Numerische Mathematik (2019) 141:917–966
https://doi.org/10.1007/s00211-018-1012-0

Numerische Mathematik

Complete radiation boundary conditions for the Helmholtz equation I: waveguides

Thomas Hagstrom[1] · **Seungil Kim**[2]

Received: 31 July 2017 / Revised: 17 July 2018 / Published online: 1 January 2019

Abstract
We consider the use of complete radiation boundary conditions for the solution of the Helmholtz equation in waveguides. A general analysis of well-posedness, convergence, and finite element approximation is given. In addition, methods for the optimization of the boundary condition parameters are considered. The theoretical results are illustrated by some simple numerical experiments.

Keywords Helmholtz equation · Complete radiation boundary condition · Waveguide

Mathematics Subject Classification 65N12 · 65N30

1 Introduction

In this paper, we shall study time-harmonic wave propagation problems in unbounded waveguides. Waveguides are an important technology with a variety of applications in acoustics, optical communications and so on. Many applications of waveguides

The authors acknowledge the support of ARO Grant W911NF-09-1-0344. The first author was also partially supported by NSF Grant OCI-0904773 and DMS-1418871. This research of the second author was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2015R1D1A1A01057350). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Army Research Office or the National Science Foundation.

✉ Seungil Kim
sikim@khu.ac.kr

Thomas Hagstrom
thagstrom@smu.edu

[1] Department of Mathematics, Southern Methodist University, PO Box 750156, Dallas, TX 75275-0156, USA

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

are found to be posed in large, effectively unbounded, domains. A challenge for the numerical solution of wave propagation problems posed in large domains is the construction and application of domain truncation techniques with high accuracy. The boundary conditions imposed on the artificial boundaries resulting from domain truncation, so-called absorbing boundary conditions (ABCs), should have the following properties

- the artificial boundary produces as little reflection as we wish and so the solution on the truncated domain can be made arbitrarily close to the solution on the original unbounded domain,
- the artificial boundary conditions are easy to implement in the discretized problems using, e.g., finite elements method (FEM) or finite difference method (FDM),
- the numerical methods incorporated with the artificial boundary conditions are stable and robust.

Many ABCs satisfying the properties listed above have been developed, for example, nonlocal boundary conditions based on Dirichlet-to-Neumann (DtN) mappings [3,13,19], high-order local boundary conditions [24,26,27,31], and perfectly matched layers (PMLs) [2,29]. We note that the design of efficient ABCs is also important for scattering problems in exterior domains, which we will consider in a subsequent paper. For general reviews of this subject, see [4,9,15,25,35].

This paper is devoted to developing local high-order absorbing boundary conditions for time-harmonic wave propagation problems in waveguides motivated by complete radiation boundary conditions (CRBCs) for wave propagation problems in the time-domain [17,18]. For time-domain calculations, CRBCs exploit the auxiliary function formulation proposed in [17], which leads to a more efficient and natural implementation of high order radiation conditions than those proposed by Higdon [20,21] and by Givoli and Neta [11]. In addition, it is shown in [17] how optimal parameters can be chosen based on the simulation time, T, the separation, b, of sources and inhomogeneities from the artificial boundary, and the error tolerance, τ. The parameterizations are quite efficient, with the total number of auxiliary functions, P, obeying

$$P \propto \ln\left(\frac{1}{\tau}\right) \cdot \ln\left(\frac{cT}{b}\right), \tag{1.1}$$

with a positive constant c.

The new method that we shall investigate not only fulfills the necessary requirements for ABCs but also has certain advantages. First of all, compared with methods based on DtN mappings [3,13,19], CRBCs do not need the knowledge of eigenfunctions of the transverse Laplace operator on the cross-section of waveguides and the number of propagating modes, though easily-obtained partial information on the distribution of the eigenvalues can be used to improve efficiency.

In addition, as CRBCs are local, the sparsity of the system matrix is retained. In contrast with earlier local boundary condition sequences or PML, CRBCs are constructed to treat evanescent modes as well as propagating modes. Thus they can be placed quite close to wave sources or scatterers without compromising accuracy. This fact will be illustrated in the numerical examples later. Here we note that to handle

evanescent modes the PML width needs to be inversely proportional to the smallest decay rate of evanscent modes so that it can be arbitrarily wide, whereas in such a case we can use suitably chosen nodes, e.g., Newman nodes, and guarantee accuracy independent of how small the smallest decay rate is.

Via the introduction of auxiliary variables, CRBCs, as well as some of the other methods mentioned above, avoid the higher order derivatives involved in product boundary operators of Higdon. Hence, these boundary conditions are compatible with FEM. The literature [10,12,16,31] shows many computational results of these ABCs for wave propagation problems in time- and frequency- domains incorporated with FEM. However, the analysis for finite element problems, e.g., well-posedness and quasi-optimal convergence, has not been available in any case. In the present paper, we will provide an improved analysis for the finite element application to time-harmonic wave propagation problems with CRBCs in waveguides. In general, the unique solvability and quasi-optimal convergence of finite element approximations to solutions of indefinite problems satisfying a Gårding type inequality and the regularity of the adjoint problem is obtained by an argument of Schatz [33]. Schatz's argument requires that the regularity of the continuous variational problem be established and that the mesh size h be small enough. That is, $0 < h < h_0$, where h_0 depends on the regularity constant of the elliptic problem. In CRBC applications, it turns out that the regularity constant may increase polynomially as P grows (a PML application has the similar result that the stability constant depends on the width of the layer polynomially [5]), which means that for large P a smaller mesh h may be required to retain the unique solvability and quasi-optimal convergence. As the error due to the approximate boundary condition typically converges exponentially with increasing order, this possible restriction on the mesh is not likely to be important. We note that in our numerical simulations no dependence on P of the mesh size for the solvability of the discretized problem or the quasi-optimality of the finite element approximations was observed.

This paper is organized as follows. In Sect. 2 we study analytic solutions of a time-harmonic waveguide model. We define the CRBCs for wave propagation problems in the frequency-domain in Sect. 3. Section 4 is devoted to reformulation of the model problem to a variational form and in Sect. 5 existence and uniqueness of solutions to the Helmholtz equation satisfying CRBCs is established. Section 6 includes the convergence analysis of the continuous problem and parameter optimization is discussed in Sect. 7. We analyze the stability and regularity of the variational problem in Sect. 8 and discuss the finite element analysis in Sect. 9. Finally, in Sect. 10 numerical examples that confirm the theories are presented. Note that we cannot directly use the time-domain analysis in the frequency domain, as in the time domain we use the finite simulation time, T, in an essential way. As a result the parameter optimization problem considered here is different and, in fact, more difficult.

2 Fourier series of solutions to the Helmholtz equation in waveguides

We consider a time-harmonic waveguide problem

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega_\infty \tag{2.1}$$

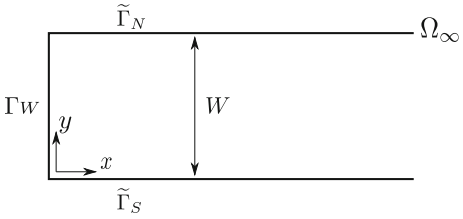


Fig. 1 Geometry of the semi-infinite waveguide Ω_∞ in $\mathbb{R}^2$, $\tilde{\Gamma}_T = \tilde{\Gamma}_N \cup \tilde{\Gamma}_S$

on a semi-infinite waveguide $\Omega_\infty = \{(x, y) \in \mathbb{R} \times \mathbb{R}^{d-1} : x > 0,\ y \in \Theta\}$, $d = 2$ or 3. Here Θ is a bounded subset of $\mathbb{R}^{d-1}$ with a smooth boundary. (For the numerical experiments we will specialize to $\mathbb{R}^2$ with $\Theta = (0, W)$. See Fig. 1). Here k is a positive wavenumber. For definiteness we assume the lateral waveguide boundary is sound-hard, i.e., the normal flux is equal to zero,

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \tilde{\Gamma}_T \equiv (0, \infty) \times \partial\Theta, \tag{2.2}$$

where ν is the outward unit normal vector on $\tilde{\Gamma}_T$. In addition, we assume that wave sources come from the west boundary Γ_W of Ω_∞ located at $x = 0$ and so it determines the boundary data on Γ_W,

$$u = f \quad \text{on } \Gamma_W. \tag{2.3}$$

This models the practically important case where more complicated physics, geometry, or distributed sources are located in the region $x < 0$.

Solutions of the Helmholtz equation (2.1) can be expressed in a Fourier series in terms of the eigenfunctions of the negative transverse Laplace operator

$$\begin{aligned} \Delta_y Y_n + \lambda_n^2 Y_n &= 0 \quad \text{in } \Theta, \\ \frac{\partial Y_n}{\partial \nu} &= 0 \quad \text{on } \partial\Theta, \end{aligned} \tag{2.4}$$

where λ_n^2 and Y_n are the nth eigenpair. We denote $\mu_n^2 = k^2 - \lambda_n^2$. By choosing normalized eigenfunctions, we have an orthonormal basis consisting of eigenfunctions Y_n. Moreover, as

$$\lim_{n\to\infty} \mu_n^2 = -\infty, \tag{2.5}$$

there are only finitely many $\mu_n^2 > 0$, infinitely many $\mu_n^2 < 0$ and there may be cutoff modes $\mu_n^2 = 0$. We also note that the asymptotic behavior of the eigenvalues is well-known (e.g. [6, Ch. VI, Thm. 20–21]): for some constant A

$$\mu_n^2 \sim -An^{\frac{2}{d-1}}. \tag{2.6}$$

Now, under the time-harmonic assumption $e^{-i\omega t}$ with angular frequency ω, for each μ_n, we only take solutions that propagate to the right or are bounded for $x > 0$,

$$z_n(x) = e^{i\mu_n x}.$$

This represents a propagating mode for $\mu_n^2 > 0$ with $\mu_n > 0$ and an evanescent mode for $\mu_n^2 < 0$ with $\tilde{\mu}_n := \Im(\mu_n) > 0$. In some cases, there is a mode, a so-called cutoff mode, associated with $\mu_n = 0$, for which special care needs to be taken. For ease of exposition we now assume that there exists $N \geq 0$ such that $\mu_N = 0$, $\mu_n^2 > 0$ for all $n < N$ and $\mu_n^2 < 0$ for all $n > N$. However, we will make clear when the absence of such a mode yields substantial improvements in the error and stability estimates. Note that extensions to the case of multiple cutoff modes could similarly be obtained.

Thus, a general solution to the Helmholtz equation satisfying the *outgoing radiation condition* is represented by the Fourier series

$$\begin{aligned} u(x, y) &= \sum_{n=0}^{\infty} A_n e^{i\mu_n x} Y_n(y) \\ &= \sum_{n=0}^{N} A_n e^{i\mu_n x} Y_n(y) + \sum_{n=N+1}^{\infty} A_n e^{-\tilde{\mu}_n x} Y_n(y), \end{aligned} \tag{2.7}$$

which is a superposition of finitely many propagating modes (including a cutoff mode) and infinitely many evanescent modes. Here the Fourier coefficient A_n is determined by the sources from Γ_W,

$$A_n = \int_{\Theta} u(0, y) Y_n(y) \, \mathrm{d}y.$$

The constant C throughout the paper is a generic constant and may be different at different places, but it does not depend on functions. Where the dependence of constants on the parameters of the approximate radiation condition are important we will indicate the dependence via a subscript, C_a. We remark that the construction and analysis can easily be extended to problems with variable coefficients depending only on the transverse coordinates, y, including the important case of layered materials. Also, the theory can be established for a case where the domain Ω_∞ includes any bounded smooth cavity with any inhomogeniety in $x < 0$, and the analysis for this case can be found in [23].

3 Complete radiation boundary conditions

Complete radiation boundary conditions were introduced in [17,18] to provide a rapidly convergent local boundary condition sequence for time-domain calculations. Fundamental differences between the time-domain and frequency-domain cases are:

i. In the frequency domain only a discrete set of modes exists, while in the time domain we must consider the continuum of modes present as k varies along an entire inversion contour;
ii. In the time domain we are only concerned about accuracy up to the simulation time, T, which allows for the continuation of k in the complex plane. In the frequency domain this would be akin to solving a limiting absorption approximation to the

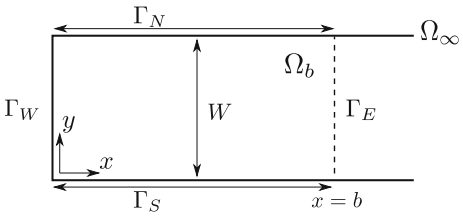


Fig. 2 Geometry of the truncated computational domain Ω_b, $\Gamma_T = \Gamma_N \cup \Gamma_S$

Helmholtz system, and thus the size of the imaginary part would be tied to the accuracy.

Directly, the conditions proposed in the time domain can be simply translated to the frequency domain by the replacement $c^{-1}\frac{\partial}{\partial t} \to -ik$, where c is the wave speed. However, both the analysis and parameter optimization differ.

We truncate the unbounded strip Ω_∞ to a bounded region $\Omega_b = (0, b) \times \Theta$, whose east boundary Γ_E is located at $x = b$ (see Fig. 2). The problem in the finite computational domain Ω_b is

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega_b, \tag{3.1}$$

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{in } \Gamma_T = (0, b) \times \partial\Theta, \tag{3.2}$$

$$u = f \quad \text{on } \Gamma_W. \tag{3.3}$$

To close the problem, we need to supplement it with the CRBC on the east boundary Γ_E. The boundary condition is defined by the following recursive formulas satisfied by auxiliary variables ϕ_j, that also satisfy the Helmholtz equation (3.1) with the sound-hard boundary condition (3.2) on Γ_T:

$$\begin{aligned} \phi_0 &= u, \\ \left(\frac{\partial}{\partial x} + a_j\right)\phi_j &= \left(-\frac{\partial}{\partial x} + a_j\right)\phi_{j+1}, \end{aligned} \tag{3.4}$$

for $j = 0, 1, 2, \ldots$, where a_j are parameters to be chosen for reducing reflection from the artificial boundary. As motivation we note that the recursion terminates if u is a superposition of modes annihilated by one of the operators $(\frac{\partial}{\partial x} + a_j)$. The parameters a_j are chosen as follows:

$$a_j = \begin{cases} -ikc_j & \text{for } j = 0, \ldots, n_p - 1, \\ \sigma_{j-n_p} & \text{for } j = n_p, \ldots, n_p + n_e \end{cases} \tag{3.5}$$

with

$$0 < c_j \le 1 \text{ for } j = 0, \ldots, n_p - 1, \quad \text{and} \quad 0 < \sigma_j \text{ for } j = n_p, \ldots, n_p + n_e. \tag{3.6}$$

In practice, the parameters we take satisfy

$$\mu_{N-1} \le kc_j \le k \quad \text{and} \quad \tilde{\mu}_{N+1} \le \sigma_j \le M_\sigma, \tag{3.7}$$

where μ_{N-1} represents the smallest axial frequency of propagating modes and $\tilde{\mu}_{N+1}$ is the smallest decay rate of evanescent modes. Also, M_σ is an upper bound for the decay rates σ_j of evanescent modes that the CRBC can damp effectively and it can be chosen so that $e^{-M_\sigma b}$ is less than an error tolerance of numerical simulations. These bounds and selection of parameters in practice will be discussed in more detail in Sect. 7. We could choose repeated parameters a_j, however from now on we assume that a_j are all distinct since the parameters in the optimal selection are all different. These recursions are terminated by

$$\phi_{n_p+n_e+1} = 0 \quad \text{on } \Gamma_E. \tag{3.8}$$

Here (n_p, n_e) is called the order of CRBCs and let $P = n_p + n_e$. If a_j is selected to be purely imaginary so that $kc_j = \mu_n > 0$, then the recursion exactly eliminates the corresponding propagating mode, and if a_j is chosen to be real so that σ_j equals the decay rate $\tilde{\mu}_n$ of an evanescent mode, then it does not produce reflection of the corresponding evanescent mode.

