COMMUNICATIONS IN COMPUTATIONAL PHYSICS
Vol. **6**, No. 2, pp. 290-318

Commun. Comput. Phys.
August 2009

# An Adaptive Finite Element Method with a Modified Perfectly Matched Layer Formulation for Diffraction Gratings

Jie Chen[1], Desheng Wang[1,*] and Haijun Wu[2]

[1] *Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371.*
[2] *Department of Mathematics, Nanjing University, Jiangsu 210093, China.*

Received 26 April 2008; Accepted (in revised version) 29 August 2008

Communicated by Gang Bao

Available online 15 December 2008

**Abstract.** For numerical simulation of one-dimensional diffraction gratings both in TE and TM polarization, an enhanced adaptive finite element method is proposed in this paper. A modified perfectly matched layer (PML) formulation is proposed for the truncation of the unbounded domain, which results in a homogeneous Dirichlet boundary condition and the corresponding error estimate is greatly simplified. The a posteriori error estimates for the adaptive finite element method are provided. Moreover, a lower bound is obtained to demonstrate that the error estimates obtained are sharp.

**AMS subject classifications**: 65N30, 78A45, 35Q60

**Key words**: Diffraction grating, adaptive finite element method, PML, a posteriori error estimates.

## 1 Introduction

Due to its wide applications in micro-optics, diffraction gratings have recently received considerable attentions in both engineering and computational sciences [1, 2, 14]. There are various methods for the numerical simulation of diffraction gratings; among which the finite element method is one of the most popular approaches due to its capability in handling complicated geometries and boundary conditions. There are two challenges in applying the finite element method to diffraction grating simulation. One is to truncate the unbounded domain into a bounded one with some adequate approximation accuracy,

*Corresponding author. *Email addresses:* chen0437@ntu.edu.sg (J. Chen), desheng@ntu.edu.sg (D. S. Wang), hjw@nju.edu.cn (H. J. Wu)

 

and the other is to resolve the solution singularity caused by the discontinuity of the dielectric coefficient. To address these two issues, the perfectly matched layer (PML) [7] technique combined with a posterior error estimate based adaptive finite element method have been applied [4, 16].

Since the pioneering work of Babuška and Rheinboldt [6], the adaptive finite element methods based on a posteriori error estimates have become a central theme in scientific and engineering computations. For appropriately designed adaptive finite element procedures, the meshes and the associated numerical complexity are quasi-optimal, see, e.g., [4,9,11–13,15,16]. This makes the adaptive finite element method attractive for grating problems, which is often combined with the PML technique. In [16], Chen and Wu introduced an adaptive linear finite element algorithm with PML for domain truncation. A posteriori error estimate is derived to determine the PML thickness parameters automatically. Moreover, an exponential decay factor is introduced so that the a posteriori error estimate decays exponentially with respect to the distance to the computational domain, which makes the computational cost insensitive to the thickness of the absorbing layer. Later in [4], a second-order adaptive finite element method with error control was developed by Bao et al. for one-dimensional grating problems. The method has been applied to solve problems such as the 2D acoustic problem [17] and the 3D electromagnetic scattering problem [8].

Based on the work of Chen and Wu [16], several important improvements on the PML-based adaptive finite element method will be made in this paper:

(a) The PML formulation in [16] is modified by subtracting an auxiliary function from the electric field variable which satisfies the Helmholtz equation. The modification results in a homogeneous Dirichlet condition for all the boundaries, while in [16], the boundary condition on the upper boundary is not homogeneous. As a consequence, the exponential decay factor used for the error estimation in [16] is deleted here and accordingly the error analysis and practical implementation of the PML algorithm are greatly simplified.

(b) Furthermore, the error estimate for the PML is improved on the situation where the imaginary part of dilectric coefficient is small and positive, and the error bound is much better than that in [16]. The derived error estimate also implies that the solution of the PML problem converges exponentially to the solution of the grating problem when either the thickness of the PML layers or the PML medium parameters approaches infinity.

(c) A posteriori error estimates between the solution of the grating problem and the finite element approximation of the PML problem are derived. Since the modification of PML formulation results in a homogeneous Dirichlet boundary condition, both the estimations and proving are much simpler than that in [16]. And a lower bound of the a posterior error estimates is obtained, which is missed in [16]. The lower bound is not used in the practical adaptive finite element procedure, however it illustrates that the derived error estimates are sharp.

The remainder of this paper is organized as follows. In Section 2 the 1D diffraction gratings model is presented. In Section 3 the modified PML formulation is introduced

first, and then the finite element discretization with the error analysis are given. In Section 4 a sharp a posteriori error estimate which lays down the basis for the adaptive method is provided. A lower bound of the error between the PML solution and its finite element approximation is also derived. In Section 5 the results on the TM polarization are presented. The adaptive algorithm is outlined in Section 6 and a numerical example is included as well. Finally the conclusion and future work are contained in Section 7.

## 2 The 1D diffraction gratings model

The diffraction grating problem that arises when an electromagnetic wave is incident on a periodic structure is considered. The time harmonic Maxwell equations can be written as

$$\nabla\times\mathbf{E}-\mathbf{i}\omega\mu\mathbf{H}=0, \tag{2.1}$$

$$\nabla\times\mathbf{H}+\mathbf{i}\omega\varepsilon\mathbf{E}=0. \tag{2.2}$$

Here $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic field vectors, and $\varepsilon(x)$ represents the dielectric coefficient, where $x=(x_1,x_2,x_3)$. We will consider only the one-dimensional (1D) grating problem in which the medium and the grating structure are assumed to be constant in the $x_2$ direction. The dielectric coefficient $\varepsilon(x)=\varepsilon(x_1,x_3)$ is assumed to be periodic in the $x_1$ direction with the period $L>0$:

$$\varepsilon(x_1+nL,x_3)=\varepsilon(x_1,x_3)\qquad \forall\, x_1,x_3\in\mathbb{R},\ n\,\text{integer}.$$

The dielectric coefficient $\varepsilon(x)$ can be complex. And it is assumed that $\mathrm{Im}\varepsilon(x)\geq 0$ and $\mathrm{Re}\varepsilon(x)>0$ when $\mathrm{Im}\varepsilon(x)=0$. Also $\varepsilon$ is supposed to be constant away from the region $\{(x_1,x_3):b_2<x_3<b_1\}$ (see Fig. 1) in the sense that there exist constants $\varepsilon_1$ and $\varepsilon_2$ which satisfies

$$\varepsilon(x_1,x_3)=\varepsilon_1\qquad \text{in}\ \Omega_1=\{(x_1,x_3):x_3\geq b_1\},$$
$$\varepsilon(x_1,x_3)=\varepsilon_2\qquad \text{in}\ \Omega_2=\{(x_1,x_3):x_3\leq b_2\}.$$

In a practical application, we have $\varepsilon_1>0$, but $\varepsilon_2$ may be complex depending on property of the substrate material used in $\Omega_2$. Based on the direction and polarization of the incident plane wave, the Maxwell equations can be simplified by considering two fundamental polarizations: the transverse electric (TE) polarization and the transverse magnetic (TM) polarization. In the TE case, the electric field $\mathbf{E}$ is parallel to the $x_2$ axis: $\mathbf{E}=(0,u,0)^T\in\mathbb{R}^3$, where $u=u(x_1,x_3)$ satisfies the Helmholtz equation

$$\Delta u+k^2(x)u=0\quad\text{in}\quad\mathbb{R}^2. \tag{2.3}$$

Here $k^2(x)=\omega^2\varepsilon(x)\mu$ is the magnitude of the wave vector. Similarly, in the TM case, the magnetic field $\mathbf{H}$ is parallel to the $x_2$ axis: $\mathbf{H}=(0,u,0)^T\in\mathbb{R}^3$, where $u=u(x_1,x_3)$ satisfies the equation

$$\mathrm{div}\Big(\frac{1}{k^2(x)}\nabla u\Big)+u=0\quad\text{in}\quad\mathbb{R}^2. \tag{2.4}$$

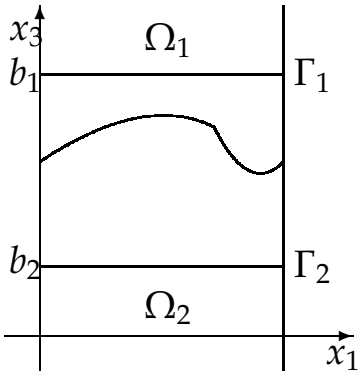


Figure 1: Geometry of the grating problem.

# 3 The modified PML formulation and the discrete problem

## 3.1 The modified PML formulation

Modified variational formulations for the grating problem (2.3) and (2.4) using PML techniques are proposed in this section. As the techniques for both the TE and TM polarization are similar, we shall concentrate on the TE polarization in this section, and then state the main results for the TM polarization in section 5.

Before introducing the absorbing PML layers, some definitions and results about the variational formulation with transparent boundary conditions are presented. For a more general discussion, please refer to [4, 16].

Let $u_{\rm I}=e^{{\rm i}\alpha x_1-{\rm i}\beta x_3}$ be the incoming plane wave which is incident upon the grating surface from the top, where $\alpha=k_1\sin\theta$, $\beta=k_1\cos\theta$, and $-\pi/2<\theta<\pi/2$. Here $\theta$ is the incident angle. We are concerned about a quasi-periodic solution $u$ of (2.3) in the sense that $u_\alpha=ue^{-{\rm i}\alpha x_1}$ and it is periodic in $x_1$ with a period $L>0$.

Let $\Gamma_j=\{(x_1,x_3):0<x_1<L,x_3=b_j\}$, $j=1,2$. Then the domain of the problem is reduced to

$$\Omega=\{(x_1,x_3):0<x_1<L,b_2<x_3<b_1\}.$$

For each integer $n$, let $\alpha_n=2\pi n/L$. As $u_\alpha$ is periodic in the $x_1$ direction, it has a Fourier series expansion

$$u_\alpha(x_1,x_3)=\sum_{n\in Z}u_\alpha^{(n)}(x_3)e^{{\rm i}\alpha_n x_1},\qquad u_\alpha^{(n)}=\frac{1}{L}\int_0^L u_\alpha e^{-{\rm i}\alpha_n x_1}{\rm d}x_1.$$

Thus we have the expansion

$$u(x_1,x_3)=u_\alpha e^{{\rm i}\alpha x_1}=\sum_{n\in Z}u_\alpha^{(n)}(x_3)e^{{\rm i}(\alpha_n+\alpha)x_1}.$$

For any integer $n\in\mathbb{Z}$ and $j=1,2$, $\beta_j^n$ is defined as

$$(\beta_j^n)^2=k_j^2-(\alpha_n+\alpha)^2,\qquad \mathrm{Im}\beta_j^n\geq 0,$$

where $k_j^2=\omega^2\varepsilon_j\mu$, and $k_j^2\neq(\alpha_n+\alpha)^2,\quad\forall n\in\mathbb{Z}$.

Substituting the above expansion into the Helmholtz equation (2.3) and noticing that the radiation condition for the diffraction problem implies that $u$ is composed of bounded outgoing plane waves in $\Omega_1$ (and $\Omega_2$), the following Rayleigh expansion can be obtained in $\Omega_1$:

$$u=u_{\mathrm{I}}+\sum_{n\in Z}A_1^n e^{\mathrm{i}(\alpha_n+\alpha)x_1+\mathrm{i}\beta_1^n x_3},\qquad x\in\Omega_1.$$

Similarly, the following Rayleigh expansion can be obtained in $\Omega_2$:

$$u=\sum_{n\in Z}A_2^n e^{\mathrm{i}(\alpha_n+\alpha)x_1-\mathrm{i}\beta_2^n x_3},\qquad x\in\Omega_2.$$

Usually, the grating structure is not very close to the upper boundary $\Gamma_1$ of the domain $\Omega$. Thus it follows that there exists a constant $\delta>0$ such that the grating structure is located in $\{(x_1,x_3):0<x_1<L,\,b_2<x_3<b_1-\delta\}$. Furthermore, a new variable is introduced by assuming

$$v(x_1,x_3)=u(x_1,x_3)-w(x_3)u_{\mathrm{I}}. \tag{3.1}$$

Here $w(x_3)\in C^2(\mathbb{R})$ is chosen as:

$$w(x_3)=\begin{cases}1 & \text{if}\quad b_1\leq x_3,\\ \rho(\dfrac{1}{\delta}(x_3-b_1+\delta)) & \text{if}\quad b_1-\delta\leq x_3<b_1,\\ 0 & \text{if}\quad x_3<b_1-\delta,\end{cases}$$

where $\rho(\tau)=6\tau^5-15\tau^4+10\tau^3$, $v=u-u_{\mathrm{I}}$ in $\Omega_1$ and $v=u$ in $\Omega_2$. We require $\rho(\tau)$ satisfies $\rho(0)=0$, $\rho(1)=1$ and $\rho\in C^2[0,1]$. For simpleness, we choose it as a polynomial.

