$ defined in (5.3) is a Fredholm operator with an index zero. Moreover, the system (5.3) admits a unique solution if $k^2\neq\gamma^2$ and $\lambda<0$.*

Proof. It suffices to prove the Fredholm property of $A+B$, since the second assertion of Theorem 5.1 follows directly from the Fredholm alternative combined with Theorem 3.1. To do this, we introduce the adjoint operator H of H', given by

$$(Hg)(x):=2\int_\Gamma \frac{\partial G(x,y)}{\partial\tau(y)}g(y)ds(y),\quad x\in\Gamma.$$

It is known that the adjoint operator of K is just K'. Since the operator S is compact from $H^{-1/2}(\Gamma)\to H^{-1/2}(\Gamma)$, we only need to justify the Fredholm property of the adjoint operator A^* of A, given by

$$A^*=\begin{pmatrix}\lambda(I+K) & -cH\\ dH & I+K\end{pmatrix}:\quad H^{1/2}(\Gamma)^2\to H^{1/2}(\Gamma)^2.$$

It is easy to see that the operator $H_1=H+j$ with the rank 1 operator

$$ju=(u,e)_{L^2(\Gamma)}e,\quad e=1\in\mathbb{C},$$

is invertible in $H^{1/2}(\Gamma)$. We will show that the operator

$$A_1:=\begin{pmatrix}\lambda(I+K) & -cH_1\\ dH_1 & I+K\end{pmatrix}:\quad H^{1/2}(\Gamma)^2\to H^{1/2}(\Gamma)^2$$

is a Fredholm operator with the index zero. Simple calculations show that the operator

$$B_1:=\begin{pmatrix}-(dH_1)^{-1}(I+K) & I\\ I & 0\end{pmatrix}=\begin{pmatrix}0 & I\\ I & (dH_1)^{-1}(I+K)\end{pmatrix}^{-1},$$

is invertible, and that

$$A_1B_1=\begin{pmatrix}-\lambda(I+K)(dH_1)^{-1}(I+K)-cH_1 & \lambda(I+K)\\ 0 & dH_1\end{pmatrix}. \tag{5.4}$$

Using the relations $HK=-KH$ and $(I+K)e=0$ (see [25]), we get

$$(I+K)H_1=H_1(I-K)-j(I-K),$$

and thus

$$(dH_1)^{-1}(I+K)=d^{-1}(I-K)H_1^{-1}-(dH_1)^{-1}[j(I-K)]H_1^{-1}. \tag{5.5}$$

Inserting (5.5) into (5.4) gives

$$A_1B_1=\begin{pmatrix}-d^{-1}[\lambda(I-K^2)+cdH_1^2]H_1^{-1}+j_1 & \lambda(I+K)\\ 0 & dH_1\end{pmatrix},$$

with $j_1 := \lambda(I+K)(dH_1)^{-1}[j(I-K)]H_1^{-1}$ being a rank one operator. Hence, A_1 is Fredholm with index zero if this is true for the operator $\lambda(I-K^2)+cdH_1^2$. Making use of $K^2-H^2=I$ and the definitions of c and d, we find

$$\lambda(I-K^2)+cdH_1^2=(cd-\lambda)H_1^2+j_2=-\lambda\cos^2\phi\,H_1^2+j_2, \tag{5.6}$$

where j_2 is some operator with rank one. Since $|\phi|<\pi/2$, we finally conclude that the operator (5.6) is Fredholm with index zero. Theorem 5.1 is thus proven. ∎

Acknowledgments. The work of G. Hu is partially supported by the National Natural Science Foundation of China (No. 12425112) and the Fundamental Research Funds for Central Universities in China (No. 63233071). The integral equation method of this paper was based on helpful discussions with Dr. G. Schmidt which are greatly appreciated.

REFERENCES

[1] T. Abboud, *Formulation variationnelle des équations de Maxwell dans un réseau bipériodique de* $\mathbf{R}^3$, C.R. Acad. Sci. Paris Sér. I Math., 317:245–248, 1993. 1

[2] H. Ammari and G. Bao, *Maxwell's equations in a perturbed periodic structure*, Adv. Comput. Math., 16:99–112, 2002. 1

[3] T. Arens, *Scattering by biperiodic layered media: The integral equation approach*, Habilitationsschrift at Karlsruhe Institute of Technology, 2010. 1, 5

[4] T. Arens, S.N. Chandler-Wilde, and J.A. DeSanto, *On integral equation and least squares methods for scattering by diffraction gratings*, Commun. Comput. Phys., 1:1010–1042, 2006. 1, 5

[5] R. Aylwin, C. Jerez-Hanckes, and J. Pinto, *On the properties of quasi-periodic boundary integral operators for the Helmholtz equation*, Integr. Equ. Oper. Theory, 92:1–41, 2020. 5

