Hybrid absorbing boundary conditions of PML and CRBC

Seungil Kim[a]

[a]*Department of Mathematics and Research Institute for Basic Sciences, Kyung Hee University, Seoul 02447, Republic of Korea*

Abstract

In this paper we introduce a hybrid absorbing boundary condition (HABC) by combining perfectly matched layer(PML) and complete radiation boundary condition(CRBC) for solving a one-dimensional diffraction grating problem. The new boundary condition is devised in such a way that it can enjoy relative advantages from both methods. The well-posedness of the problem with HABC and the convergence of approximate solutions will be analyzed. Numerical examples to illustrate the efficiency of HABC are also presented.

Key words: Absorbing boundary condition, perfectly matched layer, complete radiation boundary condition, Helmholtz equation, diffraction grating

1. Introduction

In this paper, we will develop a new absorbing boundary condition for time-harmonic wave scattering problems by combining two well-known absorbing boundary conditions, perfectly matched layers(PML) and complete radiation boundary conditions(CRBC). In particular, we consider a model problem governed by the Helmholtz equation, of which solutions propagate in one direction, for example, model problems describing wave phenomena arising in guided structures or periodic diffraction gratings. Since solutions to these kinds of model problems can be represented as a superposition of finite number of propagating modes and infinite number of evanescent modes, promising absorbing boundary conditions are required to handle both types of modes efficiently.

The PML is an absorbing boundary layer that wave fields propagating into decay exponentially without spurious reflections. Even though the PML needs to be truncated in actual involvement in numerical computations, it produces only exponentially small reflection depending on how strong PML strength and how large PML width are used. Due to its easy implementation and strong effectiveness it has been actively engaged in numerical computations related with wave propagation ever since it was proposed by Bérenger in [4, 5] to solve

Email address: sikim@khu.ac.kr (Seungil Kim)

electromagnetic wave propagation problems. On the other hand, CRBC was introduced recently for high-order absorbing boundary conditions in [8, 10]. It can be interpreted as a rational approximation to the square root function involved in the radiation condition and parameters of CRBC can be chosen in an optimal way so that it reduces reflection errors exponentially with respect to the number of parameters and uniformly in the spectral range of interest.

Although both PML and CRBC have been utilized successfully as efficient absorbing boundary conditions, one has the superiority over the other in dealing with some types of modes. In case that solutions contain grazing modes such as propagating modes of large wavelength or slowly decaying evanescent modes generated close to artificial boundaries, it is observed that PML may require inefficiently large computational resources to obtain approximations of reasonable accuracy, see e.g., [9]. In contrast, CRBC can not only act as an exact radiation condition for such modes by taking appropriate parameters but also reduce the reflection errors of modes in a certain spectral range uniformly and optimally. However since the Dirichlet-to-Neumann(DtN) operator for CRBC does not converge to the DtN operator for the exact radiation condition in a Sobolev norm, its performance may be deteriorated by evanescent modes of high axial frequency unless a gap δ between wave sources and the artificial boundary is sufficiently large for the evanescent modes to be ignorable on the artificial boundary. PML can get rid of this concern by accelerating the decay rate of the evanescent modes via PML coordinate stretchings. This observation motivates us to design a hybrid absorbing boundary condition(HABC) taking advantage of two methods, i.e, the CRBC component of the absorbing boundary condition takes care of a small number of low frequency modes and the PML counterpart takes the responsibility of diminishing the reflection errors of the rest of infinite number of modes.

Once the hybrid absorbing boundary condition is defined, we will discuss the well-posedness of the model problem supplemented with HABC and the convergence of approximate solutions. It is also worth noting that a recent effort to improve the performance of PML has been made by cooperating with a truncated DtN boundary condition in [14]. However, typically the truncated DtN boundary condition destroys the sparsity of system matrices resulting from discretization techniques as opposed to CRBC preserving the sparsity.

The rest of the paper is organized as follows. In Section 2 we introduce a one-dimensional diffraction grating problem as the model problem. In section 3, the hybrid absorbing boundary condition as a combination of CRBC and PML is defined to solve the diffraction grating problem in a finite computational domain. Section 4 is devoted to establishing the well-posedness of the Helmholtz equation with HABC. A discussion on how CRBC parameters are chosen to minimize the reflection error for given number of grid points along the direction of wave propagation is given in Section 5. Finally, numerical experiments illustrating the performance of HABC will be presented in Section 6.

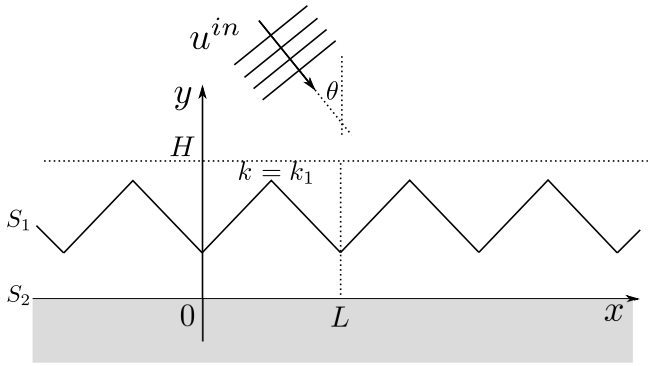


Figure 1: Geometric configuration of the one-dimensional diffraction grating

2. Model problem: one-dimensional diffraction grating

As a model problem, we consider one-dimensional periodic diffraction grating problem in $\mathbb{R}^2$. A diffraction grating is an optical device that decomposes polychromatic light into light components with different wavelength. For the transverse magnetic polarization of electromagnetic fields in $\mathbb{R}^3$, the longitudinal component (z-component) u of electric fields solves the Helmholtz equation

$$\Delta u + k^2 u = 0, \tag{2.1}$$

where $k = \omega c^{-1}\sqrt{\epsilon\mu}$ with angular frequency ω, speed of light in air c, magnetic permeability μ, and electric permittivity $\epsilon(x, y)$ depending on the spatial variable $(x, y) \in \mathbb{R}^2$. Assume that a periodic grating with period $L > 0$ along the x-axis is bounded above by a piecewise smooth periodic curve S_1, whose y component is in $0 < y < H$ and below by the curve S_2 of $y = 0$. The periodic grating structure is characterized by an inhomogeneous wavenumber $k \in L^\infty$. The medium above S_1 is assumed to be homogeneous with constant wavenumber $k = k_1$. For instance, the medium above S_1 is filled up by air and so k_1 is real and positive. The one below of S_2 is, in general, occupied by absorbing materials, which corresponds to complex wavenumber with positive real and imaginary parts, however in order to focus on the performance of the absorbing boundary condition we impose the perfectly conducting boundary condition $u = 0$ on S_2.

Now, the incident wave is given by a plane wave $u^{in}(x, y) = e^{i(\alpha x - \beta y)}$ propagating from the top with the incident angle $\theta \in (-\pi/2, \pi/2)$, $\alpha = k_1 \sin(\theta)$ and $\beta = k_1 \cos(\theta)$, see Figure 1. The periodic diffraction grating can produce an α-quasiperiodic scattered field u^{sc} satisfying

$$u^{sc}(x + L, y) = e^{i\alpha L} u^{sc}(x, y)$$

and a radiation condition for $y \to \infty$ to be defined below. Due to the α-quasiperiodicity, the problem for α-quasiperiodic solutions can be restricted to

the semi-infinite strip region

$$\Omega_\infty = \{(x,y) \in \mathbb{R}^2 \ : \ 0 < x < L, \quad 0 < y < \infty\}.$$

For the radiation condition, we will write α-quasiperiodic solutions to the Helmholtz equation in a series form with the help of a complete set of orthonormal eigenfunctions in $L^2(0,L)$ of the α-quasiperiodic eigenvalue problem in $L^2_{loc}(\mathbb{R})$

$$\begin{aligned} -\phi''(x) &= \lambda^2 \phi(x), \\ \phi(x+L) &= e^{i\alpha L}\phi(x). \end{aligned}$$

Here eigenfunctions are given by $\phi_n(x) = \frac{1}{\sqrt{L}} e^{i\lambda_n x}$ for the eigenvalues λ_n^2 with $\lambda_n = \alpha + \frac{2n\pi}{L}$, $n \in \mathbb{Z}$. Under the time-harmonic convention $e^{-i\omega t}$ of the time variable t, the scattered field u^{sc} propagating toward the positive infinity along the y-axis is the one represented by

$$u^{sc}(x,y) = \sum_{n\in\mathbb{Z}} A_n e^{i\mu_n(y-H)}\phi_n(x) \quad \text{ for } y > H \tag{2.2}$$

for some constants A_n, where $\mu_n = \sqrt{k_1^2 - \lambda_n^2}$ with the negative real axis branch cut so that μ_n is real and positive for $k_1^2 - \lambda_n^2 > 0$ and purely imaginary with positive imaginary part $\tilde{\mu}_n$ (that is, $\mu_n = i\tilde{\mu}_n$) for $k_1^2 - \lambda_n^2 < 0$. In some cases, μ_n is possibly zero, for which the mode is called the cutoff mode. For a complete analysis, we assign the index N to the cutoff mode, i.e., $\mu_N = 0$.

Thus, denoting the domain of interest by $\Omega = (0,L)\times(0,H)$ and its boundary for the radiation condition by $\Gamma = \{(x,y) \ : \ 0 < x < L, \quad y = H\}$, the radiation condition can be interpreted on the boundary Γ as

$$\partial_y u^{sc} = T(u^{sc}) \tag{2.3}$$

in terms of the DtN operator $T : H_\alpha^{1/2}(\Gamma) \to H_\alpha^{-1/2}(\Gamma)$ defined by

$$T(v) = \sum_{n\in\mathbb{Z}} i\mu_n v_n \phi_n \tag{2.4}$$

for $v = \sum_{n\in\mathbb{Z}} v_n\phi_n$, where $H_\alpha^s(\Gamma)$ for $s \in \mathbb{R}$ is a Sobolev space consisting of functions v satisfying

$$\|v\|^2_{H_\alpha^s(\Gamma)} = \sum_{n\in\mathbb{Z}} (1+\lambda_n^2)^s |v_n| < \infty.$$

Here $H_\alpha^{1/2}(\Gamma)$ can be seen as a trace space of $H_\alpha^1(\Omega)$, the subspace of functions with α-quasiperiodicity on the lateral boundaries in $H^1(\Omega)$, and it holds that

$$\|v|_\Gamma\|_{H_\alpha^{1/2}(\Gamma)} \le C\|v\|_{H^1(\Omega)}$$

for $v \in H_\alpha^1(\Omega)$ (see e.g., [11] for a similar analysis).