For any quasi-periodic function $f$ with an expansion

$$f=\sum_{n\in Z}f^{(n)}e^{\mathrm{i}(\alpha_n+\alpha)x_1},$$

the following Dirichlet to Neumann operator $T_j$ is introduced in [3]:

$$(T_jf)(x_1)=\sum_{n\in Z}\mathbf{i}\beta_j^n f^{(n)}e^{\mathrm{i}(\alpha_n+\alpha)x_1},\quad 0<x_1<L,\ j=1,2.$$

With the above notations, the Rayleigh expansion of $u$ shows that $v$ satisfies

$$\frac{\partial v}{\partial\nu}-T_1v=0\quad\text{on}\quad\Gamma_1,\qquad\frac{\partial v}{\partial\nu}-T_2v=0\quad\text{on}\quad\Gamma_2, \tag{3.2}$$

where $\nu$ stands for the unit outer normal to $\partial\Omega$. And the above two equations are similar to transparent boundary conditions used in [4]. To define a variational formulation for the 1D grating problem (2.3) with the boundary conditions (3.2), the following subspace of $H^1(\Omega)$ is introduced, which includes all the quasi-periodic functions:

$$X(\Omega)=\{u\in H^1(\Omega):u(0,x_3)=e^{-\mathrm{i}\alpha L}u(L,x_3)\quad \text{for}\, b_2<x_3<b_1\}.$$

Define the sesquilinear form $b:X(\Omega)\times X(\Omega)\longrightarrow\mathbb{C}$ as follows:

$$b(\varphi,\psi)=\int_\Omega(\nabla\varphi\nabla\overline{\psi}-k^2(x)\varphi\overline{\psi})\,\mathrm{d}x-\sum_{j=1}^{2}\int_{\Gamma_j}(T_j\varphi)\overline{\psi}\,\mathrm{d}x_1. \tag{3.3}$$

Since the equation of $u$ in the domain $\Omega$ is the original Helmholtz equation $\Delta u+k^2(x)u=0$, it follows that

$$\Delta v+k^2(x)v=-g,\qquad \text{in}\quad \Omega,$$

where

$$g=\begin{cases}\Delta(w(x_3)u_{\mathrm{I}})+k^2(x)w(x_3)u_{\mathrm{I}} & \text{if}\quad b_1-\delta\le x_3\le b_1,\\ 0 & \text{otherwise.}\end{cases} \tag{3.4}$$

Thus the weak formulation of the 1D grating problem in the TE polarization reads as follows: Given an incoming plane wave $u_{\mathrm{I}}=e^{\mathrm{i}\alpha x_1-\mathrm{i}\beta x_3}$, seek $v\in X(\Omega)$ such that

$$b(v,\psi)=\int_\Omega f\overline{\psi}\,\mathrm{d}x\qquad\forall\,\psi\in X(\Omega). \tag{3.5}$$

The existence of a unique solution $v$ to (3.5) is proved for all but a sequence of countable frequencies $\omega_j$ with $|\omega_j|\to+\infty$. The existence issue is not going to be elaborated here and we just assume that (3.5) has a unique solution. And the general theory in Babuška and Aziz [5, Chapter 5] implies that there exists a constant $\gamma>0$ such that the following inf-sup condition holds:

$$\sup_{0\neq\psi\in H^1(\Omega)}\frac{|b(\varphi,\psi)|}{\|\psi\|_{H^1(\Omega)}}\ge\gamma\|\varphi\|_{H^1(\Omega)}\qquad\forall\,\varphi\in X(\Omega). \tag{3.6}$$

Now we turn to the introduction of absorbing PML layers. The computational domain $\Omega$ is surrounded by two PML layers with thickness $\delta_1$ and $\delta_2$ in $\Omega_1$ and $\Omega_2$ respectively. Let $s(x_3)=s_1(x_3)+\mathbf{i}s_2(x_3)$ be the model medium property and it satisfies

$$s_1,s_2\in C(\mathbb{R}),\quad s_1\ge1,s_2\ge0,\quad\text{and}\quad s(x_3)=1\ \text{for}\ b_2\le x_3\le b_1. \tag{3.7}$$

Here we remark that, in contrast to the original PML condition which takes $s_1\equiv1$ in the PML region, a variable $s_1$ is used here to attenuate both the outgoing and evanescent waves. The advantage of this extension makes our method insensitive to the distance

of the PML region from the structure. Following the general idea in designing PML absorbing layers, we introduce the PML regions

$$\Omega_1^{\mathrm{PML}}=\{(x_1,x_3):0<x_1<L \text{ and } b_1<x_3<b_1+\delta_1\},$$
$$\Omega_2^{\mathrm{PML}}=\{(x_1,x_3):0<x_1<L \text{ and } b_2-\delta_2<x_3<b_2\},$$

and the PML differential operator

$$\mathcal{L}:=\frac{\partial}{\partial x_1}\Big(s(x_3)\frac{\partial}{\partial x_1}\Big)+\frac{\partial}{\partial x_3}\Big(\frac{1}{s(x_3)}\frac{\partial}{\partial x_3}\Big)+k^2(x)s(x_3).$$

Then the PML equation assumes the form (see [4,16])

$$\mathcal{L}(\hat{u}-u_{\mathrm{I}})=0 \qquad \text{in } \Omega_1^{\mathrm{PML}},$$
$$\mathcal{L}\hat{u}=0 \qquad \text{in } \Omega_2^{\mathrm{PML}},$$

where $\hat{u}$ satisfies the Helmholtz equation $\Delta\hat{u}+k^2\hat{u}=0$ in $\Omega$. Assume

$$\hat{v}(x_1,x_3)=\hat{u}(x_1,x_3)-w(x_3)u_{\mathrm{I}}, \tag{3.8}$$

and let $D=\{(x_1,x_3):0<x_1<L,b_2-\delta_2<x_3<b_1+\delta_1\}$. Applying (3.7), we can formulate a modified PML model:

$$\mathcal{L}\hat{v}=-g \quad \text{in} \quad D, \tag{3.9}$$

with a quasi-periodic boundary condition in $x_1$ direction:

$$\hat{v}(0,x_3)=e^{-\mathrm{i}\alpha L}\hat{v}(L,x_3) \quad \text{for} \quad b_2-\delta_2<x_3<b_1+\delta_1,$$

and a Dirichlet boundary condition:

$$\hat{v}=0 \quad \text{on} \quad \Gamma_1^{\mathrm{PML}}=\{(x_1,x_3):0<x_1<L, x_3=b_1+\delta_1\},$$
$$\hat{v}=0 \quad \text{on} \quad \Gamma_2^{\mathrm{PML}}=\{(x_1,x_3):0<x_1<L, x_3=b_2-\delta_2\},$$

where the source function $g$ is defined in (3.4). For any $G\subset D$, define

$$X(G)=\{u\in H^1(G):u_\alpha=ue^{-\mathrm{i}\alpha x_1} \text{ is periodic in } x_1 \text{ with period } L\},$$

and the sesquilinear form $a_G:X(G)\times X(G)\longrightarrow\mathbb{C}$ can be introduced as

$$a_G(\varphi,\psi)=\int_G(s(x_3)\frac{\partial\varphi}{\partial x_1}\frac{\partial\overline{\psi}}{\partial x_1}+\frac{1}{s(x_3)}\frac{\partial\varphi}{\partial x_3}\frac{\partial\overline{\psi}}{\partial x_3}-k^2s(x_3)\varphi\overline{\psi})\mathrm{d}x.$$

Define $\overset{\circ}{X}(D)=\{u\in X(D),u=0\,\text{on}\,\Gamma_1^{\mathrm{PML}}\cup\Gamma_2^{\mathrm{PML}}\}$. Then the weak formulation of the PML model reads as follows: Find $\hat{v}\in\overset{\circ}{X}(D)$ such that

$$a_D(\hat{v},\psi)=\int_D g\overline{\psi}\mathrm{d}x \quad \forall\,\psi\in\overset{\circ}{X}(D). \tag{3.10}$$

Our next objective is to prove the existence and uniqueness of the above problem and derive the error estimate between $\hat{v}$ and $v=u-w(x_3)u_I$. Here we remind that $v-\hat{v}=u-\hat{u}$ is the error between the solutions of the original PML problem and the 1D grating problem. To achieve the goal, we first find an equivalent formulation of (3.10) in the domain $\Omega$. Similar to the derivation of the variational formulation with transparent boundary condition, we get the following:

$$\hat{v}=\sum_{n\in Z}\frac{\zeta_1^n(x_3)}{\zeta_1^n(b_1)}\hat{u}_\alpha^{(n)}(b_1)e^{\mathrm{i}(\alpha_n+\alpha)x_1} \quad \text{in} \quad \Omega_1^{\mathrm{PML}},$$
$$\hat{v}=\sum_{n\in Z}\frac{\zeta_2^n(x_3)}{\zeta_2^n(b_2)}\hat{u}_\alpha^{(n)}(b_2)e^{\mathrm{i}(\alpha_n+\alpha)x_1} \quad \text{in} \quad \Omega_2^{\mathrm{PML}},$$

where

$$\zeta_1^n(x_3)=\exp\Big(-\mathrm{i}\beta_1^n\int_{x_3}^{b_1+\delta_1}s(\tau)d\tau\Big)-\exp\Big(\mathrm{i}\beta_1^n\int_{x_3}^{b_1+\delta_1}s(\tau)d\tau\Big),$$
$$\zeta_2^n(x_3)=\exp\Big(-\mathrm{i}\beta_2^n\int_{b_2-\delta_2}^{x_3}s(\tau)d\tau\Big)-\exp\Big(\mathrm{i}\beta_2^n\int_{b_2-\delta_2}^{x_3}s(\tau)d\tau\Big).$$

For any quasi-periodic function $f$ having the expansion

$$f=\sum_{n\in Z}f^{(n)}e^{\mathrm{i}(\alpha_n+\alpha)x_1},$$

the Dirichlet to Neumann operator $T_j^{\mathrm{PML}}$ is defined as:

$$(T_j^{\mathrm{PML}}f)(x_1)=\sum_{n\in Z}\mathbf{i}\beta_j^n\coth(-\mathbf{i}\beta_j^n\sigma_j)f^{(n)}e^{\mathrm{i}(\alpha_n+\alpha)x_1},\quad 0<x_1<L,\ j=1,2,$$

where $\coth(\tau)=(e^\tau+e^{-\tau})/(e^\tau-e^{-\tau})$ and

$$\sigma_1=\int_{b_1}^{b_1+\delta_1}s(\tau)d\tau,\qquad \sigma_2=\int_{b_2-\delta_2}^{b_2}s(\tau)d\tau. \tag{3.11}$$

Then it follows that

$$\frac{\partial\hat{v}}{\partial\nu}-T_1^{\mathrm{PML}}\hat{v}=0\quad\text{on}\quad\Gamma_1,\qquad \frac{\partial\hat{v}}{\partial\nu}-T_2^{\mathrm{PML}}\hat{v}=0\quad\text{on}\quad\Gamma_2.$$

This motivates us to introduce the sesquilinear form $b^{\mathrm{PML}}:X(\Omega)\times X(\Omega)\longrightarrow\mathrm{C}$,

$$b^{\mathrm{PML}}(\varphi,\psi)=\int_\Omega(\nabla\varphi\nabla\overline{\psi}-k^2(x)\varphi\overline{\psi})\mathrm{d}x-\sum_{j=1}^{2}\int_{\Gamma_j}(T_j^{\mathrm{PML}}\varphi)\overline{\psi}\mathrm{d}x_1, \tag{3.12}$$

and introduce the following variational problem: Find $\vartheta\in X(\Omega)$ such that

$$b^{\mathrm{PML}}(\vartheta,\psi)=\int_\Omega g\overline{\psi}\mathrm{d}x\qquad\forall\,\psi\in X(\Omega), \tag{3.13}$$

where $g$ is defined in (3.4).