Remark 3.1 As suggested for time-domain problems in [17], we may also use parameters a_j of the form

$$a_j = \sigma_j - ikc_j \tag{3.9}$$

for $j = 0, \ldots, P$ with the conditions (3.6). In this case, although the recursions do not annihilate any mode exactly, they damp reflection of propagating modes and evanescent modes simultaneously. In this paper, however, we only investigate CRBCs employing a_j as given in (3.5), which are generally more effective for frequency-domain problems.

For numerical implementation of these boundary conditions, we need to eliminate the derivative of the auxiliary variables with respect to the normal direction from the recursive formulas (3.4). To do this, we apply the operator $\partial/\partial x$ to the Eq. (3.4) for the $(j-1)$th and jth recursion, which yields

$$\frac{\partial^2}{\partial x^2}\phi_{j-1} + \frac{\partial^2}{\partial x^2}\phi_j = a_{j-1}\frac{\partial}{\partial x}\phi_j - a_{j-1}\frac{\partial}{\partial x}\phi_{j-1}, \tag{3.10}$$

and

$$\frac{\partial^2}{\partial x^2}\phi_j + \frac{\partial^2}{\partial x^2}\phi_{j+1} = a_j\frac{\partial}{\partial x}\phi_{j+1} - a_j\frac{\partial}{\partial x}\phi_j. \tag{3.11}$$

Here we eliminate $\partial\phi_{j-1}/\partial x$ from (3.10) and $\partial\phi_{j+1}/\partial x$ from (3.11) by using (3.4) for the $(j-1)$th and jth recursion, respectively, which shows that

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\phi_{j-1}+\frac{\partial^2}{\partial x^2}\phi_j &= a_{j-1}\frac{\partial}{\partial x}\phi_j - a_{j-1}\left(-\frac{\partial}{\partial x}\phi_j + a_{j-1}\phi_j - a_{j-1}\phi_{j-1}\right)\\ &= 2a_{j-1}\frac{\partial}{\partial x}\phi_j - a_{j-1}^2\phi_j + a_{j-1}^2\phi_{j-1},\end{aligned} \tag{3.12}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\phi_j+\frac{\partial^2}{\partial x^2}\phi_{j+1} &= a_j\left(-\frac{\partial}{\partial x}\phi_j + a_j\phi_{j+1} - a_j\phi_j\right) - a_j\frac{\partial}{\partial x}\phi_j\\ &= -2a_j\frac{\partial}{\partial x}\phi_j + a_j^2\phi_{j+1} - a_j^2\phi_j.\end{aligned} \tag{3.13}$$

Now, multiplying (3.12) by $1/a_{j-1}$ and (3.13) by $1/a_j$ and subsequently adding them together produces

$$\begin{aligned}&L_{j,j-1}\frac{\partial^2}{\partial x^2}\phi_{j-1} + L_{j,j}\frac{\partial^2}{\partial x^2}\phi_j + L_{j,j+1}\frac{\partial^2}{\partial x^2}\phi_{j+1}\\ &+ M_{j,j-1}\phi_{j-1} + M_{j,j}\phi_j + M_{j,j+1}\phi_{j+1} = 0,\end{aligned} \tag{3.14}$$

where

$$\begin{aligned}&L_{j,j-1} = \frac{1}{a_{j-1}}, \qquad L_{j,j} = \frac{1}{a_{j-1}} + \frac{1}{a_j}, \qquad L_{j,j+1} = \frac{1}{a_j},\\ &M_{j,j-1} = -a_{j-1}, \qquad M_{j,j} = a_{j-1} + a_j, \qquad M_{j,j+1} = -a_j.\end{aligned} \tag{3.15}$$

To find the connection between the solution $u(=\phi_0)$ and the auxiliary variables on Γ_E, as in the above derivation, we have

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\phi_0+\frac{\partial^2}{\partial x^2}\phi_1 &= a_0\frac{\partial}{\partial x}\phi_1 - a_0\frac{\partial}{\partial x}\phi_0\\ &= a_0\left(-\frac{\partial}{\partial x}\phi_0 + a_0\phi_1 - a_0\phi_0\right) - a_0\frac{\partial}{\partial x}\phi_0\\ &= -2a_0\frac{\partial}{\partial x}\phi_0 + a_0^2\phi_1 - a_0^2\phi_0.\end{aligned}$$

Therefore,

$$-2\frac{\partial}{\partial x}\phi_0 = \frac{1}{a_0}\left(\frac{\partial^2}{\partial x^2}\phi_0 + \frac{\partial^2}{\partial x^2}\phi_1\right) + a_0\phi_0 - a_0\phi_1. \tag{3.16}$$

To obtain our final system, with

$$L_{0,0} = \frac{1}{a_0} \quad \text{and} \quad M_{0,0} = a_0,$$

we define L and M by the $(P+1)\times(P+1)$ symmetric (but not Hermitian) tridiagonal matrices whose non-zero elements $L_{i,j}$ and $M_{i,j}$ are given as above, respectively. We can write the boundary condition in matrix form

$$-\left(2\frac{\partial u}{\partial x}\right)\boldsymbol{e}_0 = L\frac{\partial^2}{\partial x^2}\Phi + M\Phi,$$

where $\boldsymbol{e}_j$ is the standard $(P+1)\times 1$ basis vector whose non-zero element is one at the jth component and $\Phi = (\phi_0, \ldots, \phi_P)^t$ with $\phi_0 = u$ on Γ_E.

Finally, the Helmholtz equation removes all x-derivatives in the equation,

$$-\left(2\frac{\partial u}{\partial x}\right)\boldsymbol{e}_0 = -L\Delta_y\Phi + (-k^2L + M)\Phi.$$

Thus the model problem completed by the CRBCs on Γ_E is to find functions u defined in Ω_b and $\Phi = (\phi_0, \ldots, \phi_P)^t$ defined on Γ_E with $u = \phi_0$ on Γ_E such that

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega_b, \tag{3.17}$$

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma_T, \tag{3.18}$$

$$u = f \quad \text{on } \Gamma_W, \tag{3.19}$$

$$\frac{\partial u}{\partial x}\boldsymbol{e}_0 = \frac{-1}{2}(-L\Delta_y\Phi + (-k^2L + M)\Phi) \quad \text{on } \Gamma_E \tag{3.20}$$

with

$$\frac{\partial \Phi}{\partial \nu} = 0 \quad \text{on } \partial\Gamma_E. \tag{3.21}$$

Remark 3.2 A similar algebraic computation for time-domain problems, in which the contribution of evanescent modes is not negligible, can be found in [16]. For time-domain problems the process of removing the $\partial/\partial x$ operators required a seam function to transit from the recursions for propagating modes to those for evanescent modes, which is not needed in the recursions for frequency-domain problems as time derivatives are not involved and there is no difference between recursions for propagating modes and those for evanescent modes.

4 Variational reformulation

In this section, we reformulate the problem (3.17)–(3.21) to a variational form for a given order (n_p, n_e) of CRBCs with $n_p + n_e = P$. We begin by defining the appropriate Sobolev spaces,

$$\widetilde{H}^1(\Omega_b) = \{\xi \in H^1(\Omega_b) \ : \ \xi|_{\Gamma_E} \in H^1(\Gamma_E)\},$$
$$\widetilde{H}^1_0(\Omega_b) = \{\xi \in \widetilde{H}^1(\Omega_b) \ : \ \xi = 0 \text{ on } \Gamma_W\}.$$

In the sequel, we will use the notations $(\cdot,\cdot)_{\Omega_b}$ and $(\cdot,\cdot)_{\Gamma_E}$ for the L^2-inner product on Ω_b and Γ_E, respectively,

For the space of auxiliary variables, we first introduce the symmetric positive definite matrices $\mathcal{L}$ and $\mathcal{M}$, which are obtained by replacing a_j with $|a_j|$ in L and M, and define

$$\|\Phi\|^2_{\mathcal{L}} := (\mathcal{L}\Phi,\Phi)_{\Gamma_E} = \sum_{j=0}^{P} \frac{1}{|a_j|}\|\phi_j+\phi_{j+1}\|^2_{L^2(\Gamma_E)},$$
$$\|\Phi\|^2_{\mathcal{M}} := (\mathcal{M}\Phi,\Phi)_{\Gamma_E} = \sum_{j=0}^{P} |a_j|\|\phi_j-\phi_{j+1}\|^2_{L^2(\Gamma_E)}$$

and for $\ell = 1, 2$

$$\|\Phi\|^2_{\mathcal{L},\ell} := \sum_{j=0}^{P} \frac{1}{|a_j|}\|\phi_j+\phi_{j+1}\|^2_{H^\ell(\Gamma_E)}, \quad \|\Phi\|^2_{\mathcal{M},\ell} := \sum_{j=0}^{P} |a_j|\|\phi_j-\phi_{j+1}\|^2_{H^\ell(\Gamma_E)}$$

for $\Phi = (\phi_0,\dots,\phi_P)^t \in (L^2(\Gamma_E))^{P+1}$ with $\phi_{P+1} = 0$. We define the Sobolev space $\boldsymbol{V}_{\Gamma_E} = (H^1(\Gamma_E))^{P+1}$ with the norm

$$\|\Phi\|^2_{\boldsymbol{V}_{\Gamma_E}} = \|\Phi\|^2_{\mathcal{L},1} + \|\Phi\|^2_{\mathcal{M}},$$

which is equivalent to the standard product norm of $(H^1(\Gamma_E))^{P+1}$ but the constants involved in the equivalence may depend on P. Furthermore, we introduce fractional Sobolev spaces $H^s(\Gamma_E)$ for $-1 \le s \le 2$ characterized by the norm

$$\|u\|^2_{H^s(\Gamma_E)} = \sum_{n=0}^{\infty} (\lambda_n^2+1)^s |u_n|^2$$

for $u = \sum_{n=0}^{\infty} u_n Y_n$.

Remark 4.1 We note that $H^s(\Gamma_E)$ for $3/2 \le s \le 2$ in this paper is different from a usual fractional Sobolev space. In this case, $H^s(\Gamma_E)$ is the space of functions which are in a usual fractional Sobolev space obtained by real interpolation $[H^1(\Gamma_E), H^2(\Gamma_E)]_{s-1}$ and whose normal derivatives vanish on $\partial\Gamma_E$. However $H^s(\Gamma_E)$ for $-1 \le s < 3/2$ is a usual fractional Sobolev space

$$H^s(\Gamma_E) = \begin{cases} [(H^1(\Gamma_E))^*, L^2(\Gamma_E)]_{s+1}, & -1 \le s \le 0, \\ [L^2(\Gamma_E), H^1(\Gamma_E)]_s, & 0 \le s \le 1, \\ [H^1(\Gamma_E), H^2(\Gamma_E)]_{s-1}, & 1 \le s < 3/2 \end{cases}$$

with $(H^1(\Gamma_E))^*$ the dual space of $H^1(\Gamma_E)$.

If we use the same notations $\|\cdot\|_{\mathcal{L}}$ and $\|\cdot\|_{\mathcal{M}}$ for vectors in $\mathbb{C}^{P+1}$, the norm in $\boldsymbol{V}_{\Gamma_E}$ can be written as

$$\|\Phi\|^2_{\boldsymbol{V}_{\Gamma_E}} = \sum_{n=0}^{\infty}(\lambda_n^2+1)\|\Phi^n\|^2_{\mathcal{L}} + \|\Phi^n\|^2_{\mathcal{M}}$$

for functions Φ in $\boldsymbol{V}_{\Gamma_E}$ with Fourier series $\Phi = \sum_{n=0}^{\infty}\Phi^n Y_n$.

The solution space $\boldsymbol{V}$ is defined by

$$\boldsymbol{V} := \{(u,\Phi) \in \widetilde{H}^1(\Omega_b)\times \boldsymbol{V}_{\Gamma_E} \ : \ u=\phi_0 \quad \text{on } \Gamma_E \text{ for } \Phi=(\phi_0,\ldots,\phi_P)^t\},$$

which is equipped with the Sobolev norm

$$\|(u,\Phi)\|^2_{\boldsymbol{V}} = \|u\|^2_{H^1(\Omega_b)} + \|\Phi\|^2_{\boldsymbol{V}_{\Gamma_E}}.$$

We note that since $\boldsymbol{V}$ is closed in $H^1(\Omega_b)\times(H^1(\Gamma_E))^{P+1}$, it is a Hilbert space. For regularity estimates, more regular spaces $\boldsymbol{V}^2_{\Gamma_E}$ and $\boldsymbol{V}^2$ are required, where $\boldsymbol{V}^2_{\Gamma_E}$ is the set $(H^2(\Gamma_E))^{P+1}$ with the norm

$$\|\Phi\|^2_{\boldsymbol{V}^2_{\Gamma_E}} := \|\Phi\|^2_{\mathcal{L},2} + \|\Phi\|^2_{\mathcal{M},1} = \sum_{n=0}^{\infty}(\lambda_n^2+1)^2\|\Phi^n\|^2_{\mathcal{L}} + (\lambda_n^2+1)\|\Phi\|^2_{\mathcal{M}}$$

(which is also equivalent to the standard product norm in $(H^2(\Gamma_E))^{P+1}$) and $\boldsymbol{V}^2$ is a subspace of $\boldsymbol{V}$ consisting of (u,Φ) satisfying

$$\|(u,\Phi)\|^2_{\boldsymbol{V}^2} := \|u\|^2_{H^2(\Omega_b)} + \|\Phi\|^2_{\boldsymbol{V}^2_{\Gamma_E}} < \infty.$$

Finally, we introduce the test space $\boldsymbol{V}_0$, the set of functions $(\xi,\Psi)\in\widetilde{H}^1_0(\Omega_b)\times\boldsymbol{V}_{\Gamma_E}$ such that $\xi=\psi_0$ on Γ_E for $\Psi=(\psi_0,\ldots,\psi_P)^t$. Now, we take a test function $(\xi,\Psi)\in\boldsymbol{V}_0$, multiply (3.17) by 2ξ and (3.20) by 2Ψ, and integrate them by parts, which transforms the problem (3.17)–(3.21) to the variational problem of finding $(u,\Phi)\in\boldsymbol{V}$ with $u=f$ on Γ_W such that

$$A((u,\Phi),(\xi,\Psi)) = 0 \tag{4.1}$$

for all $(\xi,\Psi)\in\boldsymbol{V}_0$, where

$$A((u,\Phi),(\xi,\Psi)) = 2(\nabla u,\nabla\xi)_{\Omega_b} - 2k^2(u,\xi)_{\Omega_b} + J(\Phi,\Psi), \tag{4.2}$$

and

$$J(\Phi,\Psi) = (L\nabla_y\Phi,\nabla_y\Psi)_{\Gamma_E} + ((-k^2L+M)\Phi,\Psi)_{\Gamma_E}$$

is the sesquilinear form defined on $\boldsymbol{V}_{\Gamma_E} \times \boldsymbol{V}_{\Gamma_E}$. Also, we define

$$\widetilde{A}((u,\Phi),(\xi,\Psi)) = 2(\nabla u, \nabla\xi)_{\Omega_b} + 2(u,\xi)_{\Omega_b} + \widetilde{J}(\Phi,\Psi)$$

and

$$\widetilde{J}(\Phi,\Psi) = (L\nabla_y\Phi, \nabla_y\Psi)_{\Gamma_E} + (L\Phi,\Psi)_{\Gamma_E} + (\bar{M}\Phi,\Psi)_{\Gamma_E},$$

where $\bar{M}$ is the $(P+1)\times(P+1)$ tridiagonal symmetric matrix whose components are the complex conjugate of those of M.