Now, the incident wave u^{in} can be imposed on Γ as a source term for the boundary condition of the total field $u^{to} = u^{in} + u^{sc}$, i.e., noting that the incident wave $u^{in}(x,y) = \sqrt{L}e^{-i\beta y}\phi_0(x)$ is composed of only one mode of $n = 0$ and satisfies

$$\partial_y u^{in} = T(u^{in}) - 2i\beta u^{in} \text{ on } \Gamma, \tag{2.5}$$

(2.3) and (2.5) give the boundary condition for the total field u^{to}

$$\partial_y u^{to} = T(u^{to}) - 2i\beta u^{in} \text{ on } \Gamma.$$

Therefore, the diffraction grating problem to find the total field u^{to} for given u^{in} can be reduced to the problem posed on the bounded domain Ω, that can be written as finding $u^{to} \in H^1_\alpha(\Omega)$ satisfying

$$\begin{aligned} \Delta u^{to} + k^2 u^{to} &= 0 \quad \text{in } \Omega, \\ u^{to} &= 0 \quad \text{on } \Gamma_2, \qquad u^{to}(L,\cdot) = e^{i\alpha L} u^{to}(0,\cdot), \\ \partial_y u^{to} &= T(u^{to}) - 2i\beta u^{in} \text{ on } \Gamma. \end{aligned} \tag{2.6}$$

Here $\Gamma_2 = (0,L) \times \{0\}$. Denoting the L^2-inner product in a domain $\mathcal{D}$ by $(\cdot,\cdot)_{\mathcal{D}}$ and the duality pairing between $H^{-1/2}_\alpha(\Gamma)$ and $H^{1/2}_\alpha(\Gamma)$ with pivot space $H^0_\alpha(\Gamma) = L^2(\Gamma)$ by $\langle\cdot,\cdot\rangle_\Gamma$, and letting $H^1_{\alpha,0}(\Omega)$ be the subspace of functions vanishing on Γ_2 in $H^1_\alpha(\Omega)$, we reformulate (2.6) to a variational problem to find $u^{to} \in H^1_{\alpha,0}(\Omega)$ such that

$$a_\Omega(u^{to}, v) - \langle T(u^{to}), v\rangle_\Gamma = \langle g, v\rangle_\Gamma \text{ for } v \in H^1_{\alpha,0}(\Omega), \tag{2.7}$$

where $g = -2i\beta u^{in}$ on Γ and

$$a_\Omega(u,v) = (\nabla u, \nabla v)_\Omega - k^2 (u,v)_\Omega.$$

It is shown by the compact argument in [3] (see also [6]) that except for discrete sets of eigenvalues ω the problem (2.7) is well-posed and there exists a positive constant C such that

$$\|w\|_{H^1(\Omega)} \le C \sup_{0 \neq \psi \in H^1_{\alpha,0}(\Omega)} \frac{|a_\Omega(w,\psi) - \langle T(w), \psi\rangle_\Gamma|}{\|\psi\|_{H^1(\Omega)}}. \tag{2.8}$$

for all $w \in H^1_{\alpha,0}(\Omega)$. Thus, from now on we assume that ω is not an eigenvalue of the diffraction grating problem. For such ω, the main goal is to develop an absorbing boundary condition that can replace T and have the relative advantage from PML and CRBC.

3. Hybrid absorbing boundary condition

In this section we introduce the hybrid absorbing boundary condition based on PML and CRBC. We first review CRBC from [8, 10].

3.1. Complete radiation boundary condition

The CRBC is a high order absorbing boundary condition, which can be interpreted as a rational approximation to the square root function of the symbol $\sqrt{k^2-\partial_x^2}$ of the radiation condition, see [10]. Recall that CRBC of order (n_p, n_e) for non-negative integers n_p and n_e is defined in terms of auxiliary variables u^j for $j = 0, 1, \ldots, P$ with $P = n_p + n_e$ satisfying the Helmholtz equation with the boundary condition of α-quasiperiodicity on the lateral boundaries and the recurrence relations

$$\begin{aligned} u^0 &= u, \\ (\partial_y + a_j)u^j &= (-\partial_y + \tilde{a}_j)u^{j+1} \text{ for } j = 0, 1, \ldots, P-1 \end{aligned} \tag{3.9}$$

with the terminal condition

$$\partial_y u^P = 0 \text{ on } \Gamma, \tag{3.10}$$

where the parameters a_j and $\tilde{a}_j$ satisfy

$$\begin{aligned} a_j = -i\sigma_j, \quad \tilde{a}_j = -i\tilde{\sigma}_j \text{ for } j = 0, \ldots, n_p - 1, \\ a_{n_p+j}, \quad \tilde{a}_{n_p+j} > 0 \text{ for } j = 0, \ldots, n_e - 1 \end{aligned} \tag{3.11}$$

with $0 < \sigma_j, \tilde{\sigma}_j \le k_1$ such that $|a_j| \le |\tilde{a}_j|$.

Here $2n_p$ purely imaginary parameters are used for damping the reflection error from propagating modes and the $2n_e$ real parameters are responsible for reducing the reflection error from slowly decaying evanescent modes. For example, since u^j solves the Helmholtz equation with the α-quasiperiodic boundary condition on the lateral boundaries, u^j can be written as the series form near Γ

$$u^j(x,y) = (A_N^j + B_N^j(y-H))\phi_N(x) + \sum_{n\in\mathbb{Z}, n\neq N} (A_n^j e^{i\mu_n(y-H)} + B_n^j e^{-i\mu_n(y-H)})\phi_n(x) \tag{3.12}$$

for some constants A_n^j and B_n^j. If one parameter, for instance $\tilde{a}_\ell$, is chosen such that $\tilde{a}_\ell = -i\mu_n$ for some $n \neq N$, then one can show that $B_n^\ell = 0$ from (3.9) with $j = \ell$, which results in $B_n^j = 0$ for $j = 0, 1, \ldots, \ell$ from (3.9) with $j < \ell$ and so the n-th mode of $u = u^0$ satisfies the exact radiation condition. In actual practice, the parameters are chosen so that the reflection errors of modes in a certain spectral range are uniformly minimized.

Minimizing the reflection errors in a spectral range is related with the minimization of the reflection coefficients defined as the ratio of the amplitude of the reflected component to the that of the outgoing component,

$$\mathcal{Z}_n := \begin{cases} \dfrac{B_n^0}{A_n^0} = \displaystyle\prod_{j=0}^{P} \frac{(a_j + i\mu_n)(\tilde{a}_j + i\mu_n)}{(a_j - i\mu_n)(\tilde{a}_j - i\mu_n)} & \text{for } n \neq N \\ 0 & \text{for } n = N. \end{cases} \tag{3.13}$$

This formula can be easily obtained by applying (3.12) to the recurrence relation (3.9). Indeed, for $n \neq N$ it can be seen that

$$A_n^j = Q_n^{0,j-1} A_n^0, \qquad B_n^j = R_n^{j,P-1} B_n^P, \tag{3.14}$$

where

$$Q_n^{\ell,m} = \begin{cases} \displaystyle\prod_{j=\ell}^{m} \frac{a_j + i\mu_n}{\tilde{a}_j - i\mu_n} & \text{for } \ell \le m, \\ 1 & \text{for } \ell > m, \end{cases} \qquad R_n^{\ell,m} = \begin{cases} \displaystyle\prod_{j=\ell}^{m} \frac{\tilde{a}_j + i\mu_n}{a_j - i\mu_n} & \text{for } \ell \le m, \\ 1 & \text{for } \ell > m. \end{cases} \tag{3.15}$$

Since the terminal condition (3.10) for the n-th mode leads to $A_n^P = B_n^P$, (3.13) is obtained. In case of $n = N$, the computation similar to the above shows that

$$a_j B_N^j = \tilde{a}_j B_N^{j+1} \text{ and } (1-a_j H)B_N^j + a_j A_N^j = -(1+\tilde{a}_j H)B_N^{j+1} + \tilde{a}_j A_N^{j+1} \tag{3.16}$$

for $j = 0, \ldots, P-1$. The terminal condition (3.10) for the N-th mode shows that $B_N^P = 0$, which implies that $B_N^0 = 0$, that is, CRBC serves as the exact radiation condition for the cutoff mode. The optimal selection of the parameters that can minimize the reflection coefficients (3.13) will be discussed in Subsection 3.4.

The recurrence relations (3.9) can be rephrased to an equivalent form involving only the tangential derivatives, which is more suitable for finite element implementation,

$$-\partial_y u \boldsymbol{e}^0 = -\mathbb{L}\partial_x^2 \mathcal{U} + (-k_1^2 \mathbb{L} + \mathbb{M})\mathcal{U} - \partial_y u^P \boldsymbol{e}^P \text{ on } \Gamma, \tag{3.17}$$

where $\mathcal{U} = (u^0, u^1, \ldots, u^P)^\top$ and $\boldsymbol{e}^j$ is the j-th standard basis vector of $\mathbb{C}^{P+1}$ for $j = 0, 1, \ldots, P$. In addition, $\mathbb{L}$ and $\mathbb{M}$ are $(P+1) \times (P+1)$ symmetric tridiagonal matrices such that for $0 \le j \le P$

$$\mathbb{L}_{j,j-1} = \frac{1}{a_{j-1} + \tilde{a}_{j-1}}, \quad \mathbb{L}_{j,j} = \frac{1}{a_{j-1} + \tilde{a}_{j-1}} + \frac{1}{a_j + \tilde{a}_j}, \quad \mathbb{L}_{j,j+1} = \frac{1}{a_j + \tilde{a}_j}$$

$$\mathbb{M}_{j,j-1} = \frac{-a_{j-1}^2}{a_{j-1} + \tilde{a}_{j-1}} \quad \mathbb{M}_{j,j} = \frac{a_{j-1}\tilde{a}_{j-1}}{a_{j-1} + \tilde{a}_{j-1}} + \frac{a_j \tilde{a}_j}{a_j + \tilde{a}_j} \quad \mathbb{M}_{j,j+1} = \frac{-\tilde{a}_j^2}{a_j + \tilde{a}_j}.$$

Here the terms with indices outside the index range of $a_j, \tilde{a}_j$ $(0 \le j \le P-1)$ are ignored by convention. This can be achieved by multiplying (3.9) by the differential operator ∂_y and by eliminating ∂_y^2 using the Helmholtz equation of u^j. See [8] in more detail. Of course, the Neumann terminal condition (3.10) removes the last term in (3.17) however we keep this term in (3.17) for the derivation of the hybrid absorbing boundary condition replacing the Neumann terminal condition with PML.

3.2. Hybrid absorbing boundary condition

The new absorbing boundary condition proposed in this paper is CRBC terminated with PML, that is, PML is applied to the last auxiliary variable u^P. More precisely, let $\Omega_{\text{PML}} = (0, L) \times (H, H + M)$ be the damping zone with PML width $M > 0$ and $\Gamma_{\text{PML}} = (0, L) \times \{H + M\}$ being the boundary on which PML is truncated. For a linear PML coordinate stretching function $\tilde{y}(y) = \sigma_0(1 + i)(y - H) + H$ for $y \ge H$ with positive constant σ_0 we introduce

the PML solution $\tilde{u}^P$, associated with u^P, solving

$$\sigma\partial_x^2\tilde{u}^P + \partial_y\frac{1}{\sigma}\partial_y\tilde{u}^P + \sigma k_1^2\tilde{u}^P = 0 \text{ in } \Omega_{\text{PML}}, \tag{3.18}$$

$$\tilde{u}^P = u^P \text{ and } \frac{1}{\sigma}\partial_y\tilde{u}^P = \partial_y u^P \text{ on } \Gamma, \tag{3.19}$$

$$\partial_y\tilde{u}^P = 0 \text{ on } \Gamma_{\text{PML}} \tag{3.20}$$

with the α-quasiperiodic condition on the lateral boundaries. Here $\sigma = \sigma_0(1+i) = \partial\tilde{y}/\partial y$. A homogeneous Dirichlet condition can be an alternative to (3.20), however, the Neumann condition (3.20) allows HABC to give a better performance for cutoff modes or grazing modes.