The following lemma [16] establishes the relation between the variational problem (3.4) and the modified PML model problem (3.10).

**Lemma 3.1.** *Any solution $\hat{v}$ of the problem* (3.10) *restricted to $\Omega$ is a solution of* (3.13). *Conversely, any solution $\vartheta$ of the problem* (3.13) *can be uniquely extended to the whole domain $D$ to be a solution of* (3.10).

Let

$$\begin{aligned}&\Delta_j^n=\left|\mathrm{Re}(k_j^2)-(\alpha_n+\alpha)^2\right|^{1/2},\quad U_j=\{n:\mathrm{Re}(k_j^2)\geq(\alpha_n+\alpha)^2\},\ j=1,2,\\&\Delta_j^-=\min\{\Delta_j^n:n\in U_j\},\qquad \Delta_j^+=\min\{\Delta_j^n:n\notin U_j\}.\end{aligned}\tag{3.14}$$

The following lemma plays a key role in the subsequent analysis.

**Lemma 3.2.** *For any $\varphi,\psi\in X(\Omega)$, we have*

$$\left|\int_{\Gamma_j}(T_j\varphi-T_j^{\mathrm{PML}}\varphi)\overline{\psi}\,\mathrm{d}x_1\right|\leq M_j\|\varphi\|_{L^2(\Gamma_j)}\|\psi\|_{L^2(\Gamma_j)},$$

*and $M_j=\min(M_j^a,M_j^b)$, where*

$$M_j^a=\begin{cases}\max\left(\frac{2\Delta_j^-}{\exp(2\sigma_j^I\Delta_j^-)-1},\frac{2\Delta_j^+}{\exp(2\sigma_j^R\Delta_j^+)-1}\right) & \text{if } \sigma_2^R\Delta_j^+\geq1 \text{ and } \sigma_2^I\Delta_j^-\geq1,\\ +\infty & \text{otherwise,}\end{cases}$$

$$M_j^b=\begin{cases}\frac{2|k_j|}{\exp(2\sigma_j^R\mathrm{Im}k_j)-1} & \text{if } \mathrm{Im}\varepsilon_j>0,\\ +\infty & \text{otherwise,}\end{cases}$$

*and $\sigma_j^R,\sigma_j^I$ are the real and imaginary parts of $\sigma_j$ defined in* (3.11), *that is, $\sigma_j=\sigma_j^R+\mathrm{i}\sigma_j^I$.*

*Proof.* For any $\varphi,\psi\in X(\Omega)$, their traces on $\Gamma_j$ have the following expansions:

$$\varphi(x_1,b_j)=\sum_{n\in\mathbb{Z}}\varphi_\alpha^{(n)}(b_j)e^{\mathrm{i}(\alpha_n+\alpha)x_1},\quad \psi(x_1,b_j)=\sum_{n\in\mathbb{Z}}\psi_\alpha^{(n)}(b_j)e^{\mathrm{i}(\alpha_n+\alpha)x_1},$$

where $\varphi_\alpha^{(n)}$ and $\psi_\alpha^{(n)}$ are the Fourier coefficients of the periodic functions

$$\varphi_\alpha(x_1,b_j)=\varphi(x_1,b_j)e^{-\mathrm{i}\alpha x_1},\quad \psi_\alpha(x_1,b_j)=\psi(x_1,b_j)e^{-\mathrm{i}\alpha x_1},$$

respectively. The orthogonality property of the Fourier series yields

$$\begin{aligned}\|\varphi\|_{L^2(\Gamma_j)}^2&=\|\varphi_\alpha\|_{L^2(\Gamma_j)}^2=L\sum_{n\in\mathbb{Z}}\left|\varphi_\alpha^{(n)}(b_j)\right|^2,\\ \|\psi\|_{L^2(\Gamma_j)}^2&=\|\psi_\alpha\|_{L^2(\Gamma_j)}^2=L\sum_{n\in\mathbb{Z}}\left|\psi_\alpha^{(n)}(b_j)\right|^2,\end{aligned}$$

and

$$\int_{\Gamma_j}(T_j\varphi-T_j^{\mathrm{PML}}\varphi)\overline{\psi}\mathrm{d}x=L\sum_{n\in\mathbb{Z}}\mathrm{i}\beta_j^n(1-\coth(-\mathrm{i}\beta_j^n\sigma_j))\varphi_\alpha^{(n)}(b_j)\overline{\psi}_\alpha^{(n)}(b_j). \tag{3.15}$$

Define

$$\xi_n=\frac{1}{2}(\mathrm{Re}(k_j^2)-(\alpha_n+\alpha)^2),\qquad \eta=\frac{1}{2}\mathrm{Im}(k_j^2), \tag{3.16}$$

and observe that $(\beta_j^n)^2=2(\xi_n+\mathrm{i}\eta)$. Since $\mathrm{Im}(k_j^2)=\omega^2\mathrm{Im}(\varepsilon_j)\mu\geq 0$, we obtain that

$$\mathrm{Re}\beta_j^n=(\sqrt{\xi_n^2+\eta^2}+\xi_n)^{1/2},\qquad \mathrm{Im}\beta_j^n=(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}. \tag{3.17}$$

Thus $\mathrm{Re}(-2\mathrm{i}\beta_j^n\sigma_j)=2(\sigma_j^R\mathrm{Im}\beta_j^n+\sigma_j^I\mathrm{Re}\beta_j^n)\geq 2\sigma_j^R\mathrm{Im}\beta_j^n$, which implies that

$$\begin{aligned}\left|\mathrm{i}\beta_j^n(1-\coth(-\mathrm{i}\beta_j^n\sigma_j))\right|&=\left|\frac{2\beta_j^n}{e^{-2\mathrm{i}\beta_j^n\sigma_j}-1}\right|\\ \leq\frac{2\left|\beta_j^n\right|}{e^{2\sigma_j^R\mathrm{Im}\beta_j^n}-1}&=\frac{2\sqrt{2}(\xi_n^2+\eta^2)^{1/4}}{e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}-1}.\end{aligned}$$

For the sake of convenience, we make the following notations:

$$\begin{aligned}g(\xi_n,\eta)&:=\frac{(\xi_n^2+\eta^2)^{1/4}}{e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}-1},\\ P(\xi_n,\eta)&:=\left|\mathrm{i}\beta_j^n(1-\coth(-\mathrm{i}\beta_j^n\sigma_j))\right|\leq 2\sqrt{2}g(\xi_n,\eta).\end{aligned}$$

Simple differential computation gives

$$\begin{aligned}\frac{\partial g(\xi_n,\eta)}{\partial\xi_n}=&\frac{(\xi_n^2+\eta^2)^{-3/4}e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}}{(e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}-1)^2}\Big[(1-e^{-2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}})\xi_n/2\\ &+\sqrt{\xi_n^2+\eta^2}\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}\Big]\geq 0,\end{aligned}$$

which shows that $g(\xi_n,\eta)$ increases with respect to $\xi_n$. As $\xi_n\leq\frac{1}{2}\mathrm{Re}k_j^2$, it can be verified that for $n\in\mathbb{Z}$

$$P(\xi_n,\eta)\leq 2\sqrt{2}g(\frac{1}{2}\mathrm{Re}k_j^2,\eta)=\frac{2\left|k_j\right|}{e^{2\sigma_j^R\mathrm{Im}k_j}-1},\qquad \text{if}\quad \mathrm{Im}\varepsilon_j>0. \tag{3.18}$$

For $n \notin U_j$, we have

$$\begin{aligned}\frac{\partial g(\xi_n,\eta)}{\partial \eta} =& \frac{\eta(\xi_n^2+\eta^2)^{-3/4}e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}}{(e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}-1)^2}\Big[(1-e^{-2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}})^{1/2}\\ &-\sqrt{\xi_n^2+\eta^2}\sigma_j^R/(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}\Big]\\ \le& \frac{\eta(\xi_n^2+\eta^2)^{-3/4}e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}}{(e^{2\sigma_j^R(\sqrt{\xi_n^2+\eta^2}-\xi_n)^{1/2}}-1)^2}\Big(\frac{1}{2}-\sigma_j^R\frac{|\xi_n|^{1/2}}{\sqrt{2}}\Big).\end{aligned}$$

It is easy to see that $|\xi_n|^{1/2} \ge \frac{\Delta_j^+}{\sqrt{2}}$, so $g(\xi_n,\eta)$ decreases with respect to $\eta>0$, provided that $\sigma^R\Delta_j^+ \ge 1$. Thus we get

$$P(\xi_n,\eta) \le \frac{2\sqrt{2}|\xi_n|^{1/2}}{e^{2\sqrt{2}\sigma_j^R|\xi_n|^{1/2}}-1} \le \frac{2\Delta_j^+}{e^{2\sigma_j^R\Delta_j^+}-1}, \qquad \text{if}\,\sigma_j^R\Delta_j^+ \ge 1. \tag{3.19}$$

In the above deduction, we have used the fact that when $x>0$ the function $x/(e^{ax}-1)$ decreases with respect to $x$, where $a>0$.

Similarly, for $n \in U_j$, it follows that

$$\mathrm{Re}(-2\mathrm{i}\beta_j^n\sigma_j) = 2(\sigma_j^R\mathrm{Im}\beta_j^n + \sigma_j^I\mathrm{Re}\beta_j^n) \ge 2\sigma_j^I\mathrm{Re}\beta_j^n.$$

Consequently,

$$\begin{aligned}P(\xi_n,\eta) \le& \frac{2\sqrt{2}(\xi_n^2+\eta^2)^{1/4}}{e^{2\sigma_j^I(\sqrt{\xi_n^2+\eta^2}+\xi_n)^{1/2}}-1} \le \frac{2\sqrt{2}\xi_n^{1/2}}{e^{2\sqrt{2}\sigma_j^I\xi_n^{1/2}}-1}\\ \le& \frac{2\Delta_j^-}{e^{2\sigma_j^I\Delta_j^-}-1}, \qquad \text{if } \sigma_j^I\Delta_j^- \ge 1.\end{aligned} \tag{3.20}$$

Combining (3.19) and (3.20) gives that, for $n \in \mathbb{Z}$,

$$P(\xi_n,\eta) \le \max\left(\frac{2\Delta_j^+}{e^{2\sigma_j^R\Delta_j^+}-1}, \frac{2\Delta_j^-}{e^{2\sigma_j^I\Delta_j^-}-1}\right) \qquad \text{if } \sigma_j^R\Delta_j^+ \ge 1,\ \sigma_j^I\Delta_j^- \ge 1. \tag{3.21}$$

Applying (3.18),(3.21) and the Cauchy inequality in (3.15) leads to the completion of the proof. □

**Remark 3.1.** The estimate between $T_j$ and $T_j^{\mathrm{PML}}$ is an improvement over the result in [16]. Our error bound is similar to that in [16] if $\sigma_j^R\Delta_j^+ \ge 1$, $\sigma_j^I\Delta_j^- \ge 1$ and $\mathrm{Im}\varepsilon_j = 0$, and is not larger than that in [16] if $\mathrm{Im}\varepsilon_j > 0$. However, when $\mathrm{Im}\varepsilon_j > 0$ and $\mathrm{Im}\varepsilon_j$ is small, our estimate is much better if $\Delta_j^+$ and $\Delta_j^-$ are not small.

The following trace inequality is from [16] and the proof is omitted.

**Lemma 3.3.** *For any $\psi \in X(\Omega)$, we have*

$$\|\psi\|_{L^2(\Gamma_j)} \le \|\psi\|_{H^{1/2}(\Gamma_j)} \le \hat{C}\|\psi\|_{H^1(\Omega)},$$

*with $\hat{C} = \sqrt{1+(b_2-b_1)^{-1}}$. Here if $\psi(x_1,b_j) = \sum_{n\in Z} \psi_\alpha^{(n)}(b_j)e^{i(\alpha_n+\alpha)x_1}$ on $\Gamma_j$,*

$$\|\psi\|_{H^{1/2}(\Gamma_j)} = \left( L\sum_{n\in Z}(1+|\alpha_n+\alpha|^2)^{1/2}\left|\psi_\alpha^{(n)}(b_j)\right|^2\right)^{1/2}.$$

**Theorem 3.1.** *Let $\gamma > 0$ be the constant in the inf-sup condition (3.6) and assume that $(M_1+M_2)\hat{C}^2 < \gamma$. Then the modified PML variational problem has a unique solution $\hat{v}$. Moreover, we have the following error estimate:*

$$|||v-\hat{v}|||_\Omega := \sup_{0\neq\psi\in H^1(\Omega)} \frac{|b(v-\hat{v},\psi)|}{\|\psi\|_{H^1(\Omega)}} \le \hat{C}\sum_{j=1}^{2} M_j\|\hat{v}\|_{L^2(\Gamma_j)}. \tag{3.22}$$

**Remark 3.2.** From the definition of $v$ and $\hat{v}$ we obtain the error estimate between the solution of the grating problem $u$ and the PML solution $\hat{u}$:

$$|||u-\hat{u}|||_\Omega = |||v+w(x_3)u_I-\hat{v}-w(x_3)u_I|||_\Omega = |||v-\hat{v}|||_\Omega.$$

We remark that the error estimate (3.22) is a posteriori in nature as it depends on the modified PML solution $\hat{v}$. This makes a posteriori error control possible (see Section 3 for details). The proof for the theorem is similar to that employed in [16] and here it is omitted.