Lemma 4.2 *For Φ, Ψ in $(L^2(\Gamma_E))^{P+1}$, it holds that*

$$\begin{aligned}
|(L\Phi,\Psi)_{\Gamma_E}| &\le \|\Phi\|_{\mathcal{L}}\|\Psi\|_{\mathcal{L}},\\
|(M\Phi,\Psi)_{\Gamma_E}| &\le \|\Phi\|_{\mathcal{M}}\|\Psi\|_{\mathcal{M}},\\
|(\bar{M}\Phi,\Psi)_{\Gamma_E}| &\le \|\Phi\|_{\mathcal{M}}\|\Psi\|_{\mathcal{M}}.
\end{aligned}$$

Proof Noting the symmetry of the matrix L, application of the Cauchy–Schwarz inequality shows that

$$|(L\Phi,\Psi)_{\Gamma_E}| = \left|\sum_{j=0}^{P}\frac{1}{a_j}(\phi_j+\phi_{j+1},\psi_j+\psi_{j+1})_{\Gamma_E}\right| \le \|\Phi\|_{\mathcal{L}}\|\Psi\|_{\mathcal{L}} \tag{4.3}$$

The other cases are proved similarly. □

The boundedness of J and $\widetilde{J}$ is easily obtained from Lemma 4.2.

Lemma 4.3 *For $\Phi, \Psi \in \boldsymbol{V}_{\Gamma_E}$, it holds that*

$$\begin{aligned}
|J(\Phi,\Psi)| &\le C\|\Phi\|_{\boldsymbol{V}_{\Gamma_E}}\|\Psi\|_{\boldsymbol{V}_{\Gamma_E}},\\
|\widetilde{J}(\Phi,\Psi)| &\le C\|\Phi\|_{\boldsymbol{V}_{\Gamma_E}}\|\Psi\|_{\boldsymbol{V}_{\Gamma_E}}
\end{aligned}$$

with a positive constant C depending only on k.

The following boundedness and coercivity of the sesquilinear form $\widetilde{A}(\cdot,\cdot)$ will play an important role for the existence of solutions in the next section.

Lemma 4.4 *It holds that*

$$|\widetilde{A}((u,\Phi),(\xi,\Psi))| \le C\|(u,\Phi)\|_{\boldsymbol{V}}\|(\xi,\Psi)\|_{\boldsymbol{V}}$$

and

$$|\widetilde{A}((u,\Phi),(u,\Phi))| \ge C\|(u,\Phi)\|_{\boldsymbol{V}}^2$$

for all $(u,\Phi),(\xi,\Psi) \in \boldsymbol{V}$.

Proof The boundedness of $\widetilde{A}(\cdot,\cdot)$ is an immediate consequence of Lemma 4.3 and the Cauchy–Schwarz inequality. For the coercivity, we first examine the real and imaginary parts of $\widetilde{A}((u,\Phi),(u,\Phi))$,

$$\begin{aligned}&\Re(\widetilde{A}((u,\Phi),(u,\Phi)))\\&\quad=2\|u\|^2_{H^1(\Omega_b)}+\sum_{j=n_p}^{n_p+n_e}\left(\frac{1}{a_j}\|\nabla_y(\phi_j+\phi_{j+1})\|^2_{L^2(\Gamma_E)}\right.\\&\qquad\left.+\frac{1}{a_j}\|\phi_j+\phi_{j+1}\|^2_{L^2(\Gamma_E)}+a_j\|\phi_j-\phi_{j+1}\|^2_{L^2(\Gamma_E)}\right)\end{aligned}\tag{4.4}$$

and

$$\begin{aligned}&\Im(\widetilde{A}((u,\Phi),(u,\Phi)))\\&\quad=\sum_{j=0}^{n_p-1}\left(\frac{1}{|a_j|}\|\nabla_y(\phi_j+\phi_{j+1})\|^2_{L^2(\Gamma_E)}\right.\\&\qquad\left.+\frac{1}{|a_j|}\|\phi_j+\phi_{j+1}\|^2_{L^2(\Gamma_E)}+|a_j|\|\phi_j-\phi_{j+1}\|^2_{L^2(\Gamma_E)}\right),\end{aligned}\tag{4.5}$$

and we obtain that

$$\begin{aligned}|\widetilde{A}((u,\Phi),(u,\Phi))|&\geq C(\Re(\widetilde{A}((u,\Phi),(u,\Phi)))+\Im(\widetilde{A}((u,\Phi),(u,\Phi))))\\&=C(\|u\|^2_{H^1(\Omega_b)}+\|\Phi\|^2_{V_{\Gamma_E}}),\end{aligned}$$

which completes the proof. □

We close this section with a lemma about a property of the norms $\|\cdot\|_{\mathcal{L}}$ and $\|\cdot\|_{\mathcal{M}}$, which will be used for the stability analysis of cutoff modes.

Lemma 4.5 *Let a_j be the parameters defined by (3.5) satisfying (3.7). It holds that*

$$\|\Phi\|_{\mathcal{L}}\leq C_a(P+1)\|\Phi\|_{\mathcal{M}}$$

for $\Phi\in\mathbb{C}^{P+1}$, where C_a is a constant depending on $\max_{0\leq j\leq P}\{1/|a_j|\}$.

Proof Noting that

$$\sum_{\ell=0}^{P}|\phi_\ell+\phi_{\ell+1}|^2\leq C(P+1)^2\sum_{\ell=0}^{P}|\phi_\ell-\phi_{\ell+1}|^2$$

for $\Phi=(\phi_0,\ldots,\phi_P)^t\in\mathbb{C}^{P+1}$ with $\phi_{P+1}=0$ (see e.g., [34]), it can be proved that

$$\|\Phi\|^2_{\mathcal{L}}=\sum_{\ell=0}^{P}\frac{1}{|a_\ell|}|\phi_\ell+\phi_{\ell+1}|^2\leq C_a\sum_{\ell=0}^{P}|\phi_\ell+\phi_{\ell+1}|^2$$

$$\leq C_a(P+1)^2 \sum_{\ell=0}^{P} |\phi_\ell - \phi_{\ell+1}|^2 \leq C_a^2 (P+1)^2 \|\Phi\|_{\mathcal{M}}^2.$$

□

5 Existence and uniqueness of solutions to the Helmholtz equation with the CRBCs

This section is devoted to establishing the existence and uniqueness of solutions to the problem (3.17)–(3.21). For establishing the uniqueness of solutions, assume that $f = 0$ on Γ_W and let the solution u be represented by the Fourier series

$$u(x, y) = (A_N + B_N x) Y_N(y) + \sum_{n \neq N} (A_n e^{i\mu_n x} + B_n e^{-i\mu_n x}) Y_n(y). \tag{5.1}$$

The boundary condition on Γ_W implies

$$A_n = 0 \quad \text{for } n = N, \tag{5.2}$$

$$A_n + B_n = 0 \quad \text{for } n \neq N. \tag{5.3}$$

Let C_n^0 and D_n^0 be the Fourier coefficients of the trace of u and $\partial u/\partial x$ on Γ_E, respectively,

$$\begin{aligned} C_n^0 &= \begin{cases} B_n b & \text{for } n = N, \\ A_n e^{i\mu_n b} + B_n e^{-i\mu_n b} & \text{for } n \neq N, \end{cases} \\ D_n^0 &= \begin{cases} B_n & \text{for } n = N, \\ i\mu_n (A_n e^{i\mu_n b} - B_n e^{-i\mu_n b}) & \text{for } n \neq N. \end{cases} \end{aligned} \tag{5.4}$$

The auxiliary variable ϕ_j on Γ_E has the Fourier expansion

$$\phi_j(y) = \sum_{n=0}^{\infty} C_n^j Y_n(y).$$

Now we note that the vector $\boldsymbol{C}_n = (C_n^0, \ldots, C_n^P)^t$ consisting of the nth Fourier coefficients of the auxiliary variables satisfies

$$-2D_n^0 \boldsymbol{e}_0 = (-\mu_n^2 L + M)\boldsymbol{C}_n. \tag{5.5}$$

Indeed, since Y_n is an eigenfunction associated with the eigenvalue λ_n^2, the nth Fourier mode of the right hand side of (3.20) is

$$\frac{-1}{2}(\lambda_n^2 L \boldsymbol{C}_n + (-k^2 L + M)\boldsymbol{C}_n) Y_n = \frac{-1}{2}(-\mu_n^2 L + M)\boldsymbol{C}_n Y_n,$$

while that of the left hand side is $D_n^0 \boldsymbol{e}_0 Y_n$. Applying the inner product $(\cdot,\cdot)_{\mathbb{C}^{P+1}}$ in $\mathbb{C}^{P+1}$ of (5.5) against $\boldsymbol{C}_n$ leads to

$$\begin{aligned} -2D_n^0 \bar{C}_n^0 &= -\mu_n^2 (L\boldsymbol{C}_n, \boldsymbol{C}_n)_{\mathbb{C}^{P+1}} + (M\boldsymbol{C}_n, \boldsymbol{C}_n)_{\mathbb{C}^{P+1}} \\ &= \sum_{j=0}^{P} \left[\frac{-\mu_n^2}{a_j} |C_n^j + C_n^{j+1}|^2 + a_j |C_n^j - C_n^{j+1}|^2 \right], \end{aligned} \tag{5.6}$$

where $\bar{C}_n^j$ is the complex conjugate of C_n^j and $C_n^{P+1} = 0$. Owing to (5.3) and (5.4), the left hand side of (5.6) is given by

$$4\mu_n \Im(A_n \bar{B}_n e^{2i\mu_n b}) - 2\mu_n (|A_n|^2 - |B_n|^2) i = -4\mu_n |A_n|^2 \Im(e^{2i\mu_n b}) \tag{5.7}$$

for $n < N$ (propagating modes, $\mu_n > 0$),

$$2\tilde{\mu}_n (|A_n|^2 e^{-2\tilde{\mu}_n b} - |B_n|^2 e^{2\tilde{\mu}_n b}) + 4\tilde{\mu}_n \Im(A_n \bar{B}_n) i = 2\tilde{\mu}_n |A_n|^2 (e^{-2\tilde{\mu}_n b} - e^{2\tilde{\mu}_n b}) \tag{5.8}$$

for $n > N$ (evanescent modes, $\mu_n^2 < 0$) and

$$-2b|B_N|^2 \tag{5.9}$$

for $n = N$ (cutoff mode, $\mu_n = 0$).

Now, we are ready to prove the uniqueness of solutions.

Lemma 5.1 *Suppose that the parameters a_j are given by (3.5) and k is a positive wavenumber. Then solutions to the problem (3.17)–(3.21) are unique.*

Proof For $n < N$ ($\mu_n^2 > 0$), by (5.6) and (5.7)

$$\begin{aligned} -4\mu_n |A_n|^2 \Im(e^{2i\mu_n b}) = &\sum_{j=0}^{n_p-1} \left[\frac{-\mu_n^2}{-ikc_j} |C_n^j + C_n^{j+1}|^2 - ikc_j |C_n^j - C_n^{j+1}|^2 \right] \\ &+ \sum_{j=n_p}^{n_p+n_e} \left[\frac{-\mu_n^2}{\sigma_j} |C_n^j + C_n^{j+1}|^2 + \sigma_j |C_n^j - C_n^{j+1}|^2 \right]. \end{aligned} \tag{5.10}$$

Comparing the imaginary parts of both sides, we see that

$$C_n^j = 0 \text{ for } j = 0, \ldots, n_p \quad \text{and} \quad n = 0, \ldots, N-1. \tag{5.11}$$

In addition, since $C_n^0 = C_n^1 = 0$, it follows from the zeroth row of (5.5) that $D_n^0 = 0$, which yields that $A_n = B_n = 0$ for $n = 0, \ldots, N-1$ by solving the Eq. (5.4). Then, (5.5) becomes

$$(-\mu_n^2 L + M)\boldsymbol{C}_n = 0. \tag{5.12}$$

Since the superdiagonal entries of $-\mu_n^2 L + M$ below the $(n_p - 1)$th row are non-zero,

$$-\frac{\mu_n^2}{a_j} - a_j = -\frac{\mu_n^2}{\sigma_j} - \sigma_j < 0$$

for $j = n_p, \ldots, n_p + n_e$, applying forward substitution to (5.12) from the n_pth row by using $C_n^j = 0$ for $j = 0, \ldots, n_p$ gives $C_n^j = 0$ for $j = n_p + 1, \ldots, n_p + n_e$.

For $n > N$ ($\mu_n^2 < 0$), (5.10) with (5.8) used instead of (5.7) leads to

$$2\tilde{\mu}_n |A_n|^2 (e^{-2\tilde{\mu}_n b} - e^{2\tilde{\mu}_n b}) = \sum_{j=0}^{n_p-1} \left[\frac{-\mu_n^2}{-ikc_j} |C_n^j + C_n^{j+1}|^2 - ikc_j |C_n^j - C_n^{j+1}|^2 \right] + \sum_{j=n_p}^{n_p+n_e} \left[\frac{-\mu_n^2}{\sigma_j} |C_n^j + C_n^{j+1}|^2 + \sigma_j |C_n^j - C_n^{j+1}|^2 \right].$$

Since the real part of the left hand side is non-positive while that of the right hand side is non-negative, they need to be zero, which implies that $A_n = B_n = 0$ and $C_n^j = 0$ for $j = n_p, \ldots, n_p + n_e$. We observe that $A_n = B_n = 0$ implies $D_n^0 = 0$, and so again from (5.5) obtain the linear equation (5.12) as above. In this case, since the subdiagonal entries of $-\mu_n^2 L + M$ above the $(n_p + 1)$th row are non-zero,

$$\frac{-\mu_n^2}{a_j} - a_j = \frac{-\mu_n^2}{-ikc_j} + ikc_j \neq 0$$

for $j = 0, \ldots, n_p - 1$, we solve (5.12) by backward substitution from the n_pth row by using $C_n^j = 0$ for $j = n_p, \ldots, n_p + n_e$ and then we can see that $C_n^j = 0$ for $j = 0, \ldots, n_p - 1$.