We now analyze HABC in terms of a DtN operator. Since the auxiliary variables u^j of HABC satisfy the same recurrence relations as those of CRBC, the coefficients A_n^j and B_n^j of u^j still satisfy (3.14) and (3.16). Due to the continuity conditions (3.19) of u^P, $\tilde{u}^P$ and their normal fluxes on Γ the solution $\tilde{u}^P$ to the PML equation (3.18) can be written as

$$\tilde{u}^P(x,y) = (A_N^P + \sigma B_N^P(y-H))\phi_N(x) + \sum_{n\in\mathbb{Z},n\neq N}\left[A_n^P e^{i\mu_n\sigma(y-H)} + B_n^P e^{-i\mu_n\sigma(y-H)}\right]\phi_n(x) \tag{3.21}$$

for $H < y < H+M$ with the same coefficients A_n^P and B_n^P of u^P. Finally, it can be shown that the boundary condition (3.20) for (3.21) gives

$$B_N^P = 0 \text{ and } B_n^P = e^{2i\mu_n\sigma M}A_n^P, \tag{3.22}$$

and hence combining (3.14) and (3.22) yields the reflection coefficient of HABC for the n-th mode

$$\frac{B_n^0}{A_n^0} := \mathcal{R}_n = \mathcal{Z}_n e^{2i\mu_n\sigma M}. \tag{3.23}$$

Here we note that $\mathcal{Z}_n$ is the reflection coefficient resulting from CRBC and the exponential term $e^{2i\mu_n\sigma M}$ is the typical one for PML depending on PML strength σ and width M. It is also noticed that HABC is the exact radiation condition for the cutoff mode and the coefficients for the cutoff modes can be given as

$$A_N^j = \left(\prod_{j=0}^{j-1}\frac{a_j}{\tilde{a}_j}\right)A_N^0 = Q_N^{0,j-1}A_N^0 \text{ for } j = 1,\ldots,P. \tag{3.24}$$

By using (3.14), (3.22) and (3.24) in (3.12), the solution and the auxiliary variables u^j for $j = 0,\ldots,P$ can be written as

$$\begin{aligned} u^j(x,y) &= Q_N^{0,j-1}A_N^0\phi_N(x) \\ &+ \sum_{n\in\mathbb{Z},n\neq N}\left[Q_n^{0,j-1}e^{i\mu_n(y-H)} + R_n^{j,P-1}Q_n^{0,P-1}e^{2i\mu_n\sigma M}e^{-i\mu_n(y-H)}\right]A_n^0\phi_n(x). \end{aligned} \tag{3.25}$$

To define the DtN operator associated with HABC, for given a Dirichlet data $u = u^0 = \sum_{n\in\mathbb{Z}} u_n^0 \phi_n$ on Γ, we seek for the Neumann value of u (3.25) with $j = 0$ on Γ, which reads

$$\partial_y u = \sum_{n\in\mathbb{Z}} i\mu_n (A_n^0 - B_n^0) = \sum_{n\in\mathbb{Z}} i\mu_n \frac{1-\mathcal{R}_n}{1+\mathcal{R}_n} u_n^0 \phi_n.$$

Finally, the DtN operator for HABC is defined by the operator $T_{\mathrm{HABC}} : H_\alpha^{1/2}(\Gamma) \to H_\alpha^{-1/2}(\Gamma)$ given by

$$T_{\mathrm{HABC}}(v) = \sum_{n\in\mathbb{Z}} i\mu_n \frac{1-\mathcal{R}_n}{1+\mathcal{R}_n} v_n \phi_n \tag{3.26}$$

for $v = \sum_{n\in\mathbb{Z}} v_n \phi_n \in H_\alpha^{1/2}(\Gamma)$.

In light of (3.23), we can minimize the reflection coefficients uniformly in a certain spectral range by minimizing the maximal value of $|\mathcal{Z}_n|$ as we will see in the next subsection. Also we can expect the exact radiation condition for the modes of small $|\mu_n|$ troublesome to PML by choosing a parameter to be $-i\mu_n$ of the corresponding small axial frequency. Furthermore, even if $|\mathcal{Z}_n| \to 1$ as $|n| \to \infty$ (which is a drawback of CRBC for evanescent modes of high axial frequency since CRBC can not diminish the reflection coefficients of such modes), HABC still can play a role of an effective absorbing boundary condition for such large n owing to PML reducing the reflection coefficients exponentially at a rate of $e^{-2\tilde{\mu}_n \sigma_0 M}$. Therefore, we can design HABC such that CRBC works for modes of relatively small axial frequencies (all propagating modes and slowly decaying evanescent modes) and PML takes care of the rest of modes (all fast decaying evanescent modes). We will discuss how to choose the optimal CRBC parameters in more depth to minimize the reflection coefficients for given PML strength σ and width M in Subsection 3.4.

3.3. HABC *in a variational formulation*

As discussed in Subsection 3.1, the recurrence relations (3.9) can be transformed to a matrix form (3.17) involving only tangential derivatives. By the continuity conditions (3.19) on Γ, HABC is defined in terms of $\mathcal{U} = (u^0, \ldots, u^P)^\top$ and $\tilde{u}^P$ such that $u = u^0$ and $u^P = \tilde{u}^P$ on Γ, and

$$-\partial_y u \boldsymbol{e}^0 = -\mathbb{L}\partial_x^2 \mathcal{U} + (-k_1^2 \mathbb{L} + \mathbb{M})\mathcal{U} - \frac{1}{\sigma}\partial_y \tilde{u}^P \boldsymbol{e}^P \text{ on } \Gamma, \tag{3.27}$$

$$\sigma \partial_x^2 \tilde{u}^P + \partial_y \frac{1}{\sigma} \partial_y \tilde{u}^P + \sigma k_1^2 \tilde{u}^P = 0 \text{ in } \Omega_{\mathrm{PML}}, \tag{3.28}$$

$$\partial_y \tilde{u}^P = 0 \text{ on } \Gamma_{\mathrm{PML}} \tag{3.29}$$

with the α-quasiperiodic condition on the lateral boundaries.

The next lemma shows that both conditions (3.9) and (3.17) are in fact equivalent and they produce the same DtN operator.

Lemma 3.1. *Assume that u solves the Helmholtz equation with wavenumber k_1 near Γ and satisfies the* HABC (3.27)-(3.29) *with the α-quasiperiodic condition on the lateral boundaries. Then u has the series representation*

$$u(x,y) = A_N\phi_N(x) + \sum_{n\in\mathbb{Z},n\neq N} \big[e^{i\mu_n(y-H)} + \mathcal{R}_n e^{-i\mu_n(y-H)}\big]A_n\phi_n(x). \quad (3.30)$$

Similarly, the auxiliary functions u^j for $j = 1,\dots,P$ and $\tilde{u}^P$ can be written as

$$u^j(x) = Q_N^{0,j-1}A_N\phi_N(x) + \sum_{n\in\mathbb{Z},n\neq N} \big[Q_n^{0,j-1} + R_n^{j,P-1}Q_n^{0,P-1}e^{2i\mu_n\sigma M}\big]A_n\phi_n(x) \quad (3.31)$$

on Γ and

$$\tilde{u}^P(x,y) = Q_N^{0,P-1}A_N\phi_N(x) + \sum_{n\in\mathbb{Z},n\neq N} \big[e^{i\mu_n\sigma(y-H)} + e^{2i\mu_n\sigma M}e^{-i\mu_n\sigma(y-H)}\big]Q_n^{0,P-1}A_n\phi_n(x) \quad (3.32)$$

for $H < y < H + M$.

Proof. We consider the series representations of u and $\tilde{u}^P$,

$$u(x,y) = (A_N^0 + B_N^0(y-H))\phi_N(x) + \sum_{n\in\mathbb{Z},n\neq N} \big[A_n^0 e^{i\mu_n(y-H)} + R_n^{0,P-1}B_n^P e^{-i\mu_n(y-H)}\big]\phi_n(x),$$

$$\tilde{u}^P(x,y) = (\widetilde{A}_N^P + \widetilde{B}_N^P\sigma(y-H))\phi_N(x) + \sum_{n\in\mathbb{Z},n\neq N} \big[\widetilde{A}_n^P e^{i\mu_n\sigma(y-H)} + \widetilde{B}_n^P e^{-i\mu_n\sigma(y-H)}\big]\phi_n(x)$$

for some constants $A_N^0, B_N^0, A_n^0, B_n^P$, $\widetilde{A}_n^P$ and $\widetilde{B}_n^P$. Here we set $A_n = A_n^0$ and $B_n = B_n^0$. For $n \neq N$, the equation (3.27) is reduced to

$$-i\mu_n(A_n^0 - R_n^{0,P-1}B_n^P)\boldsymbol{e}^0 = (-\mu_n^2\mathbb{L} + \mathbb{M})\mathcal{U}_n - \frac{1}{\sigma}\partial_y\tilde{u}_n^P\boldsymbol{e}^P, \quad (3.33)$$

where $\mathcal{U}_n = (u_n^0,\dots,u_n^P)^\top$ and $\tilde{u}_n^P$ are the n-th components of $\mathcal{U}$ and $\tilde{u}^P$, respectively. It is shown in [10] that $-\mu_n^2\mathbb{L}+\mathbb{M}$ is invertible, which implies that $\mathcal{U}_n$ is uniquely determined in terms of $i\mu_n(A_n^0 - R_n^{0,P-1}B_n^P)$ and $\sigma^{-1}\partial_y\tilde{u}_n^P$. Hence it suffices to show that there exist the auxiliary variables u_n^j of the form

$$u_n^j = Q_n^{0,j-1}A_n^0 + R_n^{j,P-1}B_n^P$$

satisfying (3.33). It is easy to show that the homogeneous equations of (3.33) corresponding to the j-th row for $j = 1,\dots,P-1$ hold true. Since $\tilde{u}_n^P$ can be written as

$$\tilde{u}_n^P(y) = \widetilde{A}_n^P e^{i\mu_n\sigma(y-H)} + \widetilde{B}_n^P e^{-i\mu_n\sigma(y-H)}$$

with $\widetilde{B}_n^P = e^{2i\mu_n\sigma M}\widetilde{A}_n^P$, we can find the relations between coefficients A_n^0, B_n^P, $\widetilde{A}_n^P$ and $\widetilde{B}_n^P$ from the P-th component of (3.33) and the condition $u^P = \tilde{u}^P$ on Γ,

$$i\mu_n(Q_n^{0,P-1}A_n^0 - B_n^P) = i\mu_n(\widetilde{A}_n^P - \widetilde{B}_n^P),$$
$$Q_n^{0,P-1}A_n^0 + B_n^P = \widetilde{A}_n^P + \widetilde{B}_n^P.$$

It then follows that that $\widetilde{A}_n^P = Q_n^{0,P-1} A_n^0$ and $\widetilde{B}_n^P = B_n^P$, which gives

$$B_n^P = e^{2i\mu_n \sigma M} Q_n^{0,P-1} A_n^0$$

and hence we have

$$u_n^j = \left[Q_n^{0,j-1} + R_n^{j,P-1} Q_n^{0,P-1} e^{2i\mu_n \sigma M}\right] A_n^0 \text{ for } j = 0, \ldots, P$$

and

$$\tilde{u}_n^P(y) = \left[e^{i\mu_n \sigma (y-H)} + e^{2i\mu_n \sigma M} e^{-i\mu_n \sigma (y-H)}\right] Q_n^{0,P-1} A_n^0.$$

Now, we are left with showing the 0-th equation of (3.33) holds with $\mathcal{U}$. Noting that the right hand side of the 0-th equation of (3.33) is written as

$$\frac{-\mu_n^2 + a_0 \tilde{a}_0}{a_0 + \tilde{a}_0} (A_n^0 + R_n^{0,P-1} B_n^P) + \frac{-\mu_n^2 - \tilde{a}_0^2}{a_0 + \tilde{a}_0} (Q_n^{0,0} A_n^0 + R_n^{1,P-1} B_n^P),$$

a straightforward calculation reveals that it is indeed equal to $-i\mu_n (A_n^0 - R_n^{0,P-1} B_n^P)$.