## 3.2 The discrete problem

In this section the finite element method for the modified PML problem (3.10) is presented. Let $\mathcal{M}_h$ be a regular triangulation of the domain $D$ and remember that any element $T \in \mathcal{M}_h$ is considered as closed. Let $V_h(D) \subset X(D)$ be the $n$-th order Lagrange finite element space and $\overset{\circ}{V}_h(D) = V_h(D) \cap \overset{\circ}{X}(D)$, and denote the standard finite element interpolation operator by $I_h : C(\overline{D}) \longrightarrow V_h(D)$. Then the finite element approximation to the modified PML problem (3.10) is defined as: Find $\hat{v}_h \in \overset{\circ}{V}_h(D)$ such that

$$a_D(\hat{v}_h,\psi_h) = \int_D g\overline{\psi}_h \mathrm{d}x \quad \forall\, \psi_h \in \overset{\circ}{V}_h(D). \tag{3.23}$$

Assume that the discrete problem (3.23) has a unique solution $\hat{v}_h \in \overset{\circ}{V}_h(D)$, and let

$$A(x) = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} = \begin{pmatrix} s(x_3) & 0 \\ 0 & \dfrac{1}{s(x_3)} \end{pmatrix},$$

$$B(x) = k^2(x)s(x_3).$$

Then the definition of $\mathcal{L}$ and $a_D$ can be rewritten as

$$\mathcal{L}=\operatorname{div}(A(x)\nabla)+B(x),$$
$$a_D(\varphi,\psi)=\int_D (A(x)\nabla\varphi\nabla\overline{\psi}-B(x)\varphi\overline{\psi})\,\mathrm{d}x.$$

For any $T\in\mathcal{M}_h$, we denote its diameter by $h_T$.

Let $\mathcal{B}_h$ denote the set of all sides that do not lie on $\Gamma_j^{\mathrm{PML}}$, $j=1,2$, and $h_e$ stand for its length for any $e\in\mathcal{B}_h$ and for any $T\in\mathcal{M}_h$. We introduce the residual

$$R_T:=\mathcal{L}\hat{v}_h|_T+g|_T. \tag{3.24}$$

For any interior side $e\in\mathcal{B}_h$ which is the common side of $T_1$ and $T_2\in\mathcal{M}_h$, we define the jump residual across $e$ as

$$J_e=(A\nabla\hat{v}_h|_{T_1}-A\nabla\hat{v}_h|_{T_2})\cdot\nu_e, \tag{3.25}$$

where the unit normal vector $\nu_e$ of $e$ points from $T_2$ to $T_1$. Also we define $\Gamma_{\text{left}}=\{(x_1,x_3):x_1=0,b_2-\delta_2<x_3<b_1+\delta_1\}$ and $\Gamma_{\text{right}}=\{(x_1,x_3):x_1=L,b_2-\delta_2<x_3<b_1+\delta_1\}$. If $e=\Gamma_{\text{left}}\cap\partial T$ for some element $T\in\mathcal{M}_h$ and $e'$ is the corresponding side on $\Gamma_{\text{right}}$ which is also a side of some element $T'$, then the jump residual can be defined as

$$\begin{aligned}J_e&=A_{11}\left[\frac{\partial}{\partial x_1}(\hat{v}_h|_T)-e^{-\mathrm{i}\alpha L}\cdot\frac{\partial}{\partial x_1}(\hat{v}_h|_{T'})\right],\\ J_{e'}&=A_{11}\left[e^{\mathrm{i}\alpha L}\cdot\frac{\partial}{\partial x_1}(\hat{v}_h|_T)-\frac{\partial}{\partial x_1}(\hat{v}_h|_{T'})\right].\end{aligned} \tag{3.26}$$

For any $T\in\mathcal{M}_h$, we denote the local error estimator by $\eta_T$ of the following form:

$$\eta_T=\max_{x\in\tilde{T}}|s(x_3)|\left[h_T\|R_T\|_{L^2(T)}+(\frac{1}{2}\sum_{e\subset T}h_e\|J_e\|^2_{L^2(e)})^{1/2}\right], \tag{3.27}$$

where $\tilde{T}$ is the union of all elements in $\mathcal{M}_h$ that have nonempty intersection with $T\in\mathcal{M}_h$. With the above definitions and notations, we present the main result of this paper.

**Theorem 3.2.** *There exists a constant $C>0$, depending only on the minimum angle of the mesh $\mathcal{M}_h$, such that the following a posteriori error estimate holds:*

$$|||v-\hat{v}_h|||_\Omega\le\sum_{j=1}^{2}\hat{C}M_j\|\hat{v}_h\|_{L^2(\Gamma_j)}+C(1+C_1+C_2)\Big(\sum_{T\in\mathcal{M}_h}\eta_T^2\Big)^{1/2},$$

*where the constants $M_j$ $(j=1,2)$, $\hat{C},C_j$ are defined in Lemmas* 3.2, 3.3 *and* 4.3, *respectively.*

The proof of this theorem will be given in Section 4.

**Remark 3.3.** From the definition of $v$, it follows that

$$|||v-\hat{v}_h|||_\Omega=|||u-(\hat{v}_h+w(x_3)u_{\mathrm{I}})|||_\Omega,$$

which implies that $\hat{v}_h+w(x_3)u_{\mathrm{I}}$ can be used to approximate $u$.

# 4 Error analysis

In this section the a posteriori error estimates in Theorem 3.2 is proved and the lower bound of $\|\hat{v}-\hat{v}_h\|$ is obtained.

## 4.1 Error representation formula

For any $\psi\in X(\Omega)$, it can be extended to $X(D)$ as follows:

$$\tilde{\psi}(x_1,x_3)=\sum_{n\in Z}\frac{\overline{\zeta}_j^n(x_3)}{\overline{\zeta}_j^n(b_j)}\psi_\alpha^{(n)}(b_j)e^{\mathrm{i}(\alpha_n+\alpha)x_1}\quad\text{in}\quad\Omega_j^{\mathrm{PML}},\ j=1,2, \tag{4.1}$$

where $\overline{\zeta}_j^n(x_3)$ is the conjugation of $\zeta_j^n(x_3)$, and $\psi_\alpha^{(n)}(b_j)$ are the Fourier coefficients of the function $\psi_\alpha=\psi e^{-\mathrm{i}\alpha x_1}$ on $\Gamma_j$, and

$$\psi(x_1,b_j)=\sum_{n\in Z}\psi_\alpha^{(n)}(b_j)e^{\mathrm{i}(\alpha_n+\alpha)x_1}. \tag{4.2}$$

It is easy to see that $\tilde{\psi}=\psi$ on $\Gamma_j$ and $\mathcal{L}\overline{\tilde{\psi}}=0$ in $\Omega_j^{\mathrm{PML}}$.

**Lemma 4.1** ([16]). *Let $\nu_j$ be the unit outer normal to $\Omega_j^{\mathrm{PML}}$. Then for any $\varphi,\psi\in X(\Omega)$ we have*

$$\int_{\Gamma_j}T_j^{\mathrm{PML}}\varphi\overline{\psi}\,\mathrm{d}x_1=-\int_{\Gamma_j}\varphi\frac{\partial\overline{\tilde{\psi}}}{\partial\nu_j}\mathrm{d}x_1. \tag{4.3}$$

With no confusion of notations, we shall write $\tilde{\psi}$ as $\psi$ in $\Omega_j^{\mathrm{PML}}$ in what follows.

**Lemma 4.2** (Error representation formula). *For any $\psi\in X(\Omega)$, which is extended to $X(D)$ according to* (4.1)*, and $\psi_h\in\overset{\circ}{V}_h(D)$, we have*

$$b(v-\hat{v}_h,\psi)=\int_D g(\overline{\psi-\psi_h})\,\mathrm{d}x-a_{_D}(\hat{v}_h,\psi-\psi_h)+\sum_{j=1}^2\int_{\Gamma_j}(T_j-T_j^{\mathrm{PML}})\hat{v}_h\overline{\psi}\mathrm{d}x_1. \tag{4.4}$$

*Proof.* First by (3.5), Lemma 3.1, (3.3) and (3.12), we have

$$\begin{aligned}
&b(v-\hat{v}_h,\psi)=b(v-\hat{v},\psi)+b(\hat{v}-\hat{v}_h,\psi)\\
=&b^{\mathrm{PML}}(\hat{v},\psi)-b(\hat{v},\psi)+b^{\mathrm{PML}}(\hat{v}-\hat{v}_h,\psi)+b(\hat{v}-\hat{v}_h,\psi)-b^{\mathrm{PML}}(\hat{v}-\hat{v}_h,\psi)\\
=&\sum_{j=1}^2\int_{\Gamma_j}(T_j-T_j^{\mathrm{PML}})\hat{v}\overline{\psi}\,\mathrm{d}x_1+b^{\mathrm{PML}}(\hat{v}-\hat{v}_h,\psi)-\sum_{j=1}^2\int_{\Gamma_j}(T_j-T_j^{\mathrm{PML}})(\hat{v}-\hat{v}_h)\overline{\psi}\,\mathrm{d}x\\
=&\sum_{j=1}^2\int_{\Gamma_j}(T_j-T_j^{\mathrm{PML}})\hat{v}_h\overline{\psi}\,\mathrm{d}x_1+b^{\mathrm{PML}}(\hat{v}-\hat{v}_h,\psi).
\end{aligned}$$

Next, (3.12) and Lemma 4.1 together give

$$b^{\mathrm{PML}}(\hat{v}-\hat{v}_h,\psi)=a_\Omega(\hat{v}-\hat{v}_h,\psi)-\sum_{j=1}^{2}\int_{\Gamma_j}T_j^{\mathrm{PML}}(\hat{v}-\hat{v}_h)\overline{\psi}\,\mathrm{d}x$$
$$=a_\Omega(\hat{v}-\hat{v}_h,\psi)+\sum_{j=1}^{2}\int_{\Gamma_j}(\hat{v}-\hat{v}_h)\frac{\partial\overline{\psi}}{\partial\nu_j}\mathrm{d}x_1.$$

Since $\mathcal{L}\overline{\psi}=0$ in $\Omega_j^{\mathrm{PML}}$ according to (4.1) and (4.2), the Green formula shows that

$$\begin{aligned}a_{\Omega_j^{\mathrm{PML}}}(\hat{v}-\hat{v}_h,\psi)&=\int_{\Omega_j^{\mathrm{PML}}}\left[s(x_3)\frac{\partial(\hat{v}-\hat{v}_h)}{\partial x_1}\frac{\partial\overline{\psi}}{\partial x_1}+\frac{1}{s(x_3)}\frac{\partial(\hat{v}-\hat{v}_h)}{\partial x_3}\frac{\partial\overline{\psi}}{\partial x_3}\right.\\&\qquad\left.-k^2(x)s(x_3)(\hat{v}-\hat{v}_h)\overline{\psi}\right]\mathrm{d}x\\&=-\int_{\Omega_j^{\mathrm{PML}}}\left[(\hat{v}-\hat{v}_h)\frac{\partial}{\partial x_1}(s(x_3)\frac{\partial\overline{\psi}}{\partial x_1})+(\hat{v}-\hat{v}_h)\frac{\partial}{\partial x_3}(\frac{1}{s(x_3)}\frac{\partial\overline{\psi}}{\partial x_3})\right.\\&\qquad\left.+k^2(x)s(x_3)(\hat{v}-\hat{v}_h)\overline{\psi}\right]\mathrm{d}x+\int_{\partial\Omega_j^{\mathrm{PML}}}(\hat{v}-\hat{v}_h)A(x)\nabla\overline{\psi}\cdot\nu_j\mathrm{d}s\\&=\int_{\Gamma_j}(\hat{v}-\hat{v}_h)\frac{\partial\overline{\psi}}{\partial\nu_j}\mathrm{d}x_1.\end{aligned}$$