For $n = N$ ($\mu_n^2 = 0$), (5.6) becomes

$$-2b|B_N|^2 = \sum_{j=0}^{n_p-1} -ikc_j |C_n^j - C_n^{j+1}|^2 + \sum_{j=n_p}^{n_p+n_e} \sigma_j |C_n^j - C_n^{j+1}|^2.$$

By comparing the real and imaginary parts of both sides, it can be easily shown that $C_n^j = 0$ for all $j = 0, \ldots, P$. In addition, due to $C_N^0 = B_N b$ and (5.2), we have $A_N = B_N = 0$.

Finally, the fact that $A_n = B_n = 0$ and $C_n^j = 0$ for all $n \geq 0$ and $j = 0, \ldots, P$ results in $u = 0$ in Ω_b and $\phi_j = 0$ on Γ_E for $j = 0, \ldots, P$, which completes the proof of the uniqueness of solutions. □

Theorem 5.2 *The problem* (3.17)–(3.21) *has a unique solution* $(u, \Phi) \in \boldsymbol{V}$.

Proof By invoking Lemma 4.3, we can show boundedness of $A(\cdot,\cdot)$, i.e., there exists a positive constant C_1 such that

$$|A((u,\Phi),(\xi,\Psi))| \le C_1\|(u,\Phi)\|_V\|(\xi,\Psi)\|_V.$$

Furthermore, Lemma 4.3 and Lemma 4.4 show that there exist positive constants C_2 and C_3 such that

$$\begin{aligned} A((u,\Phi),(u,\Phi)) &= \widetilde{A}((u,\Phi),(u,\Phi)) - 2(k^2+1)\|u\|^2_{L^2(\Omega_b)} \\ &\quad - (k^2+1)(L\Phi,\Phi)_{\Gamma_E} + ((M-\bar{M})\Phi,\Phi)_{\Gamma_E} \\ &\ge C_2\|(u,\Phi)\|^2_V - C_3\big(\|u\|^2_{L^2(\Omega_b)} + \|\Phi\|^2_{\mathcal{L}} + \|\Phi\|^2_{\mathcal{M}}\big) \end{aligned} \tag{5.13}$$

for all $(u,\Phi),(\xi,\Psi) \in \boldsymbol{V}_0$. Since $\boldsymbol{V}_0$ is compactly embedded in $L^2(\Omega_b) \times (L^2(\Gamma_E))^{P+1}$, the existence of solutions is a consequence of the Fredholm alternative theorem and the uniqueness of solutions given in Lemma 5.1. □

In the proof, it is not established how the stability constant depends on the number of parameters, $P+1$. This will be studied in more detail in Sect. 8.

Remark 5.3 Let $\mathbf{V}_0^*$ be the dual space of $\mathbf{V}_0$ with the norm

$$\|\mathcal{G}\|_{\mathbf{V}_0^*} = \sup_{0\ne(\xi,\Psi)\in\mathbf{V}_0} \frac{|\mathcal{G}(\xi,\Psi)|}{\|(\xi,\Psi)\|_{\mathbf{V}}}$$

for $\mathcal{G} \in \mathbf{V}_0^*$. The same argument used in the proof of Theorem 5.2 can show that the problem $A((u,\Phi),(\xi,\Psi)) = \mathcal{G}(\xi,\Psi)$ for all $(\xi,\Psi) \in \mathbf{V}_0$ admits a unique solution in $\mathbf{V}_0$.

We can find a formula for the approximate solution u and ϕ_j satisfying the CRBC on Γ_E in terms of a prescribed condition $f \in H^{1/2}(\Gamma_W)$. To this end, let $f \in H^{1/2}(\Gamma_W)$ be a boundary datum, which has a Fourier series

$$f(y) = \sum_{n=0}^{\infty} f_n Y_n(y),$$

and introduce

$$Q^n_{j,m} = \begin{cases} \displaystyle\prod_{\ell=j}^{m} \frac{a_\ell + i\mu_n}{a_\ell - i\mu_n} & \text{for } m \ge j, \\ 1 & \text{for } m < j, \end{cases} \tag{5.14}$$

for $n \ne N$. Now, ϕ_j in the recursions (3.4) are represented by a Fourier series similar to (5.1),

$$\phi_j(x,y) = (A^j_N + B^j_N x)Y_N(y) + \sum_{n\ne N}(A^j_n e^{i\mu_n x} + B^j_n e^{-i\mu_n x})Y_n(y)$$

with $A^0_n = A_n$ and $B^0_n = B_n$.

Non-cutoff modes, $n \neq N$: By (3.4) it is easily shown that

$$\begin{aligned}(a_j - i\mu_n)A_n^{j+1} &= (a_j + i\mu_n)A_n^j,\\ (a_j + i\mu_n)B_n^{j+1} &= (a_j - i\mu_n)B_n^j\end{aligned} \tag{5.15}$$

for all j. If $a_j + i\mu_n \neq 0$ for all $0 \leq j \leq P$, then it holds that

$$A_n^j = Q_{0,j-1}^n A_n \quad \text{and} \quad B_n^j = \frac{1}{Q_{0,j-1}^n} B_n \quad \text{for} \quad 0 \leq j \leq P.$$

The coefficients A_n and B_n of the approximate solution u in (5.1) are determined by the system of linear equations

$$\begin{aligned}A_n + B_n &= f_n,\\ e^{i\mu_n b} Q_{0,P}^n A_n + (e^{i\mu_n b} Q_{0,P}^n)^{-1} B_n &= 0,\end{aligned}$$

from which one can easily see that

$$A_n = \frac{f_n}{1 - (e^{i\mu_n b} Q_{0,P}^n)^2} \quad \text{and} \quad B_n = \frac{-(e^{i\mu_n b} Q_{0,P}^n)^2 f_n}{1 - (e^{i\mu_n b} Q_{0,P}^n)^2}. \tag{5.16}$$

If $a_j + i\mu_n = 0$ for some j, then a similar computation shows that $A_n^j = Q_{0,j-1}^n A_n$ and $B_n^j = 0$ for all j and hence (5.16) is still valid.

Cutoff modes, $n = N$: By the recursive relations (3.4), we observe

$$B_N^j = B_N^{j+1}, \quad B_N^j + a_j A_N^j = -B_N^{j+1} + a_j A_N^{j+1}, \tag{5.17}$$

which implies

$$B_N^j = B_N \quad \text{and} \quad A_N^j = A_N + 2\sum_{\ell=0}^{j-1} \frac{1}{a_\ell} B_N$$

for $j = 1, \ldots, P$. From the boundary condition $A_N^0 = f_N$ and the terminal condition

$$A_N^P + B_N^P b = 0, \tag{5.18}$$

we find

$$A_N = f_N \quad \text{and} \quad B_N = \frac{-f_N}{b + 2\sum_{j=0}^P a_j^{-1}}. \tag{5.19}$$

The formula (5.19) reveals the convergence of cutoff modes provided $\sum_{j=0}^P a_j^{-1} \to \infty$.

We note that better results if a cutoff mode is known to be present could be obtained by changing the termination condition (3.8) to

$$\frac{\partial}{\partial x}\phi_{P+1} = 0, \tag{5.20}$$

since cutoff modes do not have any variation along the axis of the waveguide. In fact, the CRBC terminated by (5.20) yields coefficients A_n and B_n of approximate solutions such that

$$A_n = \frac{f_n}{1 + (e^{i\mu_n b} Q^n_{0,P})^2} \quad \text{and} \quad B_n = \frac{-(e^{i\mu_n b} Q^n_{0,P})^2 f_n}{1 + (e^{i\mu_n b} Q^n_{0,P})^2},$$

which converge to the exact coefficients at the same rate as those of (5.16) by the Dirichlet condition, but $A_N = f_N$ and $B_N = 0$, which coincide with those of the exact solution. However this would change the form of the boundary system and require further analysis. Thus we do not consider it here but refer readers to [23].

Alternatively, we can guarantee rapid convergence independent of the distribution of eigenvalues by using Newman nodes which converge to 0 geometrically, for example Newman's nodes $a_j = -ike^{j/\sqrt{P}}$ for propagating modes and/or their analogous form in the evanescent regime [7,22]. Even though it turns out that with such a choice our bounds on the stability constants degenerate with $e^{\sqrt{P}}$, our experiments, presented in Sect. 10, indicate the discretized problem keeps a convergence rate expected in the continuous level with increasing P as long as the problem is discretized with small mesh size compensating the degenerating stability constants.

6 Convergence of approximate solutions satisfying CRBCs

In this section, we show convergence of approximate solutions satisfying CRBCs. As we have seen above, the error of the cutoff mode is estimated in terms of

$$S_P = |b + 2\sum_{j=0}^{P} a_j^{-1}|^{-1},$$

which approaches zero as the order P increases. For non-cutoff modes the error is controlled by the following factor

$$\left|-e^{i\mu_n b}(Q^n_{0,P})^2\right| = \begin{cases} \displaystyle\prod_{j=0}^{n_p-1} \left|\frac{a_j + i\mu_n}{a_j - i\mu_n}\right|^2 & \text{for } 0 \le n \le N-1, \\ e^{-\tilde{\mu}_n b} \displaystyle\prod_{j=n_p}^{n_p+n_e} \left|\frac{a_j - \tilde{\mu}_n}{a_j + \tilde{\mu}_n}\right|^2 & \text{for } N+1 \le n. \end{cases}$$

Since $\lim_{n\to\infty} |Q^n_{0,P}| = 1$, the error does not decay exponentially as a function of P. However, since the factor $e^{i\mu_n b}$ decays exponentially for large n, we can bound the error almost by an exponential function of P (except for the cutoff mode) in the sense of the following theorem. The optimal choice of parameters would depend on a knowledge of the axial frequencies μ_n, $\tilde{\mu}_n$. Later on we will advocate a simpler approach based only on the knowledge of intervals containing the axial frequencies. We then introduce the min–max problems determining the reflection coefficients for each $n \neq N$,

$$\rho_p = \min_{a_0,\ldots,a_{n_p-1}\in i\mathbb{R}_-} \max_{\mu_{N-1}\le\eta\le k} \prod_{j=0}^{n_p-1} \left|\frac{a_j + i\eta}{a_j - i\eta}\right|^2, \tag{6.1}$$

$$\rho_e = \min_{a_{n_p},\ldots,a_{n_p+n_e}\in\mathbb{R}_+} \max_{\tilde{\mu}_{N+1}\le\tilde{\eta}\le M_\sigma} e^{-\tilde{\eta}b} \prod_{j=n_p}^{n_p+n_e} \left|\frac{a_j - \tilde{\eta}}{a_j + \tilde{\eta}}\right|^2. \tag{6.2}$$

Here we recall that M_σ is determined by $e^{-M_\sigma b}$ less than an error tolerance. It is shown in [30] that the reflection coefficients can be reduced at an exponential rate with respect to the number of parameters used,

$$\begin{aligned} \rho_p &\le e^{-Cn_p/\ln(k/\mu_{N-1})}, \\ \rho_e &\le e^{-\tilde{\mu}_{N+1}b} e^{-Cn_e/\ln(M_\sigma/\tilde{\mu}_{N+1})}. \end{aligned} \tag{6.3}$$

by selecting parameters which satisfy (6.1)–(6.2). These are easy to compute in practice using the Remez algorithm, and in the case of (6.1) they are known analytically (see [7]).

Theorem 6.1 *Suppose that f is in $H^{1/2}(\Gamma_W)$, u^{ex} is the exact radiating solution to the problem (2.1)–(2.3) and u is the solution to the problem (3.17)–(3.21). Then it holds that*

$$\|u - u^{ex}\|_{H^1(\Omega_b)} \le C\rho(M_\sigma, n_p, n_e)\|f\|_{H^{1/2}(\Gamma_W)}, \tag{6.4}$$

where

$$\rho(M_\sigma, n_p, n_e) = \max\{S_P, e^{-Cn_p/\ln(k/\mu_{N-1})}, e^{-\tilde{\mu}_{N+1}b} e^{-Cn_e/\ln(M_\sigma/\tilde{\mu}_{N+1})}, e^{-M_\sigma b}\}.$$

Remark 6.2 We have not attempted to sharply estimate the dependence of the inequality (6.4) on the wave number k, or on the k-dependence of inequalities (8.3), (8.4), (9.3), or (9.4). From the arguments given we can only derive bounds which grow very rapidly with k. Numerical experiments with k as large as 100 show that the actual k-dependence of the stability and error constants is in fact quite mild.

Remark 6.3 Note that the term S_P is absent if no cutoff modes exist. Then we have that with node choices satisfying (6.1)–(6.2) and an error tolerance τ

$$P \propto \ln\left(\frac{1}{\tau}\right) \cdot \ln\left(\ln\left(\frac{1}{\tau}\right)\right) \tag{6.5}$$

suffices.

To prove Theorem 6.1, we start by studying the regularity of solutions satisfying the exact radiation condition given by the Dirichlet-to-Neumann map on the artificial boundary Γ_E. For $0 \le s \le 2$, let $T : H^s(\Gamma_E) \to H^{s-1}(\Gamma_E)$ be the Dirichlet-to-Neumann map defined by

$$Tv = \sum_{n=0}^{\infty} i\mu_n v_n Y_n$$

for $v = \sum_{n=0}^{\infty} v_n Y_n$ in $H^s(\Gamma_E)$. We consider the problem with the exact boundary condition associated with the Dirichlet-to-Neumann map T: For $g_{in} \in H^s(\Omega_b)$ and $g_{bd} \in H^{s+1/2}(\Gamma_E)$ with $-1 \le s \le 0$,

$$\begin{aligned} \Delta u + k^2 u &= g_{in} \quad \text{in} \quad \Omega_b, \\ u &= 0 \quad \text{on } \Gamma_W, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma_T, \\ \frac{\partial u}{\partial x} - Tu &= g_{bd} \quad \text{on } \Gamma_E. \end{aligned} \tag{6.6}$$

As in [3], it can be shown that the regularity of solutions satisfying the exact boundary condition holds by transforming the problem to one without the Dirichlet condition on Γ_W via the odd reflection with respect to Γ_W.

Lemma 6.4 *For $g_{in} \in H^s(\Omega_b)$ and $g_{bd} \in H^{s+1/2}(\Gamma_E)$ with $-1 \le s \le 0$, the problem (6.6) admits a unique solution in $H^{s+2}(\Omega_b)$. Moreover, there exists a positive constant C such that*

$$\|u\|_{H^{s+2}(\Omega_b)} \le C(\|g_{in}\|_{H^s(\Omega_b)} + \|g_{bd}\|_{H^{s+1/2}(\Gamma_E)}).$$

Now, the proof of Theorem 6.1 is as follows.

Proof of Theorem 6.1 We first note that the error function $z = u - u^{ex}$ satisfies

$$\begin{aligned} \Delta z + k^2 z &= 0 \quad \text{in } \Omega_b, \\ z &= 0 \quad \text{on } \Gamma_W, \quad \frac{\partial z}{\partial \nu} = 0 \quad \text{on } \Gamma_T, \\ \frac{\partial z}{\partial x} - Tz &= g_{bd} \quad \text{on } \Gamma_E, \end{aligned}$$

where g_{bd} has the Fourier series

$$g_{bd} = \frac{\partial u}{\partial x} - Tu = \frac{-1}{b + 2\sum_{j=0}^{P} a_j^{-1}} f_N Y_N + \sum_{n \ne N} \frac{2i\mu_n e^{i\mu_n b}(Q_{0,P}^n)^2}{1 - (e^{i\mu_n b} Q_{0,P}^n)^2} f_n Y_n \tag{6.7}$$

by using (5.16) and (5.19).