For the cutoff mode, $n = N$, since $\partial_y \tilde{u}_N^P = 0$ on Γ_{PML}, we have $\widetilde{B}_N^P = 0$. Also, the boundary conditions $\partial_y u_N = B_N^0$ and $\partial_y \tilde{u}_N^P = 0$ on Γ show that $\mathcal{U}_N$ solves the problem $\mathbb{M}\mathcal{U}_N = -B_N^0 \boldsymbol{e}^0$. Noting that the 0-th and P-th components of $\mathcal{U}_N$ are $u_N^0 = A_N^0$ and $u_N^P = A_N^P = \widetilde{A}_N^P$, respectively, we apply Gaussian eliminations from the P-th row to the 0-th row and see that $\mathbb{M}\mathcal{U}_N = -B_N^0 \boldsymbol{e}^0$ has a solution only if $B_N^0 = 0$ and u_N^j is uniquely determined as $u_N^j = Q_N^{0,j-1} A_N$, which completes the proof. □

Now, we define the Sobolev spaces $H^1_\alpha(\Omega_{\mathrm{PML}})$ by the subspace of $H^1(\Omega_{\mathrm{PML}})$ with α-quasiperiodicity on the lateral boundaries. Denoting $\boldsymbol{H}^1_\alpha(\Gamma) = (H^1_\alpha(\Gamma))^{P+1}$, we let $\boldsymbol{X}$ be the subspace of functions $(u, \mathcal{U}, \tilde{u}^P)$ in $H^1_{\alpha,0}(\Omega) \times \boldsymbol{H}^1_\alpha(\Gamma) \times H^1_\alpha(\Omega_{\mathrm{PML}})$ such that $u|_\Gamma = u^0$ and $u^P = \tilde{u}^P|_\Gamma$, which is equipped with the weighted norm depending on a_j, $\tilde{a}_j$ and σ,

$$\|(u, \mathcal{U}, \tilde{u}^P)\|_{\boldsymbol{X}}^2 = \|u\|_{H^1(\Omega)}^2 + \|\mathcal{U}\|_{\boldsymbol{H}^1_\alpha(\Gamma)}^2 + \|\tilde{u}^P\|_{H^1_\sigma(\Omega_{\mathrm{PML}})}^2,$$

where

$$\|\mathcal{U}\|_{\boldsymbol{H}^1_\alpha(\Gamma)}^2 = \sum_{j=0}^{P-1} \frac{1}{|a_j + \tilde{a}_j|} \|u^j + u^{j+1}\|_{H^1_\alpha(\Gamma)}^2,$$

$$\|\tilde{u}^P\|_{H^1_\sigma(\Omega_{\mathrm{PML}})}^2 = |\sigma| \|\partial_x \tilde{u}^P\|_{L^2(\Omega_{\mathrm{PML}})}^2 + \frac{1}{|\sigma|} \|\partial_y \tilde{u}^P\|_{L^2(\Omega_{\mathrm{PML}})}^2 + |\sigma| \|\tilde{u}^P\|_{L^2(\Omega_{\mathrm{PML}})}^2.$$

If we denote

$$\|\Xi\|_{\mathbb{L}}^2 = \sum_{j=0}^{P-1} \frac{1}{|a_j + \tilde{a}_j|} |\xi^j + \xi^{j+1}|^2$$

for $\Xi = (\xi^0, \ldots, \xi^P)^\top \in \mathbb{C}^{P+1}$, then $\|\mathcal{U}\|_{\boldsymbol{H}^1_\alpha(\Gamma)}^2$ can be written as

$$\|\mathcal{U}\|_{\boldsymbol{H}^1_\alpha(\Gamma)}^2 = \sum_{n \in \mathbb{Z}} (1 + \lambda_n^2) \|\mathcal{U}_n\|_{\mathbb{L}}^2.$$

Then the problem (2.7) with the radiation condition replaced by HABC (3.27)-(3.29) is written in a weak sense as finding $(u, \mathcal{U}, \tilde{u}^P) \in \boldsymbol{X}$ satisfying

$$a_\Omega(u, v) + b_\Gamma(\mathcal{U}, \mathcal{V}) + a_{\Omega_{\rm PML}}(\tilde{u}^P, \tilde{v}^P) = \langle g, v\rangle_\Gamma \text{ for } (v, \mathcal{V}, \tilde{v}^P) \in \boldsymbol{X}, \qquad (3.34)$$

where

$$b_\Gamma(\mathcal{U}, \mathcal{V}) = (\mathbb{L}\partial_x\mathcal{U}, \partial_x\mathcal{V})_\Gamma + ((-k_1^2\mathbb{L} + \mathbb{M})\mathcal{U}, \mathcal{V})_\Gamma,$$
$$a_{\Omega_{\rm PML}}(\tilde{u}^P, \tilde{v}^P) = (\sigma\partial_x\tilde{u}^P, \partial_x\tilde{v}^P)_{\Omega_{\rm PML}} + (\frac{1}{\sigma}\partial_y\tilde{u}^P, \partial_y\tilde{v}^P)_{\Omega_{\rm PML}} - k_1^2(\sigma\tilde{u}^P, \tilde{v}^P)_{\Omega_{\rm PML}}.$$

3.4. Minimizing the reflection coefficients

Let $\mu_{\min}$ and $\mu_{\max}$ be the smallest and the largest axial frequencies of all propagating modes,

$$\mu_{\min} = \min\{\mu_n \ : \ \mu_n^2 > 0\} \ \text{ and } \ \mu_{\max} = \max\{\mu_n \ : \ \mu_n^2 > 0\},$$

respectively. Also, we let $\tilde{\mu}_{\min} = \min\{|\mu_n| \ : \ \mu_n^2 < 0\}$ be the smallest decay rate of evanescent modes.

Choosing the optimal CRBC parameters is, in fact, is related to the minimization of the maximal $|\mathcal{Z}_n|$, known as the third Zolotarev problem, and it has a long history dating back the late 1870s [15], see also [1, 12, 13]. It turns out that the optimal imaginary CRBC parameters for propagating modes whose axial frequencies are in $[\mu_{\min}, \mu_{\max}]$ are obtained by solving the min-max problem

$$\rho_{prop} = \min_{a_j, \tilde{a}_j \in i\mathbb{R}^-} \max_{\mu_{\min}\le\mu\le\mu_{\max}} \prod_{j=0}^{n_p-1} \left|\frac{(a_j + i\mu)(\tilde{a}_j + i\mu)}{(a_j - i\mu)(\tilde{a}_j - i\mu)}\right|. \qquad (3.35)$$

The min-max problem (3.35) can be solved by using Remez algorithm [13] numerically or by using elliptic functions [1, 7, 10] analytically. In addition, the minimal values decay exponentially with respect to the number of parameters,

$$\rho_{prop} \le 2e^{-\frac{n\pi^2}{\ln(16/\gamma)}} \qquad (3.36)$$

with $\gamma = \mu_{\min}/\mu_{\max}$, $n = n_p$.

With regard to the real CRBC parameters for evanescent modes we first determine an upper bound $\tilde{\mu}_{\max}$ of the decay rates of evanescent modes that CRBC is responsible for by solving the inequality

$$e^{-2\tilde{\mu}_{\max}\sigma_0 M} \le e^{-2\mu_{\min}\sigma_0 M}\rho_{prop},$$

that is,

$$\tilde{\mu}_{\max} = \frac{1}{2\sigma_0 M}\left(2\mu_{\min}\sigma_0 M + \ln\frac{1}{\rho_{prop}}\right),$$

and the reflection coefficients of evanescent modes of $\tilde{\mu}_n$ larger than $\tilde{\mu}_{\max}$ are smaller than $e^{-2\mu_{\min}\sigma_0 M}\rho_{prop}$ due to PML without considering the effect of

CRBC. Then we solve the similar min-max problem for evanescent modes whose decay rates are in $[\tilde{\mu}_{\min}, \tilde{\mu}_{\max}]$,

$$\rho_{evan} = \min_{a_j, \tilde{a}_j \in \mathbb{R}^+} \max_{\tilde{\mu}_{\min} \le \tilde{\mu} \le \tilde{\mu}_{\max}} \prod_{j=n_p}^{n_p+n_e-1} \left| \frac{(a_j - \tilde{\mu})(\tilde{a}_j - \tilde{\mu})}{(a_j + \tilde{\mu})(\tilde{a}_j + \tilde{\mu})} \right|, \tag{3.37}$$

and its minimal value also decreases at the rate of (3.36) with $\gamma = \tilde{\mu}_{\min}/\tilde{\mu}_{\max}$, $n = n_e$. Therefore it can be concluded that the reflection coefficients are bounded by

$$|\mathcal{R}_n| \le \begin{cases} e^{-2\mu_{\min}\sigma_0 M}\rho_{prop} & \text{for } \mu_n \in [\mu_{\min}, \mu_{\max}], \\ e^{-2\tilde{\mu}_{\min}\sigma_0 M}\rho_{evan} & \text{for } \tilde{\mu}_n \in [\tilde{\mu}_{\min}, \tilde{\mu}_{\max}], \\ e^{-2\tilde{\mu}_{\max}\sigma_0 M} & \text{for } \tilde{\mu}_n \in (\tilde{\mu}_{\max}, \infty) \end{cases}$$

which approaches zero rapidly as n_p, n_e and $\sigma_0 M$ tend toward infinity. We let $\mathcal{E}_{\text{HABC}}$ be the maximal reflection coefficient of HABC,

$$\mathcal{E}_{\text{HABC}} = \max\{e^{-2\mu_{\min}\sigma_0 M}\rho_{prop}, \quad e^{-2\tilde{\mu}_{\min}\sigma_0 M}\rho_{evan}\}. \tag{3.38}$$

In the sequel, we also will use $\mathcal{E}_{\text{HABC}}$ as the maximal reflection coefficient multiplied by a generic constant C independent of CRBC and PML for the later analysis. The constant C throughout the paper may take different values in different places but does not depend on CRBC and PML parameters.

If there exist grazing modes, some CRBC parameters are set up aiming at the modes to eliminate completely reflection errors from such modes. These types of modes are related with the small frequencies μ_n such that $|e^{-2i\mu_n\sigma M}|$ is larger than desired tolerance τ from the PML point of view or are related with the small $\gamma \ll 1$ in (3.36) from the CRBC point of view. To distinguish grazing modes from others, let $\mathcal{N}_0$ be the set of indices n of grazing modes, and $\mathcal{N}_1 = \mathbb{Z} \setminus \mathcal{N}_0$. In other words, $\mathcal{N}_0$ consists of only a few indices such that $|e^{-2\mu_n\sigma M}| > \tau$ and $-i\mu_n$ for $n \in \mathcal{N}_0$ is taken as CRBC parameters. The rest of modes of $n \in \mathcal{N}_1$ are taken care of by optimized CRBC and PML. In this case, $\mu_{\min}$ and $\tilde{\mu}_{\min}$ are defined as

$$\mu_{\min} = \min\{\mu_n \ : \ \mu_n^2 > 0, n \in \mathcal{N}_1\} \quad \text{and} \quad \tilde{\mu}_{\min} = \min\{|\mu_n| \ : \ \mu_n^2 < 0, n \in \mathcal{N}_1\}.$$

4. Well-posedness and convergence

In this section we investigate the well-posedness of the problem (3.34) and the convergence of solutions u of (3.34) in Ω to the solution u^{to} of (2.7). To do this, by using the DtN operator for HABC we rewrite the problem (3.34) as an equivalent problem to find $u \in H^1_{\alpha,0}(\Omega)$ satisfying

$$a_{\Omega_\Omega}(u, v) - \langle T_{\text{HABC}}(u), v\rangle_\Gamma = \langle g, v\rangle_\Gamma \text{ for } v \in H^1_{\alpha,0}(\Omega). \tag{4.39}$$

We first show the convergence of T_{HABC} to T in a norm sense.