Therefore, by using (3.10) and (3.23), we conclude that

$$\begin{aligned}b^{\mathrm{PML}}(\hat{v}-\hat{v}_h,\psi)&=a_D(\hat{v}-\hat{v}_h,\psi)\\&=\int_D g(\overline{\psi-\psi_h})\mathrm{d}x-a_D(\hat{v}_h,\psi-\psi_h).\end{aligned}$$

This completes the proof. □

## 4.2 Estimates for the extension

**Lemma 4.3.** *For any $\psi\in X(\Omega)$, let $\psi$ be extended to the whole domain $D$ according to* (4.1). *Then we have the following estimates, for $j=1,2$:*

$$\left\|s^{-1}\nabla\psi\right\|_{L^2(\Omega_j^{\mathrm{PML}})}\le C_j\|\psi\|_{H^1(\Omega)},$$

*with $C_j=\min(C_j^a,C_j^b)$, where*

$$C_j^a=\begin{cases}\hat{C}\max\left(\frac{2\sqrt{2}|k_j|\delta_j^{1/2}}{1-\exp(-2\Delta_j^-\sigma_j^I)},\frac{2(|k_j|+\delta_j\mathrm{Re}(k_j^2)+1)^{1/2}}{1-\exp(-2\Delta_j^+\sigma_j^R)}\right) & \text{if } \Delta_j^+ \text{ and } \Delta_j^->0,\\ +\infty & \text{otherwise,}\end{cases}$$

$$C_j^b=\begin{cases}\hat{C}\frac{2[\max(1,|k_j|)(1+2\delta_j(\mathrm{Im}k_j+|k_j|))]^{1/2}}{1-\exp(-2\sigma_j^R\mathrm{Im}k_j)} & \text{if } \mathrm{Im}\varepsilon_j>0,\\ +\infty & \text{otherwise.}\end{cases}$$

*Proof.* We define

$$r_1(x_3)=\int_{x_3}^{b_1+\delta_1} s(\tau)\mathrm{d}\tau, \qquad r_2(x_3)=\int_{b_2-\delta_2}^{x_3} s(\tau)\mathrm{d}\tau.$$

Then it follows that $\zeta_j^n(x_3)=e^{-\mathbf{i}\beta_j^n r_j(x_3)}-e^{\mathbf{i}\beta_j^n r_j(x_3)}$ and consequently

$$\frac{d\overline{\zeta}_j^n}{\mathrm{d}x_3}=\mathbf{i}\overline{\beta}_j^n(-1)^j\overline{s}(x_3)\left[e^{\mathbf{i}\overline{\beta}_j^n\overline{r}_j(x_3)}+e^{-\mathbf{i}\overline{\beta}_j^n\overline{r}_j(x_3)}\right].$$

By direct calculation, we deduce from (4.1) that

$$\begin{aligned}\int_0^L|\nabla\psi|^2\mathrm{d}x_1=&L\sum_{n\in\mathbf{Z}}|\alpha_n+\alpha|^2|e^{-\mathbf{i}\beta_j^n r_j(x_3)}-e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2|\zeta_j^n(b_j)|^{-2}|\psi_\alpha^{(n)}(b_j)|^2\\&+L\sum_{n\in\mathbf{Z}}|\beta_j^n|^2|s(x_3)|^2|e^{-\mathbf{i}\beta_j^n r_j(x_3)}+e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2|\zeta_j^n(b_j)|^{-2}|\psi_\alpha^{(n)}(b_j)|^2.\end{aligned}$$

We denote $(b_1,b_1+\delta_1)$, $(b_2-\delta_2,b_2)$ by $I_1$ and $I_2$ respectively. Then

$$\begin{aligned}&\left\|s^{-1}\nabla\psi\right\|^2_{L^2(\Omega_j^{\mathrm{PML}})}=\int_{I_j}|s(x_3)|^{-2}\int_0^L|\nabla\psi(x_1,x_3)|^2\mathrm{d}x_1\mathrm{d}x_3\\&\le L\int_{I_j}\sum_{n\in\mathbf{Z}}\max_{\pm}|e^{-\mathbf{i}\beta_j^n r_j(x_3)}\pm e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2|\zeta_j^n(b_j)|^{-2}|\psi_\alpha^{(n)}(b_j)|^2\times\left[|\alpha_n+\alpha|^2+|\beta_j^n|^2\right]\mathrm{d}x_3,\end{aligned}\tag{4.5}$$

where

$$|e^{-\mathbf{i}\beta_j^n r_j(x_3)}\pm e^{\mathbf{i}\beta_j^n r_j(x_3)}|^2|\zeta_j^n(b_j)|^{-2}\le\left|\frac{e^{\mathrm{Re}(-\mathbf{i}\beta_j^n r_j(x_3))}+e^{\mathrm{Re}(\mathbf{i}\beta_j^n r_j(x_3))}}{e^{\mathrm{Re}(-\mathbf{i}\beta_j^n r_j(b_j))}-e^{\mathrm{Re}(\mathbf{i}\beta_j^n r_j(b_j))}}\right|^2.\tag{4.6}$$

For $n\in U_j$, we have $\mathrm{Re}\beta_j^n\ge\Delta_j^n\ge\Delta_j^-$, $\mathrm{Im}\beta_j^n\ge 0$; and thus

$$\mathrm{Re}(-\mathbf{i}\beta_j^n r_j(x_3))=\mathrm{Re}\beta_j^n r_j^I(x_3)+\mathrm{Im}\beta_j^n r_j^R(x_3)\ge\mathrm{Re}\beta_j^n r_j^I(x_3),$$

where $r_j^R, r_j^I$ are the real and imaginary parts of $r_j$, i.e., $r_j=r_j^R+\mathbf{i}r_j^I$. Then the right-hand side of (4.6) is bounded as

$$\begin{aligned}&\left|\frac{e^{\mathrm{Re}(-\mathbf{i}\beta_j^n r_j(x_3))}+e^{\mathrm{Re}(\mathbf{i}\beta_j^n r_j(x_3))}}{e^{\mathrm{Re}(-\mathbf{i}\beta_j^n r_j(b_j))}-e^{\mathrm{Re}(\mathbf{i}\beta_j^n r_j(b_j))}}\right|^2=e^{2\mathrm{Re}(-\mathbf{i}\beta_j^n(r_j(x_3)-r_j(b_j)))}\left|\frac{1+e^{2\mathrm{Re}(\mathbf{i}\beta_j^n r_j(x_3))}}{1-e^{2\mathrm{Re}(\mathbf{i}\beta_j^n r_j(b_j))}}\right|^2\\&\le\frac{4e^{2\mathrm{Re}\beta_j^n(-1)^j\int_{b_j}^{x_3}s_2(\tau)\mathrm{d}\tau}}{(1-e^{-2\mathrm{Re}\beta_j^n r_j^I(b_j)})^2}\le\frac{4}{(1-e^{-2\Delta_j^-\sigma_j^I})^2}\qquad\text{if }\Delta_j^->0.\end{aligned}$$

Similarly, for $n\notin U_j$, we have $\mathrm{Re}\beta_j^n\ge 0$, and $\mathrm{Im}\beta_j^n\ge\Delta_j^n\ge\Delta_j^+$; and thus

$$\mathrm{Re}(-\mathbf{i}\beta_j^n r_j(x_3))\ge\mathrm{Im}\beta_j^n r_j^R(x_3).$$

Then the right-hand side of (4.6) is bounded as

$$\begin{aligned}&\left|\frac{e^{\mathrm{Re}(-\mathrm{i}\beta_j^n r_j(x_3))}+e^{\mathrm{Re}(\mathrm{i}\beta_j^n r_j(x_3))}}{e^{\mathrm{Re}(-\mathrm{i}\beta_j^n r_j(b_j))}-e^{\mathrm{Re}(\mathrm{i}\beta_j^n r_j(b_j))}}\right|^2\\ &\leq\frac{4(e^{2\mathrm{Im}\beta_j^n(-1)^j\int_{b_j}^{x_3}s_1(\tau)\mathrm{d}\tau})}{(1-e^{-2\mathrm{Im}\beta_j^n r_j^R(b_j)})^2}\leq\frac{4e^{-2\mathrm{Im}\beta_j^n|x_3-b_j|}}{(1-e^{-2\Delta_j^+\sigma_j^R})^2}\quad\text{if }\Delta_j^+>0.\end{aligned}$$

We have used the fact that $s_1(\tau)\geq 1$ in the above deduction.

Consequently, it follows that

$$\begin{aligned}&\left\|s^{-1}\nabla\psi\right\|^2_{L^2(\Omega_j^{\mathrm{PML}})}\\ \leq& L\int_{I_j}\Bigg\{\sum_{n\in U_j}\frac{4}{(1-e^{-2\Delta_j^-\sigma_j^I})^2}|\psi_\alpha^{(n)}(b_j)|^2(|\alpha_n+\alpha|^2+|\beta_j^n|^2)\\ &+\sum_{n\notin U_j}\frac{4e^{-2\mathrm{Im}\beta_j^n|x_3-b_j|}}{(1-e^{-2\Delta_j^+\sigma_j^R})^2}|\psi_\alpha^{(n)}(b_j)|^2(|\alpha_n+\alpha|^2+|\beta_j^n|^2)\Bigg\}\mathrm{d}x_3.\end{aligned}\tag{4.7}$$

Moreover, it is easy to get the following estimate:

$$\int_{I_j}e^{-2\mathrm{Im}\beta_j^n|x_3-b_j|}\mathrm{d}x_3\leq\min\left(\delta_j,\frac{1}{2\mathrm{Im}\beta_j^n}\right).$$

Substituting the above estimate into (4.7) yields

$$\begin{aligned}&\left\|s^{-1}\nabla\psi\right\|^2_{L^2(\Omega_j^{\mathrm{PML}})}\\ \leq&\sum_{n\in U_j}\frac{4L\delta_j}{(1-e^{-2\Delta_j^-\sigma_j^I})^2}|\psi_\alpha^{(n)}(b_j)|^2(|\alpha_n+\alpha|^2+|\beta_j^n|^2)\\ &+\sum_{n\notin U_j}\frac{4L\min(\delta_j,\frac{1}{2\mathrm{Im}\beta_j^n})}{(1-e^{-2\Delta_j^+\sigma_j^R})^2}|\psi_\alpha^{(n)}(b_j)|^2(|\alpha_n+\alpha|^2+|\beta_j^n|^2)\\ :=&\mathrm{I}+\mathrm{II}.\end{aligned}$$

If $n\in U_j$, then $|\alpha_n+\alpha|^2+|\beta_j^n|^2\leq 2|k_j|^2$. Consequently,

$$\mathrm{I}\leq(C_j^a)^2\hat{C}^{-2}L\sum_{n\in U_j}|\psi_\alpha^{(n)}(b_j)|^2.$$

If $n \notin U_j$, using (3.16) and (3.17) yields

$$\begin{aligned}
&|\alpha_n+\alpha|^2+|\beta_j^n|^2=|\alpha_n+\alpha|^2+2\sqrt{\xi_n^2+\eta^2}\\
=&|\alpha_n+\alpha|^2+2[(\mathrm{Im}\beta_j^n)^2+\xi_n]=2(\mathrm{Im}\beta_j^n)^2+\mathrm{Re}(k_j^2)\\
\leq&2\mathrm{Im}\beta_j^n\cdot\sqrt{2}(\xi_n^2+\eta^2)^{\frac{1}{4}}+\mathrm{Re}(k_j^2)\\
\leq&2\mathrm{Im}\beta_j^n\left[(\mathrm{Re}(k_j^2)-|\alpha_n+\alpha|^2)^2+(\mathrm{Im}(k_j^2))^2\right]^{\frac{1}{4}}+\mathrm{Re}(k_j^2)\\
\leq&2\mathrm{Im}\beta_j^n\left[|\alpha_n+\alpha|^4+|k_j|^4\right]^{\frac{1}{4}}+\mathrm{Re}(k_j^2)\leq 2\mathrm{Im}\beta_j^n(|\alpha_n+\alpha|+|k_j|)+\mathrm{Re}(k_j^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&(|\alpha_n+\alpha|^2+\left|\beta_j^n\right|^2)\min\left(\delta_j,\frac{1}{2\mathrm{Im}\beta_j^n}\right)\\
\leq&|\alpha_n+\alpha|+|k_j|+\delta_j\mathrm{Re}(k_j^2)\\
\leq&\left(|k_j|+\delta_j\mathrm{Re}(k_j^2)+1\right)\left(1+|\alpha_n+\alpha|^2\right)^{\frac{1}{2}},
\end{aligned}$$

which yields

$$\mathrm{II}\leq(C_j^a)^2\hat{C}^{-2}L\sum_{n\notin U_j}(1+|\alpha_n+\alpha|^2)^{\frac{1}{2}}|\psi_\alpha^{(n)}(b_j)|^2.$$

Thus, applying Lemma 3.3 we get

$$\left\|s^{-1}\nabla\psi\right\|_{L^2(\Omega_j^{\mathrm{PML}})}\leq C_j^a\|\psi\|_{H^1(\Omega)}.$$

Finally, from [16], we have

$$\left\|s^{-1}\nabla\psi\right\|_{L^2(\Omega_j^{\mathrm{PML}})}\leq C_j^b\|\psi\|_{H^1(\Omega)}.$$

This completes our proof. □

### 4.3 Proof of Theorem 3.2

To interpolate non-smooth functions which satisfies quasi-periodic boundary conditions, we resort to an interpolation operator

$$\Pi_h:\overset{\circ}{X}(D)\to\overset{\circ}{V}_h(D)$$

of Scott and Zhang in [10].