Let $N_* > N$ be the largest integer such that $\tilde{\mu}_{N_*} \le M_\sigma$. Since $|1 - (e^{i\mu_n b} Q^n_{0,P})^2|$ is bounded away from zero for all $n \ge 0$ and $|\mu_n|^2 \le C(\lambda_n^2 + 1)$ for all $n \ne N$, by (6.3) we obtain

$$\begin{aligned}\|g_{bd}\|^2_{H^{-1/2}(\Gamma_E)} \le C \Bigg(& \sum_{0\le n\le N-1} e^{-2Cn_p/\ln(k/\mu_{N-1})} \frac{|\mu_n f_n|^2}{(1+\lambda_n^2)^{1/2}} \\ & + \sum_{N+1\le n\le N_*} e^{-2\tilde{\mu}_n b} e^{-2Cn_e/\ln(M_\sigma/\tilde{\mu}_{N+1})} \frac{|\mu_n f_n|^2}{(1+\lambda_n^2)^{1/2}} \\ & + \sum_{N_*+1\le n} e^{-2M_\sigma b} \frac{|\mu_n f_n|^2}{(1+\lambda_n^2)^{1/2}} + \frac{1}{|b + 2\sum_{j=0}^P a_j^{-1}|^2} |f_N|^2 \Bigg) \\ & \le C\rho(M_\sigma, n_p, n_e)^2 \|f\|^2_{H^{1/2}(\Gamma_W)}.\end{aligned}$$

Finally, Lemma 6.4 completes the proof of (6.4). □

Remark 6.5 When the parameters a_j are chosen such that

$$a_j = -i\mu_j \text{ for } j = 0, \ldots, N-1 \quad \text{and} \quad a_j = \tilde{\mu}_{j+1} \text{ for } j = N, \ldots, P,$$

the CRBCs behave as the exact boundary conditions for the important $P+1$ modes, which are all propagating modes combined with slowly decaying evanescent modes. These are the modes which would produce the largest reflections without efficient absorbing boundary conditions. Since $Q^n_{0,P} = 0$ for $n = 0, \ldots, P$ and $n \ne N$, the error is estimated as

$$\|u - u^{ex}\|_{H^1(\Omega_b)} \le C(S_P + e^{-\tilde{\mu}_{P+1} b}) \|f\|_{H^{1/2}(\Gamma_W)},$$

where again S_P is absent if there are no cutoff modes.

7 Parameter selection

The general error formulas derived in the preceding section can be used to guide the selection of optimal parameters. Experiments with an automatic parameter selection algorithm will be reported elsewhere; here we will make selections which, though suboptimal, show that the number of parameters will be small even for difficult cases.

Optimal parameters for a fixed P, chosen independent of f and minimizing the error in the Fourier coefficients at $x = b$, would be those which minimize the maximum over $n \ne N$ of

$$\rho \equiv \left| -e^{i\mu_n b} (Q^n_{0,P})^2 \right| = \begin{cases} \left| (Q^n_{0,P})^2 \right| & \text{for } \mu_n^2 > 0, \\ \left| e^{-\tilde{\mu}_n b} (Q^n_{0,P})^2 \right| & \text{for } \mu_n^2 < 0. \end{cases} \tag{7.1}$$

Note that the number of propagating modes is finite, as is the number of evanescent modes satisfying $e^{-\tilde{\mu}_n b} > \tau$ for any error tolerance τ. The remaining evanescent modes are sufficiently small at the boundary, so the value of $\left|(Q^n_{0,P})^2\right| \leq 1$ is unimportant. Moreover, the number of important modes increases with increasing k; for k small it is feasible to directly compute this small number of modes and choose parameters which are exact on these modes. (For a discussion of conditions using a different set of auxiliary variables which are exact for propagating modes, see Bendali and Guillaume [3].)

Here we look at the simpler problem of minimizing ρ over an entire interval rather than over a discrete set. We introduce the following scalings:

$$\eta \equiv \mu/k \ \ (\tilde{\eta} \equiv \tilde{\mu}/k), \ \ \tilde{a}_j \equiv a_j/k, \ \ b = 2\pi k^{-1} n_\lambda,$$

where n_λ is the number of wavelengths of the normally propagating mode, e^{ikx}, on the interval $[0, b]$. Now, we explicitly assume that $\mu_n \neq 0$; that is the cutoff mode is absent. To perform the optimizations we quantify the gap in the spectrum near 0

$$\eta^2 \geq c_0^2 \quad \text{and} \quad \tilde{\eta}^2 \geq g_0^2 \tag{7.2}$$

for some constants c_0 and g_0. In real situations, c_0 and g_0 would be some constants approximate to the smallest axial frequency, μ_{N-1}, of propagating modes and smallest decay rate, $\tilde{\mu}_{N+1}$, of evanescent modes, respectively. We then consider the reflection coefficients

$$\rho_p = \max_{c_0 \leq \eta \leq 1} \prod_{j=0}^{n_p - 1} \left| \frac{\tilde{a}_j + i\eta}{\tilde{a}_j - i\eta} \right|^2, \tag{7.3}$$

$$\rho_e = \max_{\tilde{\eta} \geq g_0} e^{-2\pi n_\lambda \tilde{\eta}} \prod_{j=n_p}^{n_p + n_e} \left| \frac{\tilde{a}_j - \tilde{\eta}}{\tilde{a}_j + \tilde{\eta}} \right|^2. \tag{7.4}$$

For fixed values of n_p and n_e, we can compute optimal parameters using the Remez algorithm (see, e.g., [30]). For instance, consider the truncated waveguide Ω_b defined with $W = 1, b = 0.1$. When the wavenumber is $k = 100$, there are 32 propagating modes involved in acoustic pressure fields. For $n_p = 4$ and $n_e = 3$, the Remez algorithm applied to minimization of the maximal reflection coefficients (7.3) and (7.4) produces the damping parameters with which the graphs of the reflection coefficients as a function of n are presented in Fig. 3. It indicates that reflection of all propagating modes and evanescent modes can be reduced up to 3.9590×10^{-6} and 5.3492×10^{-5}, respectively. Here the upper bound for $\tilde{\eta}$ in the Remez algorithm is determined in a way that the modes between the vertical green lines damped effectively. Note that our simple Matlab implementation of the Remez algorithm, which uses a geometrical sequence as an initial guess, has converged rapidly for all the cases considered here. The authors will provide it to any interested readers.

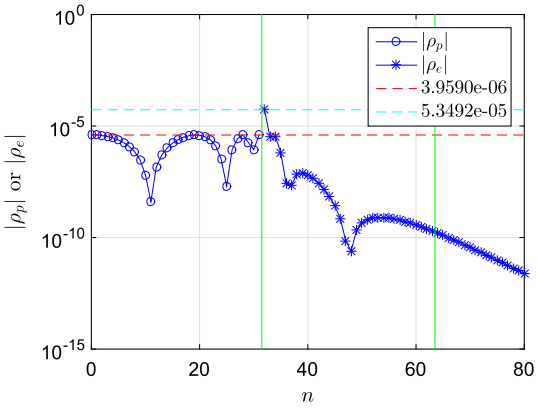


Fig. 3 Reflection coefficients of $|\rho_p|$ and $|\rho_e|$ as a function of n with the optimal parameters obtained by Remez algorithm

To determine the smallest P for a given tolerance, τ, as a function of c_0, g_0 and n_λ we simply find the smallest values of n_p and n_e such that the optimal nodes chosen by the Remez algorithm lead to $\rho_p \leq \tau$, $\rho_e \leq \tau$.

Note that these approximations can be directly related to optimal approximation of the square root function, which was solved by Zolotarev using elliptic functions [30]. The error estimates developed in [7,22] state the error in the Zolotarev approximation of degree $(d-1, d)$ on the interval $[z_0, z_1]$ to be of the order $e^{-\pi^2 d/\ln(z_1/z_0)}$. For propagating modes this implies

$$n_p \propto \ln\left(\frac{1}{\tau}\right) \cdot \ln\left(\frac{1}{c_0}\right). \tag{7.5}$$

For evanescent modes we note that the largest value of $\tilde{\eta}$ is relevant scales like $n_\lambda^{-1} \ln\left(\frac{1}{\tau}\right)$. Thus we conclude that

$$n_e \propto \ln\left(\frac{1}{\tau}\right) \cdot \ln\left(\frac{1}{n_\lambda g_0}\right) + \ln\left(\frac{1}{\tau}\right) \cdot \ln\ln\left(\frac{1}{\tau}\right). \tag{7.6}$$

We carried out the optimizations discussed above for the parameters

$$c_0 = \{10^{-2}, 10^{-4}\}, g_0 = \{10^{-2}, 10^{-4}\}, n_\lambda = \{1, 0.1\}, \tau = \{10^{-3}, 10^{-5}\}.$$

The results are shown in Table 1. Based on the Remez algorithm the results are consistent with the estimates (7.5)–(7.6). We emphasize that these results are definitely suboptimal as they do not take account of the actual modal distributions. Methods for constructing better parameters may be based, for example, on rational Krylov algorithms [8,14,28] applied to the finite element discretization of the cross-sectional Laplace operator.

In practice, then, we recommend the following procedure to select the method parameters. Given a choice of b, which can be taken as the separation between the radiation boundary and any sources, scatterers, or inhomogeneities, and an error tolerance, τ:

Table 1 Number of terms needed to meet the tolerance, τ, for select values of c_0, g_0, and n_λ

c_0	g_0	n_λ	τ	P_{opt} with (3.5)
10^{-2}	10^{-2}	1	10^{-3}	10
10^{-2}	10^{-2}	1	10^{-5}	16
10^{-2}	10^{-2}	0.1	10^{-3}	12
10^{-2}	10^{-2}	0.1	10^{-5}	18
10^{-2}	10^{-4}	1	10^{-3}	14
10^{-2}	10^{-4}	1	10^{-5}	21
10^{-2}	10^{-4}	0.1	10^{-3}	16
10^{-2}	10^{-4}	0.1	10^{-5}	24
10^{-4}	10^{-2}	1	10^{-3}	14
10^{-4}	10^{-2}	1	10^{-5}	22
10^{-4}	10^{-2}	0.1	10^{-3}	16
10^{-4}	10^{-2}	0.1	10^{-5}	25
10^{-4}	10^{-4}	1	10^{-3}	18
10^{-4}	10^{-4}	1	10^{-5}	28
10^{-4}	10^{-4}	0.1	10^{-3}	20
10^{-4}	10^{-4}	0.1	10^{-5}	31

i. If possible estimate the number of important modes; in many cases this can be done based on the frequency, k, and the geometry of the cross-section using standard inequalities on the spectrum of elliptic operators [6]. If this is small enough, for propagating modes, evanescent modes, or both, application of a Lanczos algorithm [32] will produce them at minimal cost. Then choose the parameters to exactly absorb these modes.
ii. If the use of exact conditions is deemed inefficient, again for propagating modes, evanescent modes, or both, use the Lanczos algorithm to compute the eigenvalues nearest k^2 and use that information to define the intervals for input into the Remez algorithm.

8 Stability and regularity of the variational problem

In this section, we study the stability and regularity of the variational problems

$$A((u, \Phi), (\xi, \Psi)) = (f_s, \xi)_{\Gamma_E} \tag{8.1}$$

for all $(\xi, \Psi) \in \boldsymbol{V}_0$ with $f_s \in L^2(\Omega_b)$ supported away from Γ_E, and

$$A((u, \Phi), (\xi, \Psi)) = (L\Upsilon, \Psi)_{\Gamma_E} \tag{8.2}$$

for all $(\xi, \Psi) \in \boldsymbol{V}_0$ with the source $L\Upsilon$, $\Upsilon \in (L^2(\Gamma_E))^{P+1}$ being given as auxiliary variables. The study of the problem (8.1) suffices for verification of the stability and

regularity of solutions to the problem (4.1) since the boundary value problem can be reduced to the source problem due to a lifting of the boundary condition. Also, these results will come into play in the finite element analysis.

8.1 Stability and regularity of solutions to Problem (8.1)

We note that the problem (8.1) has a unique solution in $\boldsymbol{V}_0$ by Remark 5.3. The energy norm estimates for the solution u and the auxiliary variables Φ are given in the following theorem.

Theorem 8.1 *Let a_j be the parameters defined by (3.5) satisfying (3.6). Then for any $f_s \in L^2(\Omega_b)$ supported away from Γ_E, the solution (u, Φ) to the problem (8.1) satisfies*

$$\|u\|_{H^1(\Omega_b)} \leq C\|f_s\|_{L^2(\Omega_b)}$$

and

$$\|\Phi\|_{\boldsymbol{V}_{\Gamma_E}} \leq C_a(P+1)\|f_s\|_{L^2(\Omega_b)}.$$

In addition, the regularity result holds,

$$\|u\|_{H^2(\Omega_b)} \leq C\|f_s\|_{L^2(\Omega_b)} \tag{8.3}$$

and

$$\|\Phi\|_{\boldsymbol{V}^2_{\Gamma_E}} \leq C_a(P+1)\|f_s\|_{L^2(\Omega_b)}. \tag{8.4}$$

If cutoff modes are excluded, the constants C_a for the stability and regularity estimates are independent of a_j and the exponents on $(P+1)$ are halved; that is the constants in the estimates of Φ become $C(P+1)^{1/2}$.

The proof of Theorem 8.1 proceeds based on a sequence of lemmas for solution formulas of auxiliary variables. In order to study the stability estimate of problem (8.1), it is required to analyze the auxiliary variables solving the problem

$$\begin{aligned} -L\Delta_y\Phi + (-k^2L + M)\Phi &= E_0\boldsymbol{e}_0 \quad \text{in } \Gamma_E, \\ \frac{\partial\Phi}{\partial\nu} &= 0 \quad \text{on } \partial\Gamma_E. \end{aligned} \tag{8.5}$$

The nth Fourier coefficients Φ^n of Φ satisfy the equation

$$-\mu_n^2 L\Phi^n + M\Phi^n = E_0^n\boldsymbol{e}_0. \tag{8.6}$$

We start by finding the explicit form of the solution Φ^n with $E_j^n\boldsymbol{e}_j$ for $j = 0, \ldots, P$ on the right hand side of (8.6), recalling the definition (5.14) of $Q_{j,m}^n$ for $n \neq N$.