Lemma 4.1. *Assume that* CRBC *parameters are chosen as in Subsection 3.4 and let* $\mathcal{E}_{\mathrm{HABC}}$ *be defined by* (3.38)*. Then for sufficiently large* n_p*,* n_e *and* $\sigma_0 M$*, it holds that*

$$\|(T - T_{\mathrm{HABC}})v\|_{H_\alpha^{-1/2}(\Gamma)} \leq \mathcal{E}_{\mathrm{HABC}} \|v\|_{H_\alpha^{1/2}(\Gamma)}$$

for $v \in H_\alpha^{1/2}(\Gamma)$*.*

Proof. By using the definition of T and T_{HABC} given by (2.4) and (3.26), we can show that

$$(T - T_{\mathrm{HABC}})(v) = \sum_{n \in \mathcal{N}_1} i\mu_n \frac{2\mathcal{R}_n}{1 + \mathcal{R}_n} v_n \phi_n$$

for $v = \sum_{n \in \mathbb{Z}} v_n \phi_n$. Since $\mathcal{E}_{\mathrm{HABC}} \ll 1$ for large n_p, n_e and $\sigma_0 M$, it follows that

$$\begin{aligned} \|(T - T_{\mathrm{HABC}})v\|^2_{H_\alpha^{-1/2}(\Gamma)} &\leq \mathcal{E}^2_{\mathrm{HABC}} \sum_{n \in \mathcal{N}_1} \frac{|\mu_n|^2}{1 + \lambda_n^2} (1 + \lambda_n^2)^{1/2} |v_n|^2 \\ &\leq \mathcal{E}^2_{\mathrm{HABC}} \|v\|^2_{H_\alpha^{1/2}(\Gamma)}, \end{aligned}$$

which completes the proof. □

Now we can prove the well-posedness of the problem (4.39) for the solution u defined on the physical domain Ω.

Theorem 4.2. *Assume that* ω *is not an eigenvalue of the diffraction grating problem. If* $\mathcal{E}_{\mathrm{HABC}}$ *is sufficiently small, then there exists a unique solution* $u \in H^1_{\alpha,0}(\Omega)$ *to the problem* (4.39) *satisfying*

$$\|u\|_{H^1(\Omega)} \leq C \|g\|_{H_\alpha^{1/2}(\Gamma)}.$$

Proof. We will first show that

$$\|w\|_{H^1(\Omega)} \leq C \sup_{0 \neq \psi \in H^1_{\alpha,0}(\Omega)} \frac{|a_\Omega(w, \psi) - \langle T_{\mathrm{HABC}}(w), \psi \rangle_\Gamma|}{\|\psi\|_{H^1(\Omega)}} \tag{4.40}$$

for $w \in H^1_\alpha(\Omega)$. In fact, this inf-sup condition follows from the inf-sup condition (2.8) of the problem (2.7) and the convergence result of T_{HABC}. We start with the inf-sup condition (2.8)

$$\|w\|_{H^1(\Omega)} \leq C \sup_{0 \neq \psi \in H^1_{\alpha,0}(\Omega)} \frac{|a_\Omega(w, \psi) - \langle T(w), \psi \rangle_\Gamma|}{\|\psi\|_{H^1(\Omega)}}.$$

Lemma 4.1 leads us to

$$\begin{aligned} \|w\|_{H^1(\Omega)} &\leq C \left(\sup_{0 \neq \psi \in H^1_{\alpha,0}(\Omega)} \frac{|a_\Omega(w, \psi) - \langle T_{\mathrm{HABC}}(w), \psi \rangle_\Gamma|}{\|\psi\|_{H^1(\Omega)}} + \|(T - T_{\mathrm{HABC}})(w)\|_{H_\alpha^{-1/2}(\Gamma)} \right) \\ &\leq C \sup_{0 \neq \psi \in H^1_{\alpha,0}(\Omega)} \frac{|a_\Omega(w, \psi) - \langle T_{\mathrm{HABC}}(w), \psi \rangle_\Gamma|}{\|\psi\|_{H^1(\Omega)}} + \mathcal{E}_{\mathrm{HABC}} \|w\|_{H^1(\Omega)}. \end{aligned}$$

For HABC such that $\mathcal{E}_{\mathrm{HABC}} < 1/2$, we obtain (4.40).

Since $a_\Omega(w,\psi) = a_\Omega(\overline{\psi},\overline{w})$ and $\langle T_{\mathrm{HABC}}(w),\psi\rangle_\Gamma = T_{\mathrm{HABC}}(\overline{\psi}),\overline{w}\rangle_\Gamma$ the inf-sup condition for the adjoint problem also holds true and hence the well-posedness follows. □

The convergence of approximate solutions satisfying HABC is established in the next theorem.

Theorem 4.3. *Assume the same conditions as those in Theorem 4.2, and let u^{to} and u be the solutions to the problems* (2.7) *and* (4.39)*, respectively. Then the approximate solution u converges to u^{to} exponentially,*

$$\|u^{to}-u\|_{H^1(\Omega)} \le \mathcal{E}_{\mathrm{HABC}}\|g\|_{H_\alpha^{1/2}(\Gamma)}.$$

Proof. From (2.7) and (4.39), it can be shown that

$$a_\Omega(u^{to}-u,v) - \langle T(u^{to}-u),v\rangle_\Gamma = \langle (T-T_{\mathrm{HABC}})u,v\rangle_\Gamma$$

for all $v \in H^1_{\alpha,0}(\Omega)$. Therefore, the inf-sup condition (2.8), Lemma 4.1 and Theorem 4.2 give

$$\begin{aligned}\|u^{to}-u\|_{H^1(\Omega)} &\le C\|(T-T_{\mathrm{HABC}})u\|_{H_\alpha^{-1/2}(\Gamma)}\\ &\le \mathcal{E}_{\mathrm{HABC}}\|g\|_{H_\alpha^{1/2}(\Gamma)},\end{aligned}$$

which is the desired convergence estimate. □

We are left with the stability of the CRBC auxiliary variables $\mathcal{U}$ and the PML solution $\tilde{u}^P$ to the problem (3.34). To do this, we recall (3.15) and denote $Z_n^{l,m} = Q_n^{l,m}R_n^{l,m}$. Clearly, $|Z_n^{l,m}| \le 1$. Since $|a_j| \le |\tilde{a}_j|$, $|Q_n^{l,m}| \le 1$ also holds true. In addition, it is shown in [10] that the parameters obtained for minimizing the reflection errors satisfy

$$1 < \left|\frac{\tilde{a}_j}{a_j}\right| < C, \tag{4.41}$$

from which the next lemma follows, see [10].

Lemma 4.4. *Assume that* CRBC *parameters are selected as described in Subsection 3.4. Then it holds that*

$$\frac{|1+Q_n^{j,j}|}{\sqrt{|a_j+\tilde{a}_j|}} \le \frac{C}{\sqrt{|\mu_n|}} \quad \textit{and} \quad \frac{|1+R_n^{j,j}|}{\sqrt{|a_j+\tilde{a}_j|}} \le \frac{C}{\sqrt{|\mu_n|}}$$

Now, we are ready to estimate the CRBC auxiliary variables.

Lemma 4.5. *Let $\mu_{\min} = \min\{|\mu_n| \ : \ n \ne N\}$ be the minimal nonzero axial frequency. Then the auxiliary variables $\mathcal{U} = (u^0,u^1,\dots,u^P)^\top$ solving* (3.27)-(3.29) *satisfy the estimates*

$$\|\mathcal{U}\|_{\boldsymbol{H}_\alpha^1(\Gamma)} \le \frac{C\sqrt{P}}{\sqrt{\mu_{\min}}}\|u\|_{H^1(\Omega)}$$

and

$$\|u^P\|_{H_\alpha^{1/2}(\Gamma)} \le C\|u\|_{H^1(\Omega)}.$$

Proof. We use (3.31) to see that for $n \neq N$

$$u_n^j + u_n^{j+1} = \left[(1+Q_n^{j,j})Q_n^{0,j-1} + (1+R_n^{j,j})Q_n^{0,j}Z_n^{j+1,P-1}e^{2i\mu_n\sigma M}\right] A_n.$$

Since $|Q_n^{l,m}| \leq 1$ and $|Z_n^{l,m}e^{2i\mu_n\sigma M}| \leq 1$, by Lemma 4.4 it can be shown that

$$\frac{|u_n^j + u_n^{j+1}|}{\sqrt{|a_j+\tilde{a}_j|}} \leq \left(\frac{|1+Q_n^{j,j}|}{\sqrt{|a_j+\tilde{a}_j|}} + \frac{|1+R_n^{j,j}|}{\sqrt{|a_j+\tilde{a}_j|}}\right)|A_n| \leq \frac{C}{\sqrt{|\mu_n|}}|A_n|. \tag{4.42}$$

Similarly, for $n = N$

$$u_N^j + u_N^{j+1} = (1+Q_N^{j,j})Q_N^{0,j-1}A_N = \left(1+\frac{a_j}{\tilde{a}_j}\right)Q_N^{0,j-1}A_n,$$

which yields that

$$\frac{|u_N^j + u_N^{j+1}|}{\sqrt{|a_j+\tilde{a}_j|}} \leq \frac{C}{\sqrt{|\tilde{a}_j|}}|A_N| \leq \frac{C}{\sqrt{\mu_{\min}}}|A_N|. \tag{4.43}$$

Combining (4.42) and (4.43) gives the desired estimate,

$$\begin{aligned}\|\mathcal{U}\|^2_{\boldsymbol{H}^1_\alpha(\Gamma)} &\leq C\left(\frac{P(1+\lambda_N^2)}{\mu_{\min}}|A_N|^2 + \sum_{n\in\mathbb{Z},n\neq N}\frac{P(1+\lambda_n^2)}{|\mu_n||1+\mathcal{R}_n|^2}|(1+\mathcal{R}_n)A_n|^2\right)\\ &\leq \frac{CP}{\mu_{\min}}\|u^0\|^2_{H^{1/2}_\alpha(\Gamma)} \leq \frac{CP}{\mu_{\min}}\|u\|^2_{H^1(\Omega)}.\end{aligned}$$

The second and third inequalities are obtained by using the fact that $u_n^0 = (1+\mathcal{R}_n)A_n$ and a trace inequality for $u^0 = u|_\Gamma$.