Let $N_h=\{a_i\}_{i=1}^N$ be the set of all nodes of $\mathcal{M}_h$, and $\{\phi_i\}_{i=1}^N$ be the corresponding nodal basis of $V_h(D)$. For any interior node $a_i$ in $D$, we use $\sigma_i$ to denote an edge $e$ which contains $a_i$ as its vertex. And for any node $a_i$ that is in the interior of the left boundary of the domain, i.e. $a_i=(0,z_i)$ for some $z_i\in(b_2-\delta_2,b_1+\delta_1)$, we also use $\sigma_i$ to denote an edge on the left boundary with $a_i$ as its vertex. Similar denotations can be used for the case of right boundary. Finally, in the case that $a_i$ lies on $\overline{\Gamma_1^{\rm PML}\cup\Gamma_2^{\rm PML}}$, any side on $\overline{\Gamma_1^{\rm PML}}$ or $\overline{\Gamma_2^{\rm PML}}$ containing $a_i$ as a vertex can be used.

Let $a_{i,1}=a_i$ and $\{a_{i,j}\}_{j=1}^2$ be the set of nodal points in $\sigma_i$ with nodal basis $\{\phi_{i,j}\}_{j=1}^2$. Define $\{\psi_{i,j}\}_{j=1}^2$ to be the $L^2(\sigma_i)$ dual basis:

$$\int_{\sigma_i}\psi_{i,j}(x)\phi_{i,k}(x)ds=\delta_{jk},\qquad j,k=1,2,$$

where $\delta_{jk}$ is the Kronecker delta, and let $\psi_i=\psi_{i,1}$. Then the interpolation operator $\Pi_h$: $H^1(D)\to W_h(D)$ (the confirming linear finite element space) is defined by

$$\Pi_h\varphi(x)=\sum_{i=1}^N\phi_i(x)\int_{\sigma_i}\psi_i(x)\varphi(x)ds.$$

The operator $\Pi_h$ enjoys the following estimates ( [10]):

$$\|\varphi-\Pi_h\varphi\|_{L^2(T)}\le Ch_T\|\nabla\varphi\|_{L^2(\tilde{T})},\quad \|\varphi-\Pi_h\varphi\|_{L^2(e)}\le Ch_e^{1/2}\|\nabla\varphi\|_{L^2(\tilde{e})},\tag{4.8}$$

where $\tilde{T}$ and $\tilde{e}$ are the union of the elements in $\mathcal{M}_h$, which have nonempty intersection with $T\in\mathcal{M}_h$ and the side $e$, respectively.

It remains to check whether $\Pi_h$ keeps the boundary condition. It is clear that $\Pi_h\varphi=0$ on $\Gamma_1^{\rm PML}\cup\Gamma_2^{\rm PML}$ since for any $a_i\in\overline{\Gamma_1^{\rm PML}\cup\Gamma_2^{\rm PML}}$, $\sigma_i\subset\overline{\Gamma_1^{\rm PML}}$ or $\overline{\Gamma_2^{\rm PML}}$ and $\varphi=0$ on $\Gamma_1^{\rm PML}\cup\Gamma_2^{\rm PML}$. Now let $a_i=(0,z_i)\in\Gamma_{\rm left}$ and $a_k=(L,z_i)\in\Gamma_{\rm right}$. Without loss of generality, we assume $\sigma_i=\{x\in\mathbf{R}^2:x_1=0,z_i\le x_3\le z_{i+1}\}$ and it follows that $\sigma_k=\{x\in\mathbf{R}^2:x_1=L,z_i\le x_3\le z_{i+1}\}$. The nodal basis

$$\phi_{i,1}=(z_{i+1}-x_3)/(z_{i+1}-z_i),\quad \phi_{i,2}=(x_3-z_i)/(z_{i+1}-z_i)\quad\text{in }\sigma_i,$$

and simple calculation yields the dual basis

$$\psi_{i,1}=\frac{4}{d_i}\phi_{i,1}-\frac{2}{d_i}\phi_{i,2},\qquad \psi_{i,2}=-\frac{2}{d_i}\phi_{i,1}+\frac{4}{d_i}\phi_{i,2}\quad\text{in}\quad\sigma_i,$$

where $d_i=z_{i+1}-z_i$. Similar computation implies that

$$\psi_k(L,x_3)=\psi_i(0,x_3)=\frac{4}{d_i}\phi_{i,1}(x_3)-\frac{2}{d_i}\phi_{i,2}(x_3).$$

Thus for any $\varphi\in X(D)$, $\varphi(0,x_3)=e^{-\mathrm{i}\alpha L}\varphi(L,x_3)$, we have

$$\begin{aligned}\Pi_h\varphi(a_i)&=\int_{\sigma_i}\psi_i(0,x_3)\varphi(0,x_3)\mathrm{d}x_3\\&=e^{-\mathrm{i}\alpha L}\int_{\sigma_k}\psi_k(L,x_3)\varphi(L,x_3)\mathrm{d}x_3\\&=e^{-\mathrm{i}\alpha L}\Pi_h\varphi(a_k),\end{aligned}$$

which shows that $\Pi_h\varphi\in\overset{\circ}{V}_h(D)$ if $\varphi\in\overset{\circ}{X}(D)$.

Now we prove Theorem 3.2.

*Proof of Theorem 3.2.* We choose $\psi_h$ to be $\Pi_h\psi\in\overset{\circ}{V}_h(D)$ in the error representation formula (4.4) and it follows that

$$\begin{aligned}b(v-\hat{v}_h,\psi)&=\int_D g(\overline{\psi-\Pi_h\psi})\mathrm{d}x-a_{_D}(\hat{v}_h,\psi-\Pi_h\psi)\\&\quad+\sum_{j=1}^{2}\int_{\Gamma_j}(T_j-T_j^{\mathrm{PML}})\hat{v}_h\overline{\psi}\mathrm{d}x_1\\&:=\mathrm{III}+\mathrm{IV}+\mathrm{V}.\end{aligned}\tag{4.9}$$

Performing integration by parts and applying (3.24)-(3.26) we obtain

$$\mathrm{III}+\mathrm{IV}=\sum_{T\in\mathcal{M}_h}\left(\int_T R_T(\overline{\psi-\Pi_h\psi})\mathrm{d}x+\sum_{e\subset\partial T}\frac{1}{2}\int_e J_e(\overline{\psi-\Pi_h\psi})ds\right).$$

Combining (4.8) with Lemma 4.3 gives

$$\begin{aligned}|\mathrm{III}+\mathrm{IV}|&\le C\sum_{T\in\mathcal{M}_h}\eta_T\left\|s^{-1}\nabla\psi\right\|_{L^2(\tilde{T})}\\&\le C(1+C_1+C_2)\left(\sum_{T\in\mathcal{M}_h}\eta_T^2\right)^{1/2}\|\psi\|_{H^1(\Omega)}.\end{aligned}\tag{4.10}$$

Lemmas 3.2 and 3.3 together show that

$$|\mathrm{V}|\le\sum_{j=1}^{2}M_j\|\hat{v}_h\|_{L^2(\Gamma_j)}\|\psi\|_{L^2(\Gamma_j)}\le\sum_{j=1}^{2}\hat{C}M_j\|\hat{v}_h\|_{L^2(\Gamma_j)}\|\psi\|_{H^1(\Omega)}.\tag{4.11}$$

Combining (4.9)-(4.11), we obtain the desired estimate. □

### 4.4 The lower bound of $\|\hat{v}-\hat{v}_h\|$

In this section a lower bound of $\|\hat{v}-\hat{v}_h\|$ is given which illustrates the a posteriori error estimate is a sharp one.

For any side $e\in\mathcal{B}_h$ which is an edge of an element $T_1$, define $\omega_e=T_1\cup T_2$. If $e$ is an interior edge, then $T_2$ is the other element sharing $e$; if $e\in\Gamma_{\text{left}}$ or $\Gamma_{\text{right}}$, then $T_2$ is the element whose one edge is $e'$, where $e'$ is the corresponding edge of $e$ on $\Gamma_{\text{right}}$ or $\Gamma_{\text{left}}$. For any $T\in\mathcal{M}_h$, $\omega_T$ is used to denote the domain which consists of all elements sharing at least one side with $T$, i.e., $\omega_T=\cup_{e\subset\partial T}\omega_e$ and $R_T^a\in\mathrm{P}_{n-1}(T)$ is the $L^2$-projection of $R_T$ onto $\mathrm{P}_{n-1}(T)$, the space of polynomials with degree$\leq n-1$ over $T$. Here $n$ is the order of the finite element space $\overset{\circ}{V}_h$. And we define the *oscillation* on the element $T\in\mathcal{M}_h$ by

$$\mathrm{osc}_h(T):=h_T\|R_T-R_T^a\|_{L^2(T)}. \tag{4.12}$$

For a subset $\omega\subset D$, we have

$$\mathrm{osc}_h(\omega)^2:=\sum_{T\in\mathcal{M}_h,T\subset\omega}\mathrm{osc}_h(T)^2.$$

**Theorem 4.1** (lower bound)**.** *There exist constants $C_3$ and $C_4$, depending on the minimum angle of $\mathcal{M}_h$ and the given data, such that*

$$\eta_T^2\leq C_3\|\hat{v}-\hat{v}_h\|_{H^1(\omega_T)}^2+C_4\mathrm{osc}_h(\omega_T)^2\qquad\forall\,T\in\mathcal{M}_h. \tag{4.13}$$

*Proof.* Applying (3.10) and performing integration by parts give

$$a_{_D}(\hat{v}-\hat{v}_h,\psi)=\sum_{T\in\mathcal{M}_h}\int_T(R_T^a\overline{\psi}+(R_T-R_T^a)\overline{\psi})\,\mathrm{d}x+\sum_{e\in\mathcal{B}_h}\int_e J_e\overline{\psi}\,\mathrm{d}s\qquad\forall\,\psi\in\overset{\circ}{X}(D). \tag{4.14}$$

As in [9], we proceed in three steps.