Lemma 8.2 *Suppose that $a_j \neq -i\mu_n$ and μ_n is not a cutoff axial frequency, i.e., $\mu_n \neq 0$. Let $\Phi^n \in \mathbb{C}^{P+1}$ be a solution to the linear system (8.6) with $E_j^n \boldsymbol{e}_j$ on the right hand side. Then ϕ_ℓ^n is given by the formula $\phi_\ell^n = s_{\ell,j}^n E_j^n$, where*

$$s_{\ell,j}^n = \begin{cases} \dfrac{(1+(Q_{0,\ell-1}^n)^2)Q_{\ell,j-1}^n(1-(Q_{j,P}^n)^2)}{-4i\mu_n(1+(Q_{0,P}^n)^2)} & \text{if } \ell \leq j, \\ \dfrac{(1+(Q_{0,j-1}^n)^2)Q_{j,\ell-1}^n(1-(Q_{\ell,P}^n)^2)}{-4i\mu_n(1+(Q_{0,P}^n)^2)} & \text{if } \ell \geq j. \end{cases} \tag{8.7}$$

Proof We will find the solution Φ^n in the form

$$\phi_\ell^n = \begin{cases} Q_{0,\ell-1}^n \widetilde{A}_n + \dfrac{1}{Q_{0,\ell-1}^n}\widetilde{B}_n & \text{for } \ell = 0, 1, \ldots, j, \\ Q_{j,\ell-1}^n \widetilde{C}_n + \dfrac{1}{Q_{j,\ell-1}^n}\widetilde{D}_n & \text{for } \ell = j, j+1, \ldots, P \end{cases} \tag{8.8}$$

for $0 < j < P$. When $j = 0$ or P, we assume that ϕ_ℓ^n is defined by the upper formula with $\ell = 0, 1, \ldots, P$. Here we will verify the formulas for $0 < j < P$, as the other cases can be treated with only small modifications.

By the definition of $Q_{j,m}^n$ one can easily show that the three term recursions

$$\begin{aligned} &(-\mu_n^2 L_{\ell,\ell-1} + M_{\ell,\ell-1})\phi_{\ell-1}^n + (-\mu_n^2 L_{\ell,\ell} + M_{\ell,\ell})\phi_\ell^n \\ &\quad +(-\mu_n^2 L_{\ell,\ell+1} + M_{\ell,\ell+1})\phi_{\ell+1}^n = 0 \end{aligned}$$

hold for $\ell \neq 0,\ j,\ P$. Thus, the four unknowns $\widetilde{A}_n$, $\widetilde{B}_n$, $\widetilde{C}_n$ and $\widetilde{D}_n$ are to be determined by

$$-2i\mu_n(\widetilde{A}_n - \widetilde{B}_n) = 0 \tag{8.9}$$

from the 0th equation,

$$Q_{0,\ell-1}^n \widetilde{A}_n + \frac{1}{Q_{0,\ell-1}^n}\widetilde{B}_n = \widetilde{C}_n + \widetilde{D}_n \tag{8.10}$$

from the definition of ϕ_ℓ^n with $\ell = j$,

$$\left(Q_{0,\ell-1}^n \widetilde{A}_n - \frac{1}{Q_{0,\ell-1}^n}\widetilde{B}_n\right) - (\widetilde{C}_n - \widetilde{D}_n) = \frac{1}{2i\mu_n}E_j^n \tag{8.11}$$

from the jth equation and

$$Q_{j,P}^n \widetilde{C}_n + \frac{1}{Q_{j,P}^n}\widetilde{D}_n = 0 \tag{8.12}$$

from the Pth equation. Solving the Eqs. (8.9)–(8.12) leads to

$$\widetilde{A}_n = \frac{(1-(Q^n_{j,P})^2)Q^n_{0,j-1}}{-4i\mu_n(1+(Q^n_{0,P})^2)}E^n_j, \qquad \widetilde{B}_n = \frac{(1-(Q^n_{j,P})^2)Q^n_{0,j-1}}{-4i\mu_n(1+(Q^n_{0,P})^2)}E^n_j,$$
$$\widetilde{C}_n = \frac{(1+(Q^n_{0,j-1})^2)}{-4i\mu_n(1+(Q^n_{0,P})^2)}E^n_j, \qquad \widetilde{D}_n = \frac{(1+(Q^n_{0,j-1})^2)(Q^n_{j,P})^2}{4i\mu_n(1+(Q^n_{0,P})^2)}E^n_j$$

and hence the formula (8.7) is obtained. □

The next lemma gives solution formulas when there exists an index J such that $a_J + i\mu_n = 0$. In this case the problem can be written as two block systems. The first block system is reduced to the case in Lemma 8.2, and the formulas for the second one can be derived by a similar computation to that used in Lemma 8.2 and hence we omit the proof.

Lemma 8.3 *Suppose that there exists an index J such that $a_J + i\mu_n = 0$. Let $\Phi^n \in \mathbb{C}^{P+1}$ be a solution to the linear system (8.6) with $E^n_j \boldsymbol{e}_j$ in the right hand side. Then ϕ^n_ℓ are given by the formula $\phi^n_\ell = s^n_{\ell,j}E^n_j$, where if $j \le J$*

$$s^n_{\ell,j} = \begin{cases} \dfrac{-1}{4i\mu_n}(1+(Q^n_{0,\ell-1})^2)Q^n_{\ell,j-1} & \text{if } \ell \le j, \\ \dfrac{-1}{4i\mu_n}(1+(Q^n_{0,j-1})^2)Q^n_{j,\ell-1} & \text{if } \ell \ge j \end{cases} \tag{8.13}$$

and if $j > J$

$$s^n_{\ell,j} = \begin{cases} \dfrac{-1}{4i\mu_n}Q^n_{\ell,j-1}(1-(Q^n_{j,P})^2) & \text{if } \ell \le j, \\ \dfrac{-1}{4i\mu_n}Q^n_{j,\ell-1}(1-(Q^n_{\ell,P})^2) & \text{if } \ell \ge j. \end{cases} \tag{8.14}$$

We notice that these formulas in Lemma 8.3 are consistent with (8.7) since $Q^n_{c,d} = 0$ for $c \le J \le d$.

As a special case the solution to (8.6) is given in the following lemma.

Lemma 8.4 *Let $\Phi^n \in \mathbb{C}^{P+1}$ be a solution to the linear system (8.6). For $n \ne N$, the ϕ^n_ℓ are given by*

$$\phi^n_\ell = \frac{-Q^n_{0,\ell-1}(1-(Q^n_{\ell,P})^2)}{2i\mu_n(1+(Q^n_{0,P})^2)}E^n_0 \tag{8.15}$$

and

$$\phi^n_\ell = \frac{Q^n_{0,\ell-1}(1-(Q^n_{\ell,P})^2)}{(1-(Q^n_{0,P})^2)}\phi^n_0 \tag{8.16}$$

for $\ell = 0, \ldots, P$. For $n = N$,

$$\phi^n_\ell = \sum_{j=\ell}^{P}\frac{1}{a_j}E^n_0. \tag{8.17}$$

Proof When $a_j + i\mu_n \neq 0$ for all j, the formula (8.15) is obtained from (8.7) with $j = 0$. If there exists J such that $a_J + i\mu_n = 0$, (8.15) immediately follows from (8.13) and noting that $Q^n_{\ell,P} = 0$ for $\ell \leq J$ and $Q^n_{0,\ell-1} = 0$ for $\ell \geq J+1$. In addition, we have (8.16) by rewriting E^n_0 in terms of ϕ^n_0.

The formula (8.17) for $n = N$ is obtained straightforwardly by Gaussian elimination. □

We note that by the arithmetic-geometric mean inequality

$$\begin{aligned} \frac{1}{\sqrt{|a_\ell|}}|1 + Q^n_{\ell,\ell}| &= \frac{2\sqrt{|a_\ell\mu_n|}}{|a_\ell - i\mu_n|}\frac{1}{\sqrt{|\mu_n|}} \leq \frac{C}{\sqrt{|\mu_n|}}, \\ \sqrt{|a_\ell|}|1 - Q^n_{\ell,\ell}| &= \frac{2\sqrt{|a_\ell\mu_n|}}{|a_\ell - i\mu_n|}\sqrt{|\mu_n|} \leq C\sqrt{|\mu_n|} \end{aligned} \tag{8.18}$$

and $(\lambda_n^2 + 1) \leq C|\mu_n|^2$ for $n \neq N$.

Lemma 8.5 *Let a_j be the parameters defined by (3.5) satisfying (3.6). We assume that $\Phi \in \boldsymbol{V}_{\Gamma_E}$, $\phi_0 \in H^{s+1/2}(\Gamma_E)$ and $E_0 \in H^{s-1/2}(\Gamma_E)$ for $s \geq 0$. If Φ and E_0 satisfy (8.5), then it holds that*

$$\|\Phi\|_{\boldsymbol{V}_{\Gamma_E}} \leq C_a(P+1)(\|E_0\|_{H^{-1/2}(\Gamma_E)} + \|\phi_0\|_{H^{1/2}(\Gamma_E)}) \text{ for } s = 0.$$

In addition, we have the regularity estimate

$$\|\Phi\|_{\boldsymbol{V}^2_{\Gamma_E}} \leq C_a(P+1)(\|E_0\|_{H^{1/2}(\Gamma_E)} + \|\phi_0\|_{H^{3/2}(\Gamma_E)}) \text{ for } s = 1.$$

If cutoff modes are excluded, the constants C_a for the stability and regularity estimates are independent of a_j and the exponents on $(P+1)$ are halved; that is the constants in the estimates of Φ become $C(P+1)^{1/2}$.

Proof Cutoff modes, $n = N$: By using the solution formula (8.17), we have

$$\begin{aligned} \|\Phi^N\|^2_{\mathcal{M}} &= \sum_{\ell=0}^{P} |a_\ell||\phi^N_\ell - \phi^N_{\ell+1}|^2 = \sum_{\ell=0}^{P}\frac{1}{|a_\ell|}|E^N_0|^2 = |E^N_0|\sum_{\ell=0}^{P}\frac{1}{|a_\ell|}|E^N_0| \\ &\leq \sqrt{2}|E^N_0||\phi^N_0| = \frac{\sqrt{2}}{|\sum_{\ell=0}^{P} a_\ell^{-1}|}|\phi^N_0|^2 \leq C|\phi^N_0|^2. \end{aligned} \tag{8.19}$$

Here we used (8.17) with $\ell = 0$ for the first inequality. Also, invoking Lemma 4.5 and (8.19), we are led to

$$\|\Phi^N\|^2_{\mathcal{L}} \leq C_a^2(P+1)^2\|\Phi^N\|^2_{\mathcal{M}} \leq C_a^2(P+1)^2|\phi^N_0|^2.$$

Thus, since $\lambda_N = k$ is a constant, we have

$$(\lambda_N^2+1)^s((\lambda_N^2+1)\|\Phi^N\|^2_{\mathcal{L}} + \|\Phi^N\|^2_{\mathcal{M}}) \leq C_a^2(P+1)^2(\lambda_N^2+1)^{s+1/2}|\phi^N_0|^2. \tag{8.20}$$

Non-cutoff modes, $n \neq N$: For the estimation of non-cutoff modes, we decompose $\mathbb{N}\backslash\{N\}$ into two disjoint sets $\mathcal{N}_1$ and $\mathcal{N}_2$,

$$\mathcal{N}_1 = \{n \in \mathbb{N}\backslash\{N\} \ : \ |1 + (Q^n_{0,P})^2| \geq 1\} \quad \text{and} \quad \mathcal{N}_2 = \mathbb{N}\backslash(\mathcal{N}_1 \cup \{N\}).$$

Since $|1 + (Q^n_{0,P})^2| \geq 1$ or $|1 - (Q^n_{0,P})^2| \geq 1$ for each $n \geq 0$, if $n \in \mathcal{N}_2$, then $|1 - (Q^n_{0,P})^2| \geq 1$. Therefore, for $n \in \mathcal{N}_1$ the solution formula (8.15) implies

$$|\phi^n_\ell + \phi^n_{\ell+1}| = \left| \frac{Q^n_{\ell-1}(1 - (Q^n_{\ell+1,P})^2 Q^n_{\ell,\ell})}{(1 + (Q^n_{0,P})^2)} \frac{(1 + Q^n_{\ell,\ell})E^n_0}{2i\mu_n} \right| \leq C \left| \frac{(1 + Q^n_{\ell,\ell})E^n_0}{2i\mu_n} \right|,$$

and by (8.18) we have

$$\frac{1}{|a_\ell|}|\phi^n_\ell + \phi^n_{\ell+1}|^2 \leq C \frac{|E^n_0|^2}{|\mu_n|^3}. \tag{8.21}$$

A similar computation yields that

$$\begin{aligned} |a_\ell||\phi^n_\ell - \phi^n_{\ell+1}|^2 &= |a_\ell| \left| \frac{Q^n_{\ell-1}(1 + (Q^n_{\ell+1,P})^2 Q^n_{\ell,\ell})}{(1 + (Q^n_{0,P})^2)} \frac{(1 - Q^n_{\ell,\ell})E^n_0}{2i\mu_n} \right|^2 \\ &\leq C|a_\ell||1 - Q^n_{\ell,\ell}|^2 \frac{|E^n_0|^2}{|\mu_n|^2} \leq \frac{C}{|\mu_n|}|E^n_0|^2. \end{aligned} \tag{8.22}$$

Combining (8.21) and (8.22) yields

$$\begin{aligned} (\lambda_n^2 + 1)^s((\lambda_n^2 + 1)\|\Phi^n\|^2_{\mathcal{L}} + \|\Phi^n\|^2_{\mathcal{M}}) &\leq C(P + 1)\left(\frac{(\lambda_n^2 + 1)^{s+1}}{|\mu_n|^3} + \frac{(\lambda_n^2 + 1)^s}{|\mu_n|} \right)|E^n_0|^2 \\ &\leq C(P + 1)(\lambda_n^2 + 1)^{s-1/2}|E^n_0|^2. \end{aligned} \tag{8.23}$$

On the other hand, the same calculation as above but using (8.16) instead of (8.15) shows that for $n \in \mathcal{N}_2$

$$\begin{aligned} \frac{1}{|a_\ell|}|\phi^n_\ell + \phi^n_{\ell+1}|^2 &\leq \frac{C}{|\mu_n|}|\phi^n_0|^2, \\ |a_\ell||\phi^n_\ell - \phi^n_{\ell+1}|^2 &\leq C|\mu_n||\phi^n_0|^2, \end{aligned}$$

from which it follows that

$$(\lambda_n^2 + 1)^s((\lambda_n^2 + 1)\|\Phi^n\|^2_{\mathcal{L}} + \|\Phi^n\|^2_{\mathcal{M}}) \leq C(P + 1)(\lambda_n^2 + 1)^{s+1/2}|\phi^n_0|^2. \tag{8.24}$$

Finally, we obtain the stability and regularity estimates by using (8.20), (8.23) and (8.24) for $s = 0$ and $s = 1$, respectively. □

Proof of Theorem 8.1 It suffices to prove the regularity estimates (8.3) and (8.4). Let u^{ex} be the solution to the problem (6.6) with $g_{in} = f_s$ and $g_{bd} = 0$ satisfying

$$\|u^{ex}\|_{H^2(\Omega_b)} \leq C\|f_s\|_{L^2(\Omega_b)} \tag{8.25}$$

by Lemma 6.4. Also, by u we denote the solution satisfying CRBCs, i.e.

$$\begin{aligned} \Delta u + k^2 u &= f_s \quad \text{in } \Omega_b, \\ u &= 0 \quad \text{on } \Gamma_W, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma_T, \\ B_P(u) &= 0 \quad \text{on } \Gamma_E, \end{aligned} \tag{8.26}$$

where $B_P(u) = \phi_{P+1}$ is the trace of the $(P+1)$th auxiliary variable ϕ_{P+1} on Γ_E. Since u^{ex} is expressed as $u^{ex} = \sum_{n=0}^{\infty} A_n^{ex} e^{i\mu_n x} Y_n$ beyond the support of f_s, the error function $z = u - u^{ex}$ satisfies

$$B_P(z) = B_P(-u^{ex}) = -A_N^{ex} Y_N - \sum_{n \neq N} Q_{0,P}^n A_n^{ex} e^{i\mu_n b} Y_n.$$

Assume that z is written as $z = (A_N + B_N x)Y_N + \sum_{n\neq N}(A_n e^{i\mu_n x} + B_n e^{-i\mu_n x})Y_n$. If there exists an index J such that $a_J + i\mu_n = 0$ for some n, then the error does not include the corresponding mode, i.e., $A_n = A_n^{ex}$ and $B_n = 0$. Otherwise, the boundary conditions on Γ_E and Γ_W lead to the linear problem for A_n and B_n,

$$\begin{aligned} A_n + B_n &= 0, \\ Q_{0,P}^n e^{i\mu_n b} A_n + \frac{1}{Q_{0,P}^n e^{i\mu_n b}} B_n &= -Q_{0,P}^n e^{i\mu_n b} A_n^{ex}, \end{aligned}$$

for $n \neq N$ and

$$A_N = 0 \quad \text{and} \quad A_N + B_N \left(b + 2\sum_{j=0}^{P} a_j^{-1} \right) = -A_N^{ex}$$

for $n = N$. It then follows that

$$A_n = \frac{(Q_{0,P}^n e^{i\mu_n b})^2}{1 - (Q_{0,P}^n e^{i\mu_n b})^2} A_n^{ex} \quad \text{and} \quad B_n = \frac{-(Q_{0,P}^n e^{i\mu_n b})^2}{1 - (Q_{0,P}^n e^{i\mu_n b})^2} A_n^{ex} \tag{8.27}$$

for $n \neq N$ and

$$A_N = 0 \quad \text{and} \quad B_N = \frac{-A_N^{ex}}{b + 2\sum_{j=0}^{P} a_j^{-1}}$$

for $n = N$.