[6] G. Bao, *Finite element approximation of time harmonic waves in periodic structures*, SIAM J. Numer. Anal., 32:1155–1169, 1995. 1, 4

[7] G. Bao, *Variational approximation of Maxwell's equations in biperiodic structures*, SIAM J. Appl. Math., 57:364–381, 1997. 1

[8] G. Bao and D.C. Dobson, *On the scattering by a biperiodic structure*, Proc. Amer. Math. Soc., 128:2715–2723, 2000. 1

[9] G. Bao and P. Li, *Maxwell's Equations in Periodic Structures*, Springer/Science Press Beijing, Beijing, 208, 2022. 1, 4, 4

[10] G. Bao, P. Li, and H. Wu, *An adaptive edge element method with perfectly matched absorbing layers for wave scattering by biperiodic structures*, Math. Comput., 79:1–34, 2010. 1

[11] G. Bao and H. Wu, *Convergence analysis of the perfectly matched layer problems for time-harmonic Maxwell's equations*, SIAM J. Numer. Anal., 43:2121–2143, 2005. 1

[12] G. Bao and L. Zhang, *Analysis and computation for the scattering problem of electromagnetic waves in chiral media*, Commun. Math. Sci., 22:721–746, 2024. 1

[13] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics in Applied Mathematics, SIAM, Philadelphia, PA, 40, 2002. 4

[14] D.C. Dobson and A. Friedman, *The time-harmonic Maxwell equations in a doubly periodic structure*, J. Math. Anal. Appl., 166:507–528, 1992. 1, 5

[15] J. Elschner, R. Hinder, G. Schmidt, and F. Penzel, *Existence, uniqueness and regularity for solutions of the conical diffraction problem*, Math. Models Meth. Appl. Sci., 10:317–341, 2000. 1, 2, 2, 3, 3.2, 3

[16] J. Elschner and G. Hu, *Variational approach to scattering of plane elastic waves by diffraction gratings*, Math. Meth. Appl. Sci., 33:1924–1941, 2010. 1

[17] J. Elschner and G. Hu, *Scattering of plane elastic waves by three-dimensional diffraction gratings*, Math. Models Meth. Appl. Sci., 22(4):1150019, 2012. 1

[18] J. Elschner and G. Schmidt, *Diffraction in periodic structures and optimal design of binary gratings. I. Direct problems and gradient formulas*, Math. Meth. Appl. Sci., 21:1297–1342, 1998. 1

[19] J. Elschner and G. Schmidt, *Conical diffraction by periodic structures: variation of interfaces and gradient formulas*, Math. Nachr., 252:24–42, 2003. 1

[20] L. Feng, H. Wang, and L. Zhang, *The forward and inverse problems for the scattering of obliquely incident electromagnetic waves in a chiral medium*, J. Diff. Eqs., 284:102–125, 2021. 1

[21] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, Second Edition, 2013. 3

[22] G. Hu and A. Rathsfeld, *Convergence analysis of the FEM coupled with Fourier-mode expansion for the electromagnetic scattering by biperiodic structures*, Electron. Trans. Numer. Anal., 41:350–375, 2014. 1

[23] G. Hu and A. Rathsfeld, *Scattering of time-harmonic electromagnetic plane waves by perfectly conducting diffraction gratings*, IMA J. Appl. Math., 80:508–532, 2015. 1

[24] A. Kirsch, *Diffraction by periodic structures*, in L. Päivärinta and E. Somersalo (eds.), Inverse Problems in Mathematical Physics, Lecture Notes in Phys., Springer, Berlin, 422:87–102, 1993. 1, 5

[25] R. Kress, *Linear Integral Equations*, Springer, New York, Third Edition, 82, 2014. 5, 5

[26] G. Nakamura and H. Wang, *The direct electromagnetic scattering problem from an imperfectly conducting cylinder at oblique incidence*, J. Math. Anal. Appl., 397:142–155, 2013. 1, 5

[27] J.-C. Nédélec and F. Starling, *Integral equation methods in a quasi-periodic diffraction problem for the time-harmonic Maxwell's equations*, SIAM J. Math. Anal., 22(6):1679–1701, 1991. 1, 5

[28] R. Petit, editor, *Electromagnetic Theory of Gratings*, Springer-Verlag, Berlin-New York, 22, 1980. 1

[29] G. Schmidt, *Integral equations for conical diffraction by coated grating*, J. Integral Eqs. Appl., 23:71–112, 2011. 1, 2, 5, 5

[30] H. Wang and G. Nakamura, *The integral equation method for electromagnetic scattering problem at oblique incidence*, Appl. Numer. Math., 62:860–873, 2012. 1, 5