*Step* 1. *Interior residual*. For $T\in\mathcal{M}_h$, let $\phi_T\in\overset{\circ}{X}(D)$ be a bubble function on element $T$ with the form:

$$\phi_T=27\lambda_1\lambda_2\lambda_3,$$

where $\lambda_i$ are the barycentric coordinates on element $T$, and $\phi_T$ satisfies $0\leq\phi_T\leq1$ and vanishes on $\partial T$, i.e., $\mathrm{supp}\,\phi_T\subset T$. Since $R_T^a\in\mathrm{P}_{n-1}(T)$, we have

$$C\|R_T^a\|_{L^2(T)}^2\leq\int_T\phi_T|R_T^a|^2\,\mathrm{d}x=\int_T R_T^a(\overline{\phi_T R_T^a})\,\mathrm{d}x.$$

Since $\phi_T R_T^a$ vanishes outside $T$ (and in particular on all $e\in\mathcal{B}_h$), it follows that

$$\begin{aligned}C\|R_T^a\|_{L^2(T)}^2&\leq a_{_D}(\hat{v}-\hat{v}_h,\phi_T R_T^a)+\int_T(R_T^a-R_T)\overline{\phi_T R_T^a}\,\mathrm{d}x\\&\leq C(h_T^{-1}\|\hat{v}-\hat{v}_h\|_{H^1(T)}+\|R_T-R_T^a\|_{L^2(T)})\,\|R_T^a\|_{L^2(T)},\end{aligned}$$

which is based on the following inverse inequality for $\phi_T R_T^a$:

$$\|\phi_T R_T^a\|_{H^1(T)} \le C h_T^{-1} \|\phi_T R_T^a\|_{L^2(T)} \le C h_T^{-1} \|R_T^a\|_{L^2(T)}.$$

By the triangle inequality, we get the estimate

$$h_T^2 \|R_T\|_{L^2(T)}^2 \le C(\|\hat{v}-\hat{v}_h\|_{H^1(T)}^2 + h_T^2 \|R_T - R_T^a\|_{L^2(T)}^2). \tag{4.15}$$

*Step* 2. *Jump residual*. Let $e \in \mathcal{B}_h$ be an interior side, and $T_1, T_2 \in \mathcal{M}_h$ be the two elements sharing $e$. Let $\phi_e \in \overset{\circ}{X}(D)$ be a bubble function in $\omega_e$ with the form:

$$\phi_e = 4\lambda_1\lambda_2,$$

where $\lambda_i$ are the barycentric coordinates corresponding to the vertexes of $e$, and $\phi_e$ satisfies $0 \le \phi_e \le 1$ and vanishes on $\partial\omega_e$, i.e., $\operatorname{supp}\phi_e \subset \omega_e$.

Since $\hat{v}_h$ is continuous, $[\![\nabla\hat{v}_h]\!]_e := \nabla\hat{v}_h|_{T_1} - \nabla\hat{v}_h|_{T_2}$ is parallel to $\nu_e$, i.e., $[\![\nabla\hat{v}_h]\!]_e = j_e\nu_e$. Moreover, the coefficient matrix $A(x)$ being continuous implies

$$J_e = A(x)[\![\nabla\hat{v}_h]\!]_e \cdot \nu_e = j_e A(x)\nu_e \cdot \nu_e = a(x) j_e,$$

where $a(x) = A(x)\nu_e \cdot \nu_e$ and $a(x)$ satisfies $0 < \underline{a}_e \le |a(x)| \le \overline{a}_e$ with

$$\underline{a}_e = \min_{x\in e} \frac{1}{|s|} \qquad \overline{a}_e = \max_{x\in e} |s|.$$

Consequently,

$$\|J_e\|_{L^2(e)}^2 \le \overline{a}_e^2 \int_e |j_e|^2 \,\mathrm{d}s \le C\overline{a}_e^2 \int_e |j_e|^2 \phi_e \,\mathrm{d}s = C\overline{a}_e^2 \int_e \overline{j_e\phi_e} \frac{1}{a(x)} J_e \,\mathrm{d}s, \tag{4.16}$$

where the second inequality follows from the fact that $j_e$ is a polynomial.

We point out that $j_e$ can be extended to $\omega_e$ in the following manner: first it is mapped to the reference element, then it is extended constantly along the normal to $\hat{e}$ which is the corresponding side of $e$ in the reference element, and finally we map it back to $\omega_e$. The resulting $E_h(j_e)$ is a piecewise polynomial in $\omega_e$ so that $\phi_e E_h(j_e) \in \overset{\circ}{X}(D)$, which satisfies

$$\|\phi_e E_h(j_e)\|_{L^2(\omega_e)} \le C h_e^{1/2} \|j_e\|_{L^2(e)}.$$

Since $\psi = \phi_e E_h(j_e)/\overline{a(x)}$ is an admissible test function in (4.14) which vanishes on all sides of $\mathcal{B}_h$ but $e$, we arrive at

$$\begin{aligned}
\int_e \overline{j_e\phi_e} \frac{1}{a(x)} J_e \,\mathrm{d}s &= a_D(\hat{v}-\hat{v}_h, \psi) - \int_{T_1} R_{T_1}\overline{\psi}\,\mathrm{d}x - \int_{T_2} R_{T_2}\overline{\psi}\,\mathrm{d}x \\
&\le C\|\hat{v}-\hat{v}_h\|_{H^1(\omega_e)} \|\psi\|_{H^1(\omega_e)} + \sum_{j=1}^{2} \left\| R_{T_j} \right\|_{L^2(T_j)} \|\psi\|_{L^2(T_j)} \\
&\le C\left( h_e^{-1/2} \|\hat{v}-\hat{v}_h\|_{H^1(\omega_e)} + h_e^{1/2} \sum_{i=1}^{2} \|R_{T_i}\|_{L^2(T_i)} \right) \|j_e\|_{L^2(e)}.
\end{aligned} \tag{4.17}$$

Therefore,

$$h_e\|J_e\|^2_{L^2(e)} \le C\Big(\|\hat{v}-\hat{v}_h\|^2_{H^1(\omega_e)}+\sum_{i=1}^{2}h^2_{T_i}\|R_{T_i}\|^2_{L^2(T_i)}\Big). \tag{4.18}$$

If $e\subset\Gamma_{\text{left}}$ or $\Gamma_{\text{right}}$, let $e'$ be the corresponding edge on $\Gamma_{\text{right}}$ or $\Gamma_{\text{left}}$. By noticing that $\|J_e\|_{L^2(e)}=\|J_{e'}\|_{L^2(e')}$ and using a similar argument as above, it can be shown that (4.18) holds too.

*Step* 3. *Final estimate*. To remove the interior residual from the right-hand side of (4.18), we obtain from (4.15) that

$$h_e\|J_e\|^2_{L^2(e)} \le C\Big(\|\hat{v}-\hat{v}_h\|^2_{H^1(\omega_e)}+\sum_{i=1}^{2}h^2_{T_i}\big\|R_{T_i}-R^a_{T_i}\big\|^2_{L^2(T_i)}\Big). \tag{4.19}$$

Combining (4.15) with (4.19) gives the estimate for $\eta^2_T$:

$$\begin{aligned}\eta^2_T&=\left\{\max_{x\in T}|s(x_3)|\left[h_T\|R_T\|_{L^2(T)}+(\frac{1}{2}\sum_{e\subset T}h_e\|J_e\|^2_{L^2(e)})^{1/2}\right]\right\}^2\\&\le C\Big(h^2_T\|R_T\|^2_{L^2(T)}+\sum_{e\subset T}h_e\|J_e\|^2_{L^2(e)}\Big)\\&\le C_3\|\hat{v}-\hat{v}_h\|^2_{H^1(\omega_T)}+C_4\text{osc}_h(\omega_T)^2,\end{aligned}$$

where the constants $C_3$ and $C_4$ depend only on the minimum angle of $\mathcal{M}_h$ and $s(x)$. □

# 5 TM polarization

In this section the main results of error estimates for the grating problem (2.4) in TM polarization are presented; whose proofs are omitted as they are similar to those employed in the case of TE polarization.

The sesquilinear form $b_{\text{TM}}: X(\Omega)\times X(\Omega)\longrightarrow\mathbb{C}$ is defined as:

$$b_{\text{TM}}(\varphi,\psi)=\int_\Omega(\frac{1}{k^2(x)}\nabla\varphi\nabla\overline{\psi}-\varphi\overline{\psi})\,\mathrm{d}x-\sum_{j=1}^{2}\int_{\Gamma_j}\frac{1}{k_j^2}(T_j\varphi)\overline{\psi}\,\mathrm{d}x_1. \tag{5.1}$$

Then the variational form for the 1D grating problem in the TM polarization reads as follows: Given an incoming plane wave $u_{\text{I}}=e^{\mathrm{i}\alpha x_1-\mathrm{i}\beta x_3}$, seek $v^{\text{TM}}\in X(\Omega)$ such that

$$b_{\text{TM}}(v^{\text{TM}},\psi)=\int_\Omega g_{\text{TM}}\overline{\psi}\,\mathrm{d}x\qquad\forall\,\psi\in X(\Omega), \tag{5.2}$$

where $g_{\text{TM}}=g/k_1^2$.

Assume that the above variational problem has a unique solution. Then it follows that there exists a constant $\gamma_{TM}>0$ satisfying

$$\sup_{0\neq\psi\in H^1(\Omega)}\frac{|b_{\text{TM}}(\varphi,\psi)|}{\|\psi\|_{H^1(\Omega)}}\ge\gamma_{TM}\|\varphi\|_{H^1(\Omega)}\qquad\forall\,\varphi\in X(\Omega). \tag{5.3}$$

Moreover, the sesquilinear form $a_{\rm TM}: X(D)\times X(D)\longrightarrow\mathbb{C}$ associated with the PML problem is defined as

$$a_{\rm TM}(\varphi,\psi)=\int_D\Big(\frac{1}{k^2(x)}s(x_3)\frac{\partial\varphi}{\partial x_1}\frac{\partial\overline{\psi}}{\partial x_1}+\frac{1}{k^2(x)}\frac{1}{s(x_3)}\frac{\partial\varphi}{\partial x_3}\frac{\partial\overline{\psi}}{\partial x_3}-s(x_3)\varphi\overline{\psi}\Big)\mathrm{d}x.$$

Accordingly, the weak formulation of the modified PML model reads as: Find $\hat{v}^{\rm TM}\in\overset{\circ}{X}(D)$ such that

$$a_{\rm TM}(\hat{v}^{\rm TM},\psi)=\int_D g_{\rm TM}\overline{\psi}\mathrm{d}x\quad\forall\,\psi\in\overset{\circ}{X}(D). \tag{5.4}$$

We now present the theorem for the error estimate of the PML problem, which is an analogue of Theorem 3.1.

**Theorem 5.1.** *Let $\gamma_{\rm TM}>0$ be the constant in the inf-sup condition* (5.3) *and*

$$\sum_{j=1}^{2}M_j\hat{C}^2/k_j^2<\gamma_{\rm TM}.$$

*Then the modified PML variational problem* (5.4) *has a unique solution $\hat{v}^{\rm TM}$. Moreover, we have the following error estimate:*

$$|||v^{\rm TM}-\hat{v}^{\rm TM}|||_\Omega^{\rm TM}:=\sup_{0\neq\psi\in H^1(\Omega)}\frac{|b_{\rm TM}(v^{\rm TM}-\hat{v}^{\rm TM},\psi)|}{\|\psi\|_{H^1(\Omega)}}\leq\hat{C}\sum_{j=1}^{2}\frac{M_j}{k_j^2}\|\hat{v}\|_{L^2(\Gamma_j)}. \tag{5.5}$$

The finite element approximation to the TM polarization problem (5.4) is defined as: Find $\hat{v}_h^{\rm TM}\in\overset{\circ}{V}_h(D)$ such that

$$a_{\rm TM}(\hat{v}_h^{\rm TM},\psi_h)=\int_D g_{\rm TM}\overline{\psi}_h\mathrm{d}x\quad\forall\,\psi_h\in\overset{\circ}{V}_h(D). \tag{5.6}$$

Let

$$A_{\rm TM}(x)=A(x)/k^2(x),\quad B_{\rm TM}(x)=B(x)/k^2(x),\quad\mathcal{L}_{\rm TM}=\mathrm{div}(A_{\rm TM}(x)\nabla)+B_{\rm TM}(x).$$

Then we have the following theorem analogous to Theorem 3.2.

**Theorem 5.2.** *There exists a constant $C>0$, depending only on the minimum angle of the mesh $\mathcal{M}_h$, such that the following a posteriori error estimate is valid:*

$$|||v^{\rm TM}-\hat{v}_h^{\rm TM}|||_\Omega^{\rm TM}\leq\sum_{j=1}^{2}\hat{C}\frac{M_j}{k_j^2}\left\|\hat{v}_h^{\rm TM}\right\|_{L^2(\Gamma_j)}+C(1+C_1+C_2)\Big(\sum_{T\in\mathcal{M}_h}\eta_T^2\Big)^{1/2},$$

*where the constants $M_j$ $(j=1,2)$, $\hat{C}$,$C_j$ are defined in Lemmas* 3.2, 3.3 *and* 4.3, *respectively; $\eta_T$ is defined similar to that in* (3.27), *with $A,\mathcal{L},g$, and $\hat{v}_h$ being replaced by $A_{\rm TM}$,$\mathcal{L}_{\rm TM}$,$g_{\rm TM}$, and $\hat{v}_h^{\rm TM}$, respectively.*

# 6 Adaptive algorithm and a simple numerical example

In this section an adaptive finite element algorithm is proposed, which is a modification of the algorithm in [4]. A simple numerical example will be presented to demonstrate the effectiveness of the proposed algorithm.