Similarly, we use again (3.31) with $j = P$, $|Q_n^{0,P-1}| \leq 1$ and a trace inequality to show that

$$\begin{aligned}\|u^P\|^2_{H^{1/2}_\alpha(\Gamma)} &= (1+\lambda_N^2)^{1/2}|Q_N^{0,P-1}A_N|^2\\ &\quad + \sum_{n\in\mathbb{Z},n\neq N}(1+\lambda_n^2)^{1/2}|Q_n^{0,P-1}(1+e^{2i\mu_n\sigma M})|^2|A_n|^2\\ &\leq C\|u^0\|^2_{H^{1/2}_\alpha(\Gamma)} \leq C\|u\|^2_{H^1(\Omega)},\end{aligned}$$

which completes the proof. □

For the stability result of the PML solution $\tilde{u}^P$, we show the coercivity of the sesquilinear form $a_{\Omega_{\rm PML}}(\cdot,\cdot)$ in $H^1_{\alpha,0}(\Omega_{\rm PML})$, the subspace of functions vanishing on Γ in $H^1_\alpha(\Omega_{\rm PML})$. It can be established by the same idea as that in [9].

Lemma 4.6. *The sesquilinear form $a_{\Omega_{\rm PML}}(\cdot,\cdot)$ is coercive in $H^1_{\alpha,0}(\Omega_{\rm PML}) \times H^1_{\alpha,0}(\Omega_{\rm PML})$,*

$$\|w\|_{H^1_\sigma(\Omega_{\rm PML})} \leq C\sigma_0 M|a_{\Omega_{\rm PML}}(w,w)|$$

for all $w \in H^1_{\alpha,0}(\Omega_{\rm PML})$.

Proof. Denoting $\Re = \Re(a_{\Omega_{\rm PML}}(w,w))$ and $\Im = \Im(a_{\Omega_{\rm PML}}(w,w))$, we first note that

$$\frac{1}{\sigma_0}\Re = \|\partial_x w\|^2_{L^2(\Omega_{\rm PML})} + \frac{1}{|\sigma|^2}\|\partial_y w\|^2_{L^2(\Omega_{\rm PML})} - k_1^2\|w\|^2_{L^2(\Omega_{\rm PML})}, \tag{4.44}$$

$$\frac{1}{\sigma_0}\Im = \|\partial_x w\|^2_{L^2(\Omega_{\rm PML})} - \frac{1}{|\sigma|^2}\|\partial_y w\|^2_{L^2(\Omega_{\rm PML})} - k_1^2\|w\|^2_{L^2(\Omega_{\rm PML})}, \tag{4.45}$$

which gives

$$\frac{2}{|\sigma|^2}\|\partial_y w\|^2_{L^2(\Omega_{\rm PML})} = \frac{1}{\sigma_0}\left(\Re - \Im\right). \tag{4.46}$$

Since w vanishes on Γ, we can prove the Poincaré type inequality,

$$\|w\|_{L^2(\Omega_{\rm PML})} \le M\|\partial_y w\|_{L^2(\Omega_{\rm PML})}. \tag{4.47}$$

Therefore, by substituting (4.44) into the definition of the norm $\|\cdot\|_{H^1_\sigma(\Omega_{\rm PML})}$ and then by using (4.47) and (4.46) subsequentially, we obtain

$$\begin{aligned}
\|w\|^2_{H^1_\sigma(\Omega_{\rm PML})} &= |\sigma|\|\partial_x w\|^2_{L^2(\Omega_{\rm PML})} + \frac{1}{|\sigma|}\|\partial_y w\|^2_{L^2(\Omega_{\rm PML})} + |\sigma|\|w\|^2_{L^2(\Omega_{\rm PML})} \\
&= \frac{|\sigma|}{\sigma_0}\Re + (k_1^2+1)|\sigma|\|w\|^2_{L^2(\Omega_{\rm PML})} \\
&\le \frac{2|\sigma| + (k_1^2+1)|\sigma|^3M^2}{2\sigma_0}\Re - \frac{(k_1^2+1)|\sigma|^3M^2}{2\sigma_0}\Im \\
&\le C\sigma_0^2M^2|a_{\Omega_{\rm PML}}(w,w)|^2,
\end{aligned}$$

which is the required inequality. □

The stability of the PML solution $\tilde{u}^P$ solving (3.28)-(3.29) with $\tilde{u}^P = u^P$ on Γ can be proved as follows.

Lemma 4.7. *Let $\tilde{u}^P$ be the solution to the problem* (3.28)-(3.29) *with $\tilde{u}^P = u^P$ on Γ. Then $\tilde{u}^P$ satisfies*

$$\|\tilde{u}^P\|_{H^1_\sigma(\Omega_{\rm PML})} \le C\sigma_0 M\|u\|_{H^1(\Omega)}.$$

Proof. The solution $\tilde{u}^P \in H^1_\alpha(\Omega_{\rm PML})$ satisfies

$$a_{\Omega_{\rm PML}}(\tilde{u}^P, v) = 0 \text{ for } v \in H^1_{\alpha,0}(\Omega_{\rm PML})$$

with the boundary condition $\tilde{u}^P|_\Gamma = u^P$ on Γ. Then it follows from Lemma 4.6 and Lemma 4.5 that

$$\|\tilde{u}^P\|_{H^1_\sigma(\Omega_{\rm PML})} \le C\sigma_0 M\|u^P\|_{H^{1/2}_\alpha(\Gamma)} \le C\sigma_0 M\|u\|_{H^1(\Omega)}$$

and the proof is completed. □

Finally, we can combine all estimates to obtain the well-posedness result of the problem (3.34).

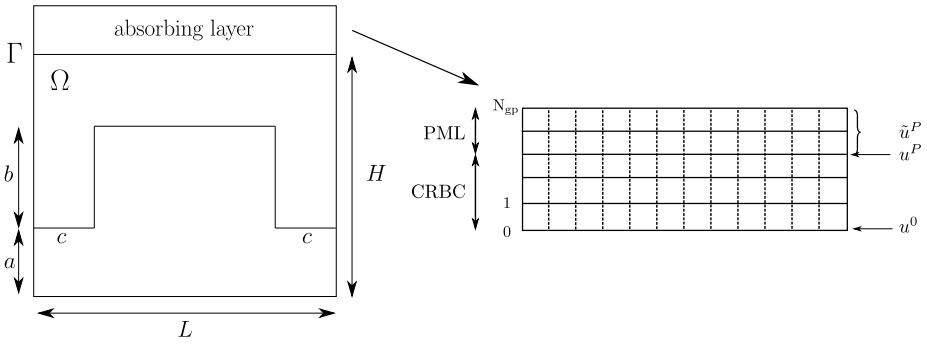


Figure 2: One cell of a diffraction grating of period L and uniform quadrilateral decomposition of the absorbing layer

Theorem 4.8. *The problem* (3.34) *has a unique solution* $(u,\mathcal{U},\tilde{u}^P) \in \boldsymbol{X}$ *satisfying*

$$\|(u,\mathcal{U},\tilde{u}^P)\|_{\boldsymbol{X}} \leq C(\frac{\sqrt{P}}{\sqrt{\mu_{\min}}} + \sigma_0 M)\|g\|_{H_\alpha^{1/2}(\Gamma)}.$$

Proof. The existence and uniqueness is guaranteed by Theorem 4.2 and Lemma 3.1. The stability follows from Theorem 4.2, Lemma 4.5 and Lemma 4.7. □

5. CRBC Parameters for finite element implementation

In this section we discuss how to select the optimal CRBC parameters in the actual finite element implementation. Assume that the absorbing layer for HABC is decomposed into uniform quadrilateral subdomains with mesh size h and we denote by $\mathrm{N_{gp}}$ the number of grid points along the axis of the propagation direction. We further assume that $\mathrm{N_{gp}} \geq \#\mathcal{N}_0$. Among $\mathrm{N_{gp}}$ horizontal grid lines, we assign P grid lines to CRBC and the rest of $\mathrm{N_{gp}} - P$ lines to PML.

The horizontal grid lines assigned to CRBC are considered as Γ for each u^j, $j = 1,\ldots,P$, see Figure 2. Here at least 1 grid point needs to be employed for PML for absorbing fast decaying evanescent modes. For given h, $\mathrm{N_{gp}}$ and σ_0, the procedure to determine the best P and CRBC parameters is described as follows.

(1) If $\mathcal{N}_0 \neq \varnothing$, then we take some a_j and $\tilde{a}_j$ such that $a_j = \tilde{a}_j = -i\mu_n$ for each $n \in \mathcal{N}_0$ and set $\mathrm{N_{gp}}$ to be $\mathrm{N_{gp}} - \#\mathcal{N}_0$.

(2) For $n_p = 0,\ldots,\mathrm{N_{gp}} - 1$,

i. when $n_p = 0$,
we define HABC as PML without CRBC and compute the maximal reflection coefficient $\varrho(n_p)$ of HABC as the maximal reflection coefficient of PML of width $M = h\mathrm{N_{gp}}$,

$$\varrho(n_p) = \max_{n\in\mathcal{N}_1}\{|e^{2i\mu_n\sigma M}|\} = e^{-2\min\{\mu_{\min},\tilde{\mu}_{\min}\}\sigma_0 M}. \tag{5.48}$$

ii. When $n_p \neq 0$,
we solve the min-max problem for propagating modes in the spectral range $[\mu_{\min}, \mu_{\max}]$,

$$\tilde{\rho}_{prop} = \min_{a_j, \tilde{a}_j \in i\mathbb{R}^-} \max_{\mu_{\min} \le \mu \le \mu_{\max}} \prod_{j=0}^{n_p-1} \left| \frac{(a_j + i\mu)(\tilde{a}_j + i\mu)}{(a_j - i\mu)(\tilde{a}_j - i\mu)} \right|. \tag{5.49}$$

It is just the maximal reflection coefficient for propagating modes without considering the effect of PML. After the PML width M is set, the actual maximal reflection coefficient affected by PML for propagating modes will be recalculated.

For $n_e = 0, \ldots, \mathrm{N_{gp}} - n_p - 1$,

(A) when $n_e = 0$,
HABC for evanescent modes acts as PML with $M = h(\mathrm{N_{gp}} - n_p)$ and we compute

$$\rho_{evan}(n_e) = \max_{n \in \mathcal{N}_1, \mu_n^2 < 0} \{|e^{2i\mu_n \sigma M}|\} = e^{-2\tilde{\mu}_{\min}\sigma_0 M}.$$

(B) When $n_e \neq 0$,

(a) set $M = h(\mathrm{N_{gp}} - n_p - n_e)$ and compute the actual maximal reflection coefficient of propagating modes affected by PML with width M,

$$\rho_{prop}(n_e) = e^{-2\mu_{\min}\sigma_0 M} \tilde{\rho}_{prop}.$$

(b) Determine an upper bound $\tilde{\mu}_{\max}$ of the spectral range of evanescent modes that CRBC takes care of by solving the inequality

$$e^{-2\tilde{\mu}_{\max}\sigma_0 M} \le \rho_{prop}(n_e), \tag{5.50}$$

that is,

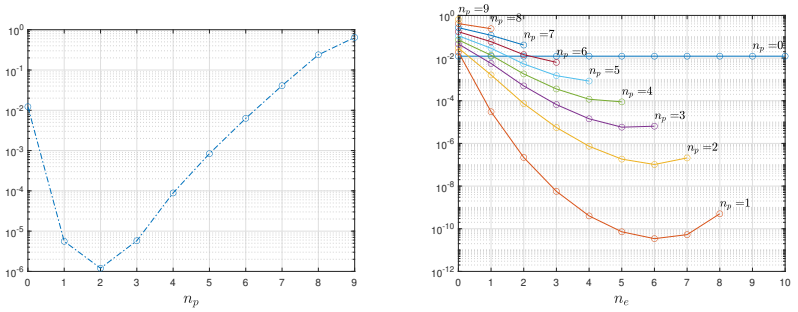
$$\tilde{\mu}_{\max} = \frac{1}{2\sigma_0 M} \ln \frac{1}{\rho_{prop}(n_e)}.$$

(c) Solve the min-max problem for evanescent modes in the spectral range $[\tilde{\mu}_{\min}, \tilde{\mu}_{\max}]$,

$$\tilde{\rho}_{evan} = \min_{a_j, \tilde{a}_j \in \mathbb{R}^+} \max_{\tilde{\mu}_{\min} \le \tilde{\mu} \le \tilde{\mu}_{\max}} \prod_{j=n_p}^{n_p+n_e-1} \left| \frac{(a_j - \tilde{\mu})(\tilde{a}_j - \tilde{\mu})}{(a_j + \tilde{\mu})(\tilde{a}_j + \tilde{\mu})} \right| \tag{5.51}$$

and set the actual maximal reflection coefficient for evanescent modes in the spectral range $[\tilde{\mu}_{\min}, \tilde{\mu}_{\max}]$,

$$\rho_{evan}(n_e) = e^{-2\tilde{\mu}_{\min}\sigma_0 M} \tilde{\rho}_{evan}.$$



(a) Maximal reflection coefficients (b) Plots of $\rho_{evan}(n_e)$ for each n_p

Figure 3: Maximal reflection coefficients for $k_1 = 30$, $\alpha = \pi/3$, $\sigma_0 = 30$.