Then z solves the problem (6.6) with $g_{in} = 0$ and

$$g_{bd} = \frac{\partial z}{\partial x} - T(z) = \frac{-1}{b + 2\sum_{j=0}^{P} a_j^{-1}} A_N^{ex} Y_N + \sum_{n \neq N} \frac{2i\mu_n e^{i\mu_n b} (Q_{0,P}^n)^2}{1 - (e^{i\mu_n b} Q_{0,P}^n)^2} A_n^{ex} Y_n.$$

Here, we note that g_{bd} is in $H^{1/2}(\Gamma_E)$. Indeed, from the boundedness of the coefficients

$$\frac{1}{b + 2\sum_{j=0}^{P} a_j^{-1}} \quad \text{and} \quad \frac{2i(Q_{0,P}^n)^2}{1 - (e^{i\mu_n b} Q_{0,P}^n)^2}$$

of g_{bd}, a trace theorem and (8.25), it follows that

$$\begin{aligned} \|g_{bd}\|^2_{H^{1/2}(\Gamma_E)} &\leq C \left[(\lambda_N^2 + 1)^{1/2} |A_N^{ex}|^2 + \sum_{n \neq N} (\lambda_n^2 + 1)^{1/2} |\mu_n|^2 |e^{i\mu_n b} A_n^{ex}|^2 \right] \\ &\leq C \|u^{ex}\|^2_{H^{3/2}(\Gamma_E)} \leq C \|f_s\|^2_{L^2(\Omega_b)}. \end{aligned}$$

Therefore, Lemma 6.4 reveals that

$$\|u - u^{ex}\|_{H^2(\Omega_b)} \leq C \|g_{bd}\|_{H^{1/2}(\Gamma_E)} \leq C \|f_s\|_{L^2(\Omega_b)},$$

which, in turn, results in (8.3)

$$\|u\|_{H^2(\Omega_b)} \leq \|z\|_{H^2(\Omega_b)} + \|u^{ex}\|_{H^2(\Omega_b)} \leq C \|f_s\|_{L^2(\Omega_b)}.$$

In addition, a trace inequality yields that

$$\|u\|_{H^{3/2}(\Gamma_E)} \text{ and } \|\frac{\partial u}{\partial x}\|_{H^{1/2}(\Gamma_E)} \leq C \|f_s\|_{L^2(\Omega_b)}, \tag{8.28}$$

and hence Lemma 8.5 with $\phi_0 = u$ and $E_0 = -2\partial u/\partial x$ on Γ_E shows (8.4). □

8.2 Regularity of solutions to Problem (8.2)

It is clear that the solution (u, Φ) to the problem (8.2) solves

$$\begin{aligned} \left(-2\frac{\partial u}{\partial x}\right) \boldsymbol{e}_0 &= -L\Delta_y \Phi + (-k^2 L + M)\Phi - \Xi \quad \text{in } \Gamma_E, \\ \frac{\partial \Phi}{\partial \nu} &= 0 \quad \text{on } \partial\Gamma_E, \end{aligned} \tag{8.29}$$

where $\Xi = L\Upsilon$.

As done in the previous subsection, we will derive explicit formulas for the solution. We know that the solution has the series representation

$$u(x,y) = (A_N + B_N x)Y_N(y) + \sum_{n\neq N}(A_n e^{i\mu_n x} + B_n e^{-i\mu_n x})Y_n(y) \tag{8.30}$$

and the linear systems for the nth Fourier coefficients

$$2i\mu_n(A_n e^{i\mu_n b} - B_n e^{-i\mu_n b})\boldsymbol{e}_0 - \mu_n^2 L\Phi^n + M\Phi^n = \Xi^n \tag{8.31}$$

for $n \neq N$ and

$$2B_n\boldsymbol{e}_0 + M\Phi^n = \Xi^n \tag{8.32}$$

for $n = N$ hold with Ξ^n being the nth Fourier coefficients of Ξ. In case of $n \neq N$, it suffices to derive the formula when $a_j \neq -i\mu_n$. Otherwise the system matrix can be written as a 2×2 block diagonal matrix and solutions of the lower block are given by the same formulas as (8.14) in Lemma 8.3.

Lemma 8.6 *Suppose that $a_j \neq -i\mu_n$ and μ_n is not a cutoff axial frequency, i.e. $\mu_n \neq 0$. Then for $\Xi = E_j\boldsymbol{e}_j = \sum_{n=0}^{\infty} E_j^n Y_n \boldsymbol{e}_j$, there exists a unique solution to the problem (8.31) given by the formula, $\phi_\ell^n = t_{\ell,j}^n E_j^n$, where*

$$t_{\ell,j}^n = \begin{cases} \dfrac{(1-e^{2i\mu_n b}(Q_{0,\ell-1}^n)^2)Q_{\ell,j-1}^n(1-(Q_{j,P}^n)^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q_{0,P}^n)^2)} & \text{if } \ell \leq j, \\ \dfrac{(1-e^{2i\mu_n b}(Q_{0,j-1}^n)^2)Q_{j,\ell-1}^n(1-(Q_{\ell,P}^n)^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q_{0,P}^n)^2)} & \text{if } \ell \geq j. \end{cases} \tag{8.33}$$

Also, the normal derivative of the nth Fourier mode u_n on Γ_E satisfies

$$\frac{\partial u_n}{\partial x} = \frac{(1+e^{2i\mu_n b})Q_{0,j-1}^n(1-(Q_{j,p}^n)^2)}{4(1-e^{2i\mu_n b}(Q_{0,P}^n)^2)}E_j^n Y_n. \tag{8.34}$$

Proof The same computation used in the proof of Lemma 8.2 will be applied. We only provide the proof of the cases for $0 < j < P$, as the other case for $j = 0, P$ can be treated with small modifications. The only difference from the proof of Lemma 8.2 is that instead of (8.9) we employ the boundary conditions

$$\begin{aligned} A_n + B_n &= 0 \quad \text{on } \Gamma_W, \\ A_n e^{i\mu_n b} + B_n e^{-i\mu_n b} &= \widetilde{A}_n + \widetilde{B}_n \quad \text{on } \Gamma_E \end{aligned} \tag{8.35}$$

and

$$2i\mu_n(A_n e^{i\mu_n b} - B_n e^{-i\mu_n b}) - 2i\mu_n(\widetilde{A}_n - \widetilde{B}_n) = 0 \tag{8.36}$$

from the 0th equation.

By solving (8.10), (8.11), (8.12), (8.35) and (8.36) in terms of E_j^n, we obtain that

$$A_n = \frac{e^{i\mu_n b}(1-(Q^n_{j,P})^2)Q^n_{0,j-1}}{4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}E^n_j, \quad B_n = \frac{-e^{i\mu_n b}(1-(Q^n_{j,P})^2)Q^n_{0,j-1}}{4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}E^n_j,$$
$$\widetilde{A}_n = \frac{e^{2i\mu_n b}(1-(Q^n_{j,P})^2)Q^n_{0,j-1}}{4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}E^n_j, \quad \widetilde{B}_n = \frac{-(1-(Q^n_{j,P})^2)Q^n_{0,j-1}}{4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}E^n_j,$$
$$\widetilde{C}_n = \frac{-(1-e^{2i\mu_n b}(Q^n_{0,j-1})^2)}{4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}E^n_j, \quad \widetilde{D}_n = \frac{(1-e^{2i\mu_n b}(Q^n_{0,j-1})^2)(Q^n_{j,P})^2}{4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}E^n_j.$$

Finally, the formulas (8.33) and (8.34) result from substituting them into (8.8) and

$$\frac{\partial u_n}{\partial x} = i\mu_n(A_n e^{i\mu_n b} - B_n e^{-i\mu_n b})Y_n,$$

which completes the proof. □

In order to study the regularity result of the problem (8.2), properties of $t^n_{\ell,j}$ are required. Let us define for $n \neq N$,

$$\mathfrak{t}_{\ell,j} = t^n_{\ell,j} + t^n_{\ell,j+1} \text{ and } \Delta^{\pm}_{\ell,j} = \mathfrak{t}_{\ell,j} \pm \mathfrak{t}_{\ell+1,j}$$

(the formula (8.33) can be extended to j or $\ell = P+1$, saying $t^n_{\ell,j} = 0$ for j or $\ell = P+1$ since $(1 - Q^n_{P+1,P}) = 0$). The following lemma provides estimates of $\Delta^{\pm}_{\ell,j}$ and its proof will be given in the Appendix.

Lemma 8.7 *The following inequalities hold,*

$$\begin{aligned} \frac{1}{\sqrt{|a_\ell|}}|\Delta^+_{\ell,j}|\frac{1}{\sqrt{|a_j|}} &\leq \frac{C}{|\mu_n|^2}, \\ \sqrt{|a_\ell|}|\Delta^-_{\ell,j}|\frac{1}{\sqrt{|a_j|}} &\leq \frac{C}{|\mu_n|}. \end{aligned} \tag{8.37}$$

Also, for the analysis in case of $a_J + i\mu_n = 0$ for some J, we need to estimate the analogues to $\Delta^{\pm}_{\ell,j}$ for $s^n_{\ell,j}$, defined in (8.14). As above, let us define

$$\mathfrak{s}_{\ell,j} = s^n_{\ell,j} + s^n_{\ell,j+1} \text{ and } \Sigma^{\pm}_{\ell,j} = \mathfrak{s}_{\ell,j} \pm \mathfrak{s}_{\ell+1,j}.$$

The same estimates of $\Sigma^{\pm}_{\ell,j}$ as those of $\Delta^{\pm}_{\ell,j}$ are in the following lemma, which can be proved in the same way as Lemma 8.7.

Lemma 8.8 *The following inequalities hold,*

$$\frac{1}{\sqrt{|a_\ell|}}|\Sigma^+_{\ell,j}|\frac{1}{\sqrt{|a_j|}} \le \frac{C}{|\mu_n|^2},$$
$$\sqrt{|a_\ell|}|\Sigma^-_{\ell,j}|\frac{1}{\sqrt{|a_j|}} \le \frac{C}{|\mu_n|}. \tag{8.38}$$

Lemma 8.9 *Let a_j be the parameters defined by (3.5) satisfying (3.6). Then for any $\Upsilon \in (L^2(\Gamma_E))^{P+1}$ the solution (u, Φ) to the problem (8.2) satisfies the regularity result,*

$$\|u\|_{H^2(\Omega_b)} \le C_a(P+1)\|\Upsilon\|_{\mathcal{L}} \tag{8.39}$$

and

$$\|\Phi\|_{V^2_{\Gamma_E}} \le C_a^2(P+1)^2\|\Upsilon\|_{\mathcal{L}}. \tag{8.40}$$

If cutoff modes are excluded, the constants C_a for the stability and regularity estimates are independent of a_j and the exponents on $(P+1)$ are halved; that is the constant in the estimate of u becomes $C(P+1)^{1/2}$ and that for Φ becomes $C(P+1)$.

Proof We first prove (8.40).

Proof of (8.40): Assume that $\Xi = \sum_{n=0}^{\infty} L\Upsilon^n Y_n$ with $\Upsilon^n = (\gamma_0^n, \gamma_1^n, \ldots, \gamma_P^n)^t$. Non-cutoff modes, $n \neq N$: We note that

$$|(1 \pm e^{2i\mu_n b}(Q^n_{p,q})^2)(1 \pm (Q^n_{r,s})^2)Q^n_{c,d}| < 4, \tag{8.41}$$

for any $0 \le p, q, r, s, c, d \le P$ and $|1 - e^{2i\mu_n b}(Q^n_{0,P})^2|$ is bounded below away from zero for all $n \neq N$.