## 6.1 Adaptive algorithm

We use the a posteriori error estimate in Theorem 3.2 to determine the PML parameters. The PML medium property $s(x_3)$ is chosen as the power function

$$s(x_3)=\begin{cases}1+\sigma_1^m\left(\dfrac{x_3-b_1}{\delta_1}\right)^m & \text{if } x_3\geq b_1,\\ 1+\sigma_2^m\left(\dfrac{b_2-x_3}{\delta_2}\right)^m & \text{if } x_3\leq b_2,\end{cases}$$

where $m\geq 1$, $\sigma_j^m, j=1,2$ are medium parameters. Thus we have

$$\sigma_j^R=\left(1+\frac{\mathrm{Re}\sigma_j^m}{m+1}\right)\delta_j,\quad \sigma_j^I=\frac{\mathrm{Im}\sigma_j^m}{m+1}\,\delta_j. \tag{6.1}$$

If follows that we only need to specify the thickness $\delta_j$ of the layers and the medium parameters $\sigma_j^m$. Recall from Theorem 3.2, we know that the a posteriori error estimate consists of two parts: the PML error $\mathcal{E}_{\mathrm{PML}}$ and the finite element discretization error $\mathcal{E}_{\mathrm{FEM}}$, where

$$\mathcal{E}_{\mathrm{PML}}=\sum_{j=1}^{2}M_j\|\hat{v}_h\|_{L^2(\Gamma_j)}, \tag{6.2}$$

$$\mathcal{E}_{\mathrm{FEM}}=\Big(\sum_{T\in\mathcal{M}_h}\eta_T^2\Big)^{1/2}. \tag{6.3}$$

Hence, $\mathcal{E}_{\mathrm{PML}}$ and $\mathcal{E}_{\mathrm{FEM}}$ should be changed accordingly in the TM case. In our implementation we first choose $\delta_j$ and $\sigma_j^m$ such that $M_jL^{1/2}\leq 10^{-8}$, which makes the PML error negligible compared with the finite element discretization errors. Once the PML region and the medium property are fixed, the standard finite element adaptive strategy is utilized to modify the mesh according to the a posteriori error estimate (6.3). For any $T\in\mathcal{M}_h$, we define the local a posteriori error estimator as $\eta_T$. The estimators are employed to make local mesh modifications by refinement to equidistribute the approximation errors and, as a consequence, to equidistribute the computational load. This naturally leads to the adaptation loop of the form

$$\mathsf{Solve}\longrightarrow\mathsf{Estimate}\longrightarrow\mathsf{Mark}\longrightarrow\mathsf{Refine}.$$

Now a modified adaptive algorithm goes as follows.

Algorithm 6.1:

Given a tolerance TOL$>0$. Let $m=2$.

- Choose $\delta_1$,$\delta_2$, and $\sigma_j^m$ such that $M_j L^{1/2} \leq 10^{-8}$ $(j=1,2)$;
- Generate an initial mesh $\mathcal{M}_h$ over $D$;
- While $\mathcal{E}_{\text{FEM}} >$ TOL do
  - Choose a set of elements $\widehat{\mathcal{M}}_h \subset \mathcal{M}_h$ such that

$$\Big(\sum_{T\in\widehat{\mathcal{M}}_h} \eta_T^2\Big)^{1/2} > 0.7\Big(\sum_{T\in\mathcal{M}_h} \eta_T^2\Big)^{1/2},$$

    then refine the elements in $\widehat{\mathcal{M}}_h$, and denote the new mesh by $\mathcal{M}_h$
  - solve the discrete problem $(3.23)$ on $\mathcal{M}_h$
  - compute error estimators on $\mathcal{M}_h$

  end while.

## 6.2 Numerical example

A simple structure, namely a lamellar grating as shown in Fig. 2, is considered here. Assume that a plane wave $u_{\text{I}} = e^{\text{i}\alpha x_1 - \text{i}\beta x_3}$ is incident on the grating, which separates two homogeneous media whose dielectric coefficients are $\varepsilon_1$ and $\varepsilon_2$, respectively.

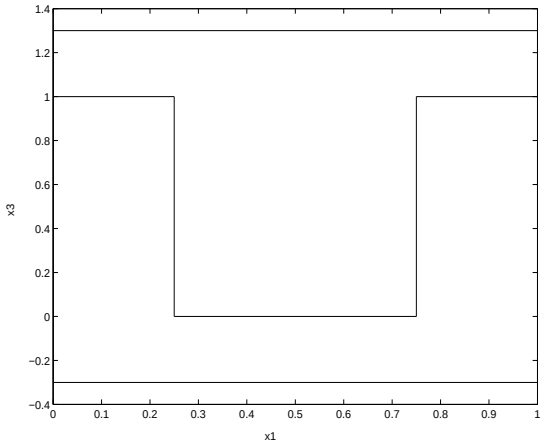


Figure 2: Grating structure.

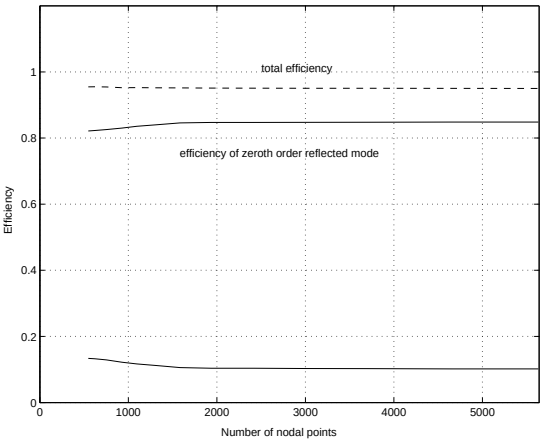


Figure 3: Grating efficiency.

In this experiment, the parameters are chosen as $\varepsilon_1 = 1$, $\varepsilon_2 = (0.22+6.71\text{i})^2$, $\theta = \pi/6$, $\omega = \pi$, $L = 1$ and the TM polarization is concerned. There are two reflected outgoing waves whose grating efficiency as well as the total grating efficiency are displayed in Fig. 3.

Fig. 4 shows the mesh after 100 adaptive iterations, which has 2822 elements and the a posteriori error estimate on the mesh is 0.0474383. It is obvious that the proposed

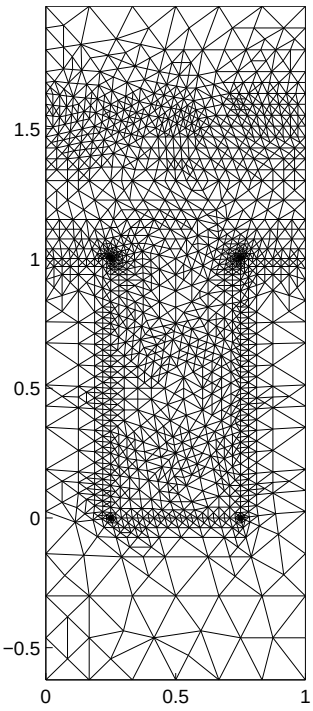


Figure 4: The adaptive mesh.

algorithm is able to capture the singularities of the problem. The meshes near the upper PML boundary are rather coarse even though we omit the exponential decay factor in the process of deducing the a posteriori error estimator. This phenomena illustrates that the property of exponential decay is equipped by the problem itself. Finally the amplitude of the associated solution is shown in Fig. 5.

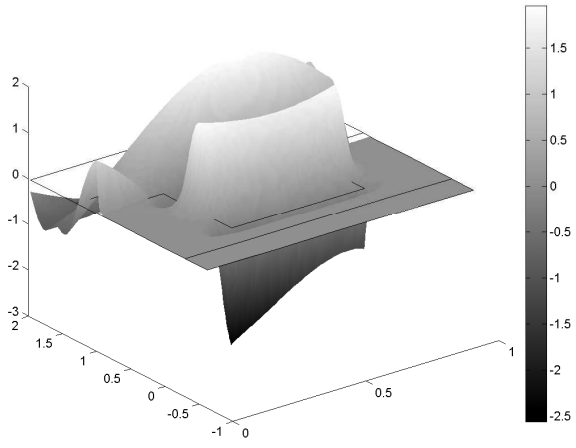


Figure 5: The surface of the amplitude of the associated solution.

## 7 Conclusion and future work

The adaptive finite element method with a PML for one-dimensional diffraction gratings both in TE and TM polarization is improved. A modified PML formulation is introduced, which renders simplified error analysis and easier numerical implementations. An improved PML error estimate on the situation $\mathrm{Im}\varepsilon_2 > 0$ is presented which results better error bound when $\mathrm{Im}\varepsilon_2$ is small and positive. A lower bound of the error between the PML solution and the finite element approximation is derived, which shows that the a posteriori error estimates we obtained in this paper are sharp.

Further numerical experiments based on our analysis are the focus of future research and the proposed algorithm will be extended to three-dimensional problems, i.e., 2D grating problems, for more realistic applications.

## Acknowledgments

The authors are grateful to the referees for the valuable suggestions on the improvement of the manuscript. The work is partially supported by the NTU start-up grant M58110011.

**References**

[1] A. Rathsfeld, G. Schmidt and B. H. Kleemann, On a fast integral equation method for diffraction gratings, Commun. Comput. Phys., 1 (2006), 984-1009.
[2] G. Bao, L. Cowsar and W. Masters, Mathematical Modeling in Optical Science, SIAM Frontier in Applied Mathematics, 2001.
[3] G. Bao, D. C. Dobson and J. A. Cox, Mathematical studies in rigorous grating theory, J. Opt. Soc. Amer. A, 12 (1995), 1029-1042.
[4] G. Bao, Z. Chen and H. Wu, Adaptive finite-element method for diffraction gratings, J. Opt. Soc. Am. A, 22 (2005), 1106-1114.
[5] I. Babuška and A. Aziz, Survey lectures on mathematical foundations of the finite element method, in: A. Aziz (Ed.), The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations, Academic Press, New York, 1973, pp. 5-359.
[6] I. Babuška and C. Rheinboldt, Error estimates for adaptive finite element computations, SIAM J. Numer. Anal., 15 (1978), 736-754.
[7] J.-P. Berenger, A perfectly matched layer for the absorption of electromagnetic waces, J. Comput. Phys., 114 (1994), 185-200.
[8] J. Chen, Adaptive finite element method for 3-D electromagnetic fields: Time-harmonic scattering problems and eddy current problems, Ph.D Thesis, the Graduate University of the Chinese Academy of Sciences.
[9] K. Mekchay and R. H. Nochetto, Convergence of adaptive finite element methods for general second order linear elliptic PDEs, SIAM J. Numer. Anal., 43 (2005), 1803-1827.
[10] L. R. Scott and S. Zhang, Finit element interpolation of nonsmooth functions satisfying boundary conditions, Math. Comp., 54 (1990), 483-493.
[11] R. Stevenson, Optimality of a standard adaptive finite element method, Found. Comput. Math., 7 (2007), 245-269.

[12] P. Morin, R. H. Nochetto and K. G. Siebert, Data oscillation and convergence of adaptive FEM, SIAM J. Numer. Anal., 38 (2000), 466-488.
[13] P. Morin, R. H. Nochetto and K. G. Siebert, Convergence of adaptive finite element methods, SIAM Rev., 44 (2002), 631-658.
[14] 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 (2006), 1010-1042.
[15] W. Dörfler, A convergent adaptive algorithm for Poisson's equation, SIAM J. Numer. Anal., 33 (1996), 1106-1124.
[16] Z. Chen and H. Wu, An adaptive finite element method with perfectly matched absorbing layers for the wave scattering by periodic structures, SIAM J. Numer. Anal., 41 (2003), 799-826.
[17] Z. Chen and X. Liu, An adaptive perfectly matched layer technique for time-harmonic scattering problems, SIAM J. Numer. Anal., 43 (2005), 645-671.