(d) Determine $\varrho(n_p) = \min_{0 \le n_e \le \mathrm{N_{gp}} - n_p - 1} \{\max\{\rho_{prop}(n_e), \rho_{evan}(n_e)\}\}$ and n_e at which the minimum is attained.

(3) Find n_p for which $\min_{0 \le n_p \le \mathrm{N_{gp}} - 1} \{\varrho(n_p)\}$ takes the minimal value and its corresponding n_e. The optimal CRBC parameters are the solutions to the min-max problems (5.49) and (5.51) for those n_p and n_e.

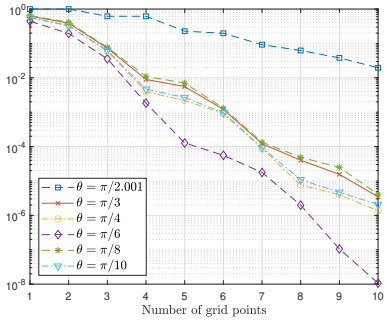
Figure 3 presents an example when $k_1 = 30$, $\alpha = \pi/3$, $\sigma_0 = 30$, $L = 1$, $h = 1/800$ and $\mathrm{N_{gp}} = 10$. According to Figure 3, the maximal reflection coefficient has the minimum value $\approx 1.201 \times 10^{-6}$ when $n_p = 2$ (from Figure 3 (a)) and $n_e = 6$ (from Figure 3 (b)). In this particular example, the best performance of HABC is achieved when optimized CRBC occupies 8 grids and PML does 2 grids. Obviously, HABC with $(n_p, n_e) = (2, 6)$ works better than the pure PML of $n_p = 0$ with the reflection coefficient $\approx 1.229 \times 10^{-2}$.

6. Numerical experiments

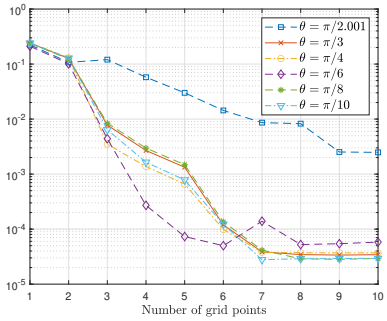
In this section, we present numerical examples that can illustrate the convergence theory for HABC developed in this paper. As the first example, we consider the waveguide problem in $\Omega = (0, L) \times (0, \delta)(L = 1$ and $\delta = 0.05)$ with the α-quasiperiodic boundary condition on the left$(x = 0)$ and the right$(x = L)$ boundaries. In order to demonstrate the convergence of approximate solutions, it is assumed that the wavenumber k_1 is constant in Ω, $k_1 = 4\pi$ or 8π, and the wave fields are given on the bottom boundary Γ_2 $(y = 0)$ as a Dirichlet condition for a wave source so that the exact radiating solution is defined by

$$u(x, y) = \sum_{n=2m_1}^{2m_2} \frac{1}{2(m_2 - m_1 + 1)} e^{i\mu_n y} \phi_n(x)$$

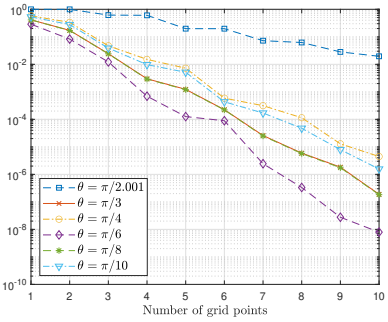
where $m_1 = \lceil -L(k_1 + \alpha)/(2\pi) \rceil$ and $m_2 = \lfloor L(k_1 - \alpha)/(2\pi) \rfloor$. Here m_1 and m_2 are the bounds of indices such that the modes of index n with $m_1 \le n \le m_2$ are propagative and others are evanescent.



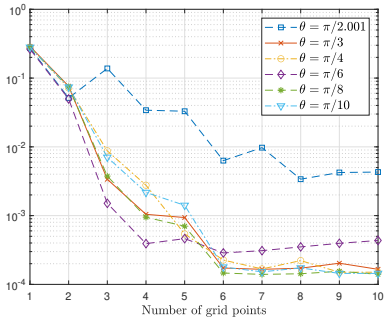
(a) Maximal reflection coefficients for $k_1 = 4\pi$



(b) Relative L^2-error plots for $k_1 = 4\pi$



(c) Maximal reflection coefficients for $k_1 = 8\pi$



(d) Relative L^2-error plots for $k_1 = 8\pi$

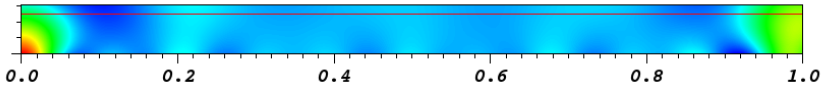
Figure 4: Relative L^2-error plots of HABC with PS-I

θ	$\pi/2.001$		$\pi/3$		$\pi/4$		$\pi/6$		$\pi/8$		$\pi/10$	
$\mu_{\min}$	0.0098		6.2832		7.6574		10.8828		5.9058		7.3863	
$\mu_{\max}$	12.5664		12.4531		12.2939		12.5664		12.4796		12.3351	
γ	7.8500E-4		0.5045		0.6229		0.8660		0.4732		0.5988	
$\tilde{\mu}_{\min}$	0.0098		6.7192		8.4961		14.0496		6.2630		8.1288	
N_{gp}	P	J	P	J	P	J	P	J	P	J	P	J
1	0	1	0	1	0	1	0	1	0	1	0	1
2	0	2	0	2	0	2	0	2	0	2	0	2
3	2	1	2	1	2	1	2	1	2	1	2	1
4	2	2	3	1	3	1	3	1	3	1	3	1
5	4	1	3	2	3	2	3	2	3	2	3	2
6	4	2	5	1	5	1	3	3	5	1	5	1
7	6	1	6	1	6	1	2	5	6	1	6	1
8	7	1	6	2	6	2	6	2	5	3	6	2
9	8	1	5	4	5	4	7	2	5	4	5	4
10	9	1	8	2	5	5	6	4	8	2	5	5

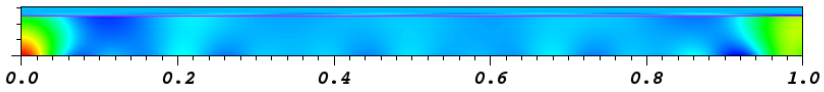
Table 1: Test for $k_1 = 4\pi$; $\gamma = \mu_{\min}/\mu_{\max}$; P and J are the numbers of grid points along the axis of the waveguide used for CRBC and PML, respectively, so that $N_{gp} = P + J$ and PML length M is $M = hJ$.

We compute bilinear finite element approximations on uniform quadrilateral mesh with the mesh size $h = 1/800$ with the help of the finite element library `deal.II` [2]. HABC with the number of grid points, N_{gp}, along the axis of the waveguide is applied to the top boundary Γ ($y = \delta$) for minimizing reflection errors of wave fields propagating along the positive y-axis, that is, the absorbing layer is the rectangular domain $(0, L) \times (\delta, \delta + hN_{gp})$. In this example, we test the performance of HABC with respect to $1 \le N_{gp} \le 10$ and with respect to the angle $\theta = \pi/m$ ($m = 2.001, 3, 4, 6, 8, 10$) of incident fields associated with the quasiperiodic constant $\alpha = k_1 \sin\theta$. The PML strength is set to be $\sigma_0 = 30$. When CRBC parameters are selected by the procedure in Section 5 assuming $\mathcal{N}_0$ is empty, we call it parameter selection I(PS-I). On the other hand, when we take care of grazing modes, i.e., $\mathcal{N}_0 \neq \varnothing$, it is called parameter selection II(PS-II).

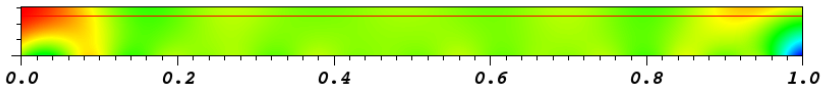
Figure 4 (a) and (c) for $k_1 = 4\pi$ and 8π, respectively, report that the maximal reflection coefficients of HABC for PS-I decay exponentially with increasing N_{gp} for all θ. Table 1 also shows how N_{gp} is split into P(number of grid points for CRBC) and J(number of grids points for PML) to achieve the maximal reflection coefficients for $k_1 = 4\pi$ depicted in Figure 4 (a). It indicates that usually more than half of the grid points are used for CRBC. Accordingly the actual convergence of approximate solutions with respect to the L^2-norm is observed in Figure 4 (b) and (d) as N_{gp} increases until finite element errors are dominant. It is worth pointing out that when $\theta = \pi/6$, there are cutoff modes for both k_1, i.e, there exist integers n such that $\lambda_n = k_1/2 + 2n\pi = \pm k_1$. We see that HABC terminated by PML with the Neumann condition can well



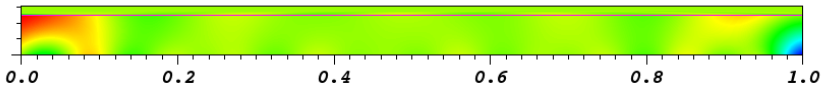
(a) Real part of the exact solution



(b) Real part of the approximate solution



(c) Imaginary part of the exact solution



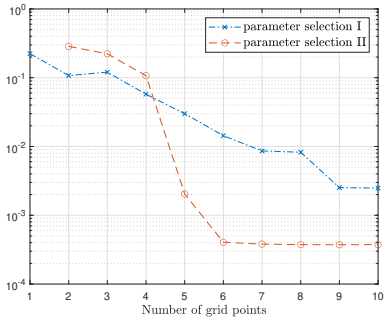
(d) Imaginary part of the approximate solution

Figure 5: Snapshots of the solutions when $k_1 = 30$, $\theta = \pi/3$, $\delta = 0.05$, $\mathrm{N_{gp}} = 10$, $h = 1/800$.

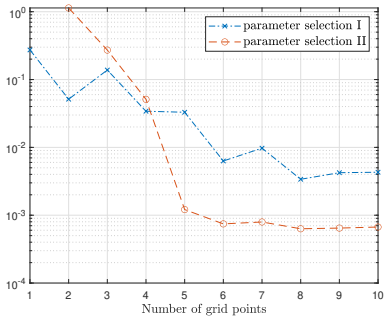
capture the character of the cutoff modes, that would become modes difficult for HABC with the Dirichlet condition to handle. The snapshots of the exact and approximate solutions for $k_1 = 30$, $\theta = \pi/3$ and $\mathrm{N_{gp}} = 10$ are presented in Figure 5. The horizontal line represents the interface between the physical domain Ω and the absorbing layer of HABC.

One of interesting results exhibited in Figure 4 is the relatively slow convergence when $\theta = \pi/2.001$ for grazing incidence. It turns out, in this case, that there exist grazing modes of small axial frequency $\mu_{\min} = \tilde{\mu}_{\min} \approx 0.0098$ of both types (propagating and evanescent) of modes as seen in Table 1. In particular, for propagating modes, we have $\gamma = \mu_{\min}/\mu_{\max}$ relatively small compared with other cases and hence the presence of the grazing modes is the reason for relatively large reflection coefficients in view of (3.36) and (5.48), which in turn results in the slow convergence. In order to improve the performance of HABC, we can apply HABC taking care of the grazing modes by using PS-II with $a_0 = \tilde{a}_0 = -i\mu_{\min}$ and $a_1 = \tilde{a}_1 = \tilde{\mu}_{\min}$ for the grazing modes. As shown in Figure 6, the errors resulting from HABC with PS-II are larger than those from HABC with PS-I for small $\mathrm{N_{gp}} \leq 4$ since two grid points are already assigned for the grazing modes and so there are not much resource that can reduce errors from other modes. However, HABC with PS-II can reduce the errors drastically as $\mathrm{N_{gp}}$ becomes larger than 4 and approximate solutions resulting from HABC with PS-II converge faster that those from HABC with PS-I.

We also compare the performance of HABC with PML truncated by the DtN



(a) Relative L^2-errors for $k_1 = 4\pi$



(b) Relative L^2-errors for $k_1 = 8\pi$

Figure 6: Relative L^2-error plots for $\theta = \pi/2.001$ of HABC with parameter selection I and II

condition for grazing modes introduced in [14]. To this end, we consider the problem with $\theta = \pi/2.001$ and $\theta = \pi/2.05$. When $\theta = \pi/2.001$, $\mu_{\min} = 0.0098$ and $\tilde{\mu}_{\min} = 0.0098$ as seen in Table 1, and when $\theta = \pi/2.05$, $\mu_{\min} = 0.9626$ and $\tilde{\mu}_{\min} = 0.9630$. We use PML of a piecewise constant stretching function $\tilde{y}(y) = \sigma y$ with

$$\sigma = \begin{cases} 1 & \text{for } y \le H \\ \sigma_r + i\sigma_i & \text{for } y > H \end{cases}$$

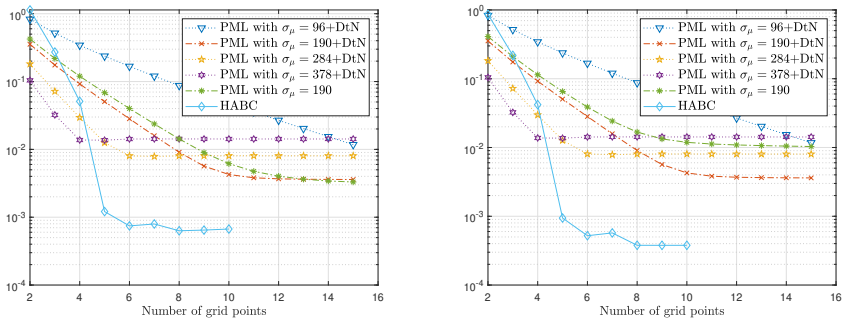
Since the DtN condition deals with the grazing modes, the coordinate stretching function σ of PML is chosen in order to reduce the reflection errors of the modes corresponding to the next smallest axial frequency and the next smallest decay rate, denoted by $\mu_{\min}$ and $\tilde{\mu}_{\min}$ again, respectively, by abuse of notation: in this experiment

$$\begin{aligned} &\theta = \pi/2.001 && \mu_{\min} = 16.6237 && \tilde{\mu}_{\min} = 18.8495, \\ &\theta = \pi/2.05 && \mu_{\min} = 16.6028 && \tilde{\mu}_{\min} = 18.8188. \end{aligned}$$

Noting that the reflection error of PML is determined by the minimal value of $e^{-2\sigma_i \mu_{\min} M}$ and $e^{-2\sigma_r \tilde{\mu}_{\min} M}$ with PML width M, the parameters σ_r and σ_i are chosen in a way that $\sigma_i \mu_{\min} = \sigma_r \tilde{\mu}_{\min}$ (which is denoted by σ_μ) so that the reflection errors from propagating modes and evanescent modes are balanced. The values of σ_μ are set to be $\sigma_\mu = 96, 190, 284$ and 378 and the corresponding stretching constants σ of PML are

$$\begin{aligned} &5.09 + 5.77i, \quad 10.07 + 11.42i, \quad 15.11 + 17.14i, \quad 20.10 + 22.79i \text{ for } \theta = \pi/2.001, \\ &5.10 + 5.78i, \quad 10.09 + 11.44i, \quad 15.09 + 17.10i, \quad 20.08 + 22.76i \text{ for } \theta = \pi/2.05, \end{aligned}$$

respectively. The relative L^2-errors with respect to number of grid points, $\mathrm{N_{gp}}$, in the absorbing layer are presented in Figure 7. Among the results for various σ_μ, PML with $\sigma_\mu = 190$ can produce the best approximate solutions for sufficiently large $\mathrm{N_{gp}}$. As noticed in [9], finite element approximate solutions



(a) $\theta = \pi/2.001$ (b) $\theta = \pi/2.05$

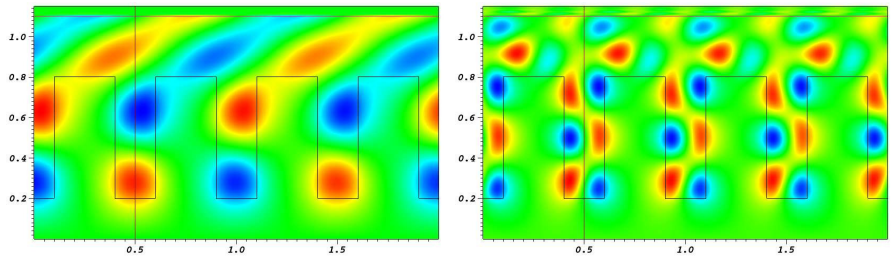
Figure 7: Comparison of HABC with PML truncated by the DtN condition in terms of relative L^2-error.

obtained by uniform mesh and PML with large σ can be worsen due to the high anisotropy of the PML problem as for the case $\sigma_\mu = 284$ and 378. Figure 7 also shows the results of PML with $\sigma_\mu = 190$ truncated simply by a homogeneous Neumann condition (green curves with $*$-marker). Since $\mu_{\min}$ and $\tilde{\mu}_{\min}$ are too small for $\theta = \pi/2.001$, the homogeneous Neumann condition can give as accurate solutions as the DtN truncation does. In contrast, it is shown that when $\theta = \pi/2.05$, PML truncated by the DtN condition is superior to PML truncated by the homogeneous Neumann condition. However, it reveals that approximate solutions obtained by HABC with parameter selection II stands out from all other results for both θ. These results are not surprising because the number of grids points P and J, used for CRBC and PML respectively, are determined in such a way that the maximal reflection coefficient is minimized.

The second experiment is conducted to illustrate the performance of HABC applied to a diffraction grating problem. The example is the diffraction grating of period $L = 0.5$ whose one cell is shown as in Figure 2 with $a = 0.2$, $b = 0.6$, $c = 0.1$ and $H = 1.1$. Assuming that wavenumber inside the grating media is $k_2 = 1$ and incident wave fields are approaching the diffraction grating structure with $\theta = \pi/6$, we use HABC with $\mathrm{N_{gp}} = 5$, $\sigma_0 = 30$, and finite elements of $h = 0.01$ to solve the α-quasiperiodic problem in one cell for $k_1 = 12.5$ and $k_1 = 8\pi+0.1$. Once finite element solutions are obtained in the cell, we construct solutions in the periodic 4 cells via the α-quasiperiodicity, which are presented in Figure 8. It illustrates well scattered fields and its α-quasiperiodicity in the diffraction grating.

7. Acknowledgment

This research of the author was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF-2018R1D1A1B07047416) funded by the Ministry of Education, Science and Technology.



(a) $k_1 = 12.5$, $k_2 = 1$, $\theta = \pi/6$ (b) $k_1 = 8\pi + 0.1$, $k_2 = 1$, $\theta = \pi/6$

Figure 8: Snapshots of real parts of approximate solutions to the diffraction grating problem

References

[1] N. I. Akhiezer, Elements of the theory of elliptic functions, American Mathematical Society, Providence, RI, 1990.

[2] W. Bangerth, R. Hartmann, G. Kanschat, deal.II—a general-purpose object-oriented finite element library, ACM Trans. Math. Software 33 (4) (2007) 24.

[3] G. Bao, Finite element approximation of time harmonic waves in periodic structures, SIAM J. Numer. Anal. 32 (4) (1995) 1155–1169.

[4] J.-P. Bérenger, A perfectly matched layer for the absorption of electromagnetic waves, J. Comput. Phys. 114 (2) (1994) 185–200.

[5] J.-P. Bérenger, Three-dimensional perfectly matched layer for the absorption of electromagnetic waves, J. Comput. Phys. 127 (2) (1996) 363–379.

[6] D. C. Dobson, Optimal design of periodic antireflective structures for the Helmholtz equation, European J. Appl. Math. 4 (4) (1993) 321–339.

[7] V. Druskin, S. Güttel, L. Knizhnerman, Near-optimal perfectly matched layers for indefinite Helmholtz problems, SIAM Rev. 58 (1) (2016) 90–116.

[8] T. Hagstrom, S. Kim, Complete radiation boundary conditions for the Helmholtz equation I: waveguides, Numer. Math. 141 (4) (2019) 917–966.

[9] S. Kim, Error analysis of PML-FEM approximations for the Helmholtz equation in waveguides, ESAIM Math. Model. Numer. Anal. 53 (4) (2019) 1191–1222.

[10] S. Kim, Analysis of complete radiation boundary conditions for the Helmholtz equation in perturbed waveguides, J. Comput. Appl. Math. 367 (2020) 112458.

[11] S. Kim, Fractional order Sobolev spaces for the Neumann Laplacian and the vector Laplacian, J. Korean. Math. Soc. In press.

[12] A. A. Medovikov, V. I. Lebedev, Variable time steps optimization of L_ω-stable Crank-Nicolson method, Russian J. Numer. Anal. Math. Modelling 20 (3) (2005) 283–303.

[13] P. Petrushev, V. Popov, Rational Approximation of Real Functions, vol. 28 of Encyclopedia of Mathematics, Cambridge University Press, Cambridge, 1987.

[14] W. Zhou, H. Wu, An adaptive finite element method for the diffraction grating problem with PML and few-mode DtN truncations, J. Sci. Comput. 76 (3) (2018) 1813–1838.

[15] E. I. Zolotarev, Applications of elliptic functions to problems on functions deviating least or most from zero (Russian), Zap. Imper. Akad. Nauk St. Petersburg, 30 (1877), no 5; reprinted in his Collected works Vol 2. Izadt. Akad. Nauk SSSR, Moscow, Ibuch. Fortschritte Math. 9 (343) (1932) 1–59.