If $a_j + i\mu_n \neq 0$ for $0 \le j \le P$, then Lemma 8.6 shows that the solution ϕ^n_ℓ can be written as $\phi^n_\ell = \sum_{j=0}^P t^n_{\ell,j}(L\Upsilon^n)_j$ and a simple computation gives

$$\begin{aligned}\phi^n_\ell &= \sum_{j=0}^P t^n_{\ell,j}\left[\frac{1}{a_{j-1}}(\gamma^n_{j-1} + \gamma^n_j) + \frac{1}{a_j}(\gamma^n_j + \gamma^n_{j+1})\right] \\ &= \sum_{j=0}^P \mathfrak{t}_{\ell,j}\frac{1}{a_j}(\gamma^n_j + \gamma^n_{j+1}).\end{aligned} \tag{8.42}$$

Now, the Cauchy–Schwarz inequality and (8.37) show that

$$\begin{aligned}\frac{1}{\sqrt{|a_\ell|}}|\phi^n_\ell + \phi^n_{\ell+1}| &\le \sum_{j=0}^P \frac{1}{\sqrt{|a_\ell|}}|\Delta^+_{\ell,j}|\frac{1}{\sqrt{|a_j|}}\frac{1}{\sqrt{|a_j|}}|\gamma^n_j + \gamma^n_{j+1}| \\ &\le \left(\sum_{j=0}^P \left|\frac{1}{\sqrt{|a_\ell|}}\Delta^+_{\ell,j}\frac{1}{\sqrt{|a_j|}}\right|^2\right)^{1/2}\left(\sum_{j=0}^P \frac{1}{|a_j|}|\gamma^n_j + \gamma^n_{j+1}|^2\right)^{1/2} \\ &\le C\frac{\sqrt{P+1}}{|\mu_n|^2}\|\Upsilon^n\|_{\mathcal{L}}\end{aligned}$$

and hence we obtain that

$$(\lambda_n^2+1)^2\|\Phi^n\|_{\mathcal{L}}^2 \le C(P+1)^2\frac{(\lambda_n^2+1)^2}{|\mu_n|^4}\|\Upsilon^n\|_{\mathcal{L}}^2 \le C(P+1)^2\|\Upsilon^n\|_{\mathcal{L}}^2. \tag{8.43}$$

In addition, the same computation as above gives that

$$\begin{aligned}\sqrt{|a_\ell|}|\phi_\ell^n-\phi_{\ell+1}^n| &\le \sum_{j=0}^{P}\sqrt{|a_\ell|}|\Delta_{\ell,j}^-|\frac{1}{\sqrt{|a_j|}}\frac{1}{\sqrt{|a_j|}}|\gamma_j^n+\gamma_{j+1}^n|\\ &\le \left(\sum_{j=0}^{P}\left|\sqrt{|a_\ell|}\Delta_{\ell,j}^-\frac{1}{\sqrt{|a_j|}}\right|^2\right)^{1/2}\left(\sum_{j=0}^{P}\frac{1}{|a_j|}|\gamma_j^n+\gamma_{j+1}^n|^2\right)^{1/2}\\ &\le C\frac{\sqrt{P+1}}{|\mu_n|}\|\Upsilon^n\|_{\mathcal{L}},\end{aligned}$$

which shows that

$$(\lambda_n^2+1)\|\Phi^n\|_{\mathcal{M}}^2 \le C(P+1)^2\frac{\lambda_n^2+1}{|\mu_n|^2}\|\Upsilon^n\|_{\mathcal{L}}^2 \le C(P+1)^2\|\Upsilon^n\|_{\mathcal{L}}^2. \tag{8.44}$$

Thus, it follows from (8.43) and (8.44) that

$$(\lambda_n^2+1)^2\|\Phi^n\|_{\mathcal{L}}^2+(\lambda_n^2+1)\|\Phi^n\|_{\mathcal{M}}^2 \le C(P+1)^2\|\Upsilon^n\|_{\mathcal{L}}^2. \tag{8.45}$$

In case when $a_J+i\mu_n=0$ for some J, the system of Eqs. (8.29) breaks into two block diagonal systems. We notice that ϕ_ℓ^n is represented by

$$\phi_\ell^n=\begin{cases}\displaystyle\sum_{j=0}^{J}t_{\ell,j}^n(L\Upsilon^n)_j & \text{for } \ell\le J,\\ \displaystyle\sum_{j=J+1}^{P}s_{\ell,j}^n(L\Upsilon^n)_j & \text{for } \ell\ge J+1,\end{cases} \tag{8.46}$$

where $t_{\ell,j}^n$ and $s_{\ell,j}^n$ are defined by (8.33) (with P replaced by J) and (8.14), respectively. By using Lemmas 8.7 and 8.8 as in the argument used above, the same result as (8.45) can be derived.

Cutoff modes, $n=N$: In this case, Φ^N satisfies (8.32), which is equivalent to

$$2b^{-1}\phi_0^N\boldsymbol{e}_0+M\Phi^N=L\Upsilon^N \tag{8.47}$$

since $A_N=0$ from the boundary condition on Γ_W and $A_N+B_Nb=\phi_0^N$. By examining the real and imaginary parts of the inner product of the left hand side of

(8.47) with Φ^N, we observe that

$$\begin{aligned}\frac{2}{b}|\phi_0^N|^2 + \|\Phi^N\|_{\mathcal{M}}^2 &\le C\left|(\frac{2}{b}\phi_0^N \boldsymbol{e}_0 + M\Phi^N, \Phi^N)_{\mathbb{C}^{P+1}}\right| \le C\|\Upsilon^N\|_{\mathcal{L}}\|\Phi^N\|_{\mathcal{L}} \\ &\le C_a(P+1)\|\Upsilon^N\|_{\mathcal{L}}\|\Phi^N\|_{\mathcal{M}}.\end{aligned} \tag{8.48}$$

The last inequality is the result from Lemma 4.5. Therefore, it follows that

$$\|\Phi^N\|_{\mathcal{M}} \le C_a(P+1)\|\Upsilon^N\|_{\mathcal{L}}. \tag{8.49}$$

Applying Lemma 4.5 again to the above inequality (8.49) yields that

$$\|\Phi^N\|_{\mathcal{L}} \le C_a^2(P+1)^2\|\Upsilon^N\|_{\mathcal{L}}$$

and hence it is concluded that

$$(\lambda_N^2+1)^2\|\Phi^N\|_{\mathcal{L}}^2 + (\lambda_N^2+1)\|\Phi^N\|_{\mathcal{M}}^2 \le C_a^4(P+1)^4\|\Upsilon^N\|_{\mathcal{L}}^2. \tag{8.50}$$

Finally, combining (8.45) and (8.50) implies

$$\|\Phi\|_{\boldsymbol{V}_{\Gamma_E}^2} \le C_a^2(P+1)^2\|\Upsilon\|_{\mathcal{L}},$$

which completes the proof of (8.40).

Proof of (8.39): We shall estimate $g_{bd} = \partial u/\partial x - T(u)$ in $H^{1/2}(\Gamma_E)$,

$$\|g_{bd}\|_{H^{1/2}(\Gamma_E)} \le C_a(P+1)\|\Upsilon\|_{\mathcal{L}}. \tag{8.51}$$

Once the inequality is established, Lemma 6.4 with (8.51) yields that

$$\|u\|_{H^2(\Omega_b)} \le C_a(P+1)\|\Upsilon\|_{\mathcal{L}},$$

which completes the proof of (8.39).

Now, we are left with proving (8.51). To do this, as done in (8.42) we use (8.33) with $\ell = 0$ and (8.34) for $n \ne N$ and $a_j + i\mu_n \ne 0$ to have

$$\frac{\partial u_n}{\partial x} - i\mu_n u_n = \sum_{j=0}^{P} \frac{Q_{0,j-1}^n(1 - Q_{j,j}^n(Q_{j+1,P}^n)^2)(1+Q_{j,j}^n)}{2(1-e^{2i\mu_n b}(Q_{0,P}^n)^2)}\frac{1}{a_j}(\gamma_j + \gamma_{j+1})Y_n.$$

The Cauchy–Schwarz inequality and (8.18) shows that

$$\|\frac{\partial u_n}{\partial x} - i\mu_n u_n\|_{L^2(\Gamma_E)}^2 \le \left(\sum_{j=0}^{P} C|1+Q_{j,j}^n|\frac{1}{|a_j|}|\gamma_j^n + \gamma_{j+1}^n|\right)^2$$

$$\leq \left(\sum_{j=0}^{P} C\frac{|1+Q_{j,j}^n|^2}{|a_j|}\right)\left(\sum_{j=0}^{P}\frac{1}{|a_j|}|\gamma_j^n+\gamma_{j+1}^n|^2\right)$$
$$\leq C\frac{P+1}{|\mu_n|}\|\Upsilon^n\|_{\mathcal{L}}^2$$

and so we obtain

$$(\lambda_n^2+1)^{1/2}\|\frac{\partial u_n}{\partial x}-i\mu_n u_n\|_{L^2(\Gamma_E)}^2 \leq C(P+1)\|\Upsilon^n\|_{\mathcal{L}}^2. \tag{8.52}$$

In case when $a_J + i\mu_n = 0$ for some n and j, since $\partial u_n/\partial x$ and u_n are affected by only the first $(J+1)$ components of Υ, it holds that

$$(\lambda_n^2+1)^{1/2}\|\frac{\partial u_n}{\partial x}-i\mu_n u_n\|_{L^2(\Gamma_E)}^2 \leq C(J+1)\|\Upsilon^n\|_{\mathcal{L}}^2 \leq C(P+1)\|\Upsilon^n\|_{\mathcal{L}}^2. \tag{8.53}$$

For the cutoff mode, i.e., $n = N$, we use (8.48) and (8.49) to see

$$\begin{aligned}\|\frac{\partial u_N}{\partial x}\|_{L^2(\Gamma_E)}^2 &= |B_N|^2 = \frac{1}{b^2}|\phi_0^N|^2\\ &\leq C_a(P+1)\|\Upsilon^N\|_{\mathcal{L}}\|\Phi^N\|_{\mathcal{M}} \leq C_a^2(P+1)^2\|\Upsilon^N\|_{\mathcal{L}}^2.\end{aligned} \tag{8.54}$$

Finally by combining (8.52), (8.53) and (8.54) we obtain

$$\|\frac{\partial u}{\partial x}-T(u)\|_{H^{1/2}(\Gamma_E)} \leq C_a(P+1)\|\Upsilon\|_{\mathcal{L}},$$

which completes the proof. □

9 Finite element approximations

Now, we are in a position to discuss the solvability and quasi-optimal convergence of the finite element approximation (u_h, Φ_h) to the solution u and the auxiliary variables $\Phi = (\phi_0, \ldots, \phi_P)^t$ to the variational problem (4.1).

Let $\mathcal{T}_h$ denote a partition of Ω_b with shape-regular meshes and let h represent the diameter of elements, e.g., $h = \max_{K\in\mathcal{T}_h}\mathrm{diam}(K)$. By extracting the boundary nodes on Γ_E generated by $\mathcal{T}_h$, we define the boundary meshes, which are denoted by $\mathcal{T}_h^b$. Let $\widetilde{S}_h$ denote a subspace of $\widetilde{H}^1(\Omega_b)$ consisting of piecewise polynomial finite element functions and S_h^0 denote the subset of functions in $\widetilde{S}_h$ which vanish on Γ_W. Also, G_h is analogously defined by a finite element subspace of $H^1(\Gamma_E)$. We assume that f is the trace of a function on Γ_W in our approximation space as the errors associated with boundary quadrature in the finite element method are well understood. Let S_h be the set of functions in $\widetilde{S}_h$ which coincide with f on Γ_W. Denoting by $\boldsymbol{V}_h$ the set of all elements (u_h, Φ_h) in $S_h \times (G_h)^{P+1}$ such that $u_h = \phi_{h,0}$ on Γ_E for $\Phi_h = (\phi_{h,0}, \ldots, \phi_{h,P})^t$ and

by $\boldsymbol{V}_h^0$ the set of all elements (u_h, Φ_h) in $S_h^0 \times (G_h)^{P+1}$ such that $u_h = \phi_{h,0}$ on Γ_E, the finite element approximation to (u, Φ) is the function $(u_h, \Phi_h) \in \boldsymbol{V}_h$ satisfying

$$A((u_h, \Phi_h), (\xi_h, \Psi_h)) = 0 \quad \text{for all } (\xi_h, \Psi_h) \in \boldsymbol{V}_h^0. \tag{9.1}$$

As mentioned earlier, we will now invoke an argument due to Schatz [33] to establish the unique solvability and quasi-optimal convergence of finite element approximations. This requires that the mesh size h satisfies $0 < h < h_0$ for a constant h_0, which may depend on the stability and regularity estimates of the elliptic problem studied in Sect. 8.

In our case, for a given order (n_p, n_e) with $P = n_p + n_e$ and the damping parameters a_j given by (3.5) satisfying (3.6), we already know that the sesquilinear form $A(\cdot, \cdot)$ is bounded,

$$|A((u, \Phi), (\xi, \Psi))| \le C\|(u, \Phi)\|_{\boldsymbol{V}} \|(\xi, \Psi)\|_{\boldsymbol{V}}.$$

Also, since

$$|((M - \bar{M})\Phi, \Phi)_{\Gamma_E}| = \sum_{j=0}^{n_p - 1} 2|a_j| \|\phi_j - \phi_{j+1}\|^2_{L^2(\Gamma_E)} \le C n_p^2 \|\Phi\|^2_{\mathcal{L}}$$

due to the fact that $|a_j| \le k$ for $j = 0, \ldots, n_p - 1$, it follows from (5.13) that the sesquilinear form $A(\cdot, \cdot)$ satisfies the inequality

$$|A((u, \Phi), (u, \Phi))| \ge C_1 \|(u, \Phi)\|^2_{\boldsymbol{V}} - C_2 n_p^2 (\|u\|^2_{L^2(\Omega_b)} + \|\Phi\|^2_{\mathcal{L}}) \tag{9.2}$$

in $\boldsymbol{V}_0 \times \boldsymbol{V}_0$ for some positive constants C_1 and C_2. Now, the solvability and quasi-optimal convergence of finite element approximations are given in the following theorem. The proof follows the same line as the standard Schatz's argument in [33] with the regularity result given in Theorem 8.1 and Lemma 8.9.

Theorem 9.1 *Let a_j be the parameters defined by (3.5) satisfying (3.6). Then there exists an $h_0 > 0$ such that for $0 < h < h_0$, (9.1) has a unique solution $(u_h, \Phi_h) \in \boldsymbol{V}_h$ satisfying*

$$\|(u, \Phi) - (u_h, \Phi_h)\|_{\boldsymbol{V}} \le Ch \|(u, \Phi)\|_{\boldsymbol{V}^2}. \tag{9.3}$$

Furthermore, the solution u_h satisfies the L^2-error estimate

$$\|u - u_h\|_{L^2(\Omega_b)} \le C_a (P + 1) h^2 \|(u, \Phi)\|_{\boldsymbol{V}^2}. \tag{9.4}$$

Here the constant C_a is independent of a_j if cutoff modes are not involved.

Proof Let $(e, E) = (u, \Phi) - (u_h, \Phi_h) \in \boldsymbol{V}_0$ be the error function. Since the sesquilinear form $A(\cdot, \cdot)$ is symmetric (not Hermitian), that is, $A((u, \Phi), (\xi, \Psi)) =$

$A((\bar{\xi}, \bar{\Psi}), (\bar{u}, \bar{\Phi}))$ for $(u, \Phi), (\xi, \Psi) \in \boldsymbol{V}_0$, the solution $(w, \Upsilon) \in \boldsymbol{V}_0$ to the dual problem

$$A((\xi, \Psi), (w, \Upsilon)) = (\xi, e)_{\Omega_b} + (L\Psi, E)_{\Gamma_E} \quad \text{for all } (\xi, \Psi) \in \boldsymbol{V}_0$$

also satisfies the regularity estimates in Theorem 8.1 and Lemma 8.9. By choosing a linear or bilinear interpolation $\Upsilon_h = (\gamma_{h,0}, \ldots, \gamma_{h,P})^t$ of $\Upsilon = (\gamma_0, \ldots, \gamma_P)^t$, it is obvious that

$$\begin{aligned}\|\Upsilon - \Upsilon_h\|^2_{\mathcal{L},1} &= \sum_{j=0}^{P} \frac{1}{|a_j|} \|\gamma_j + \gamma_{j+1} - \gamma_{h,j} - \gamma_{h,j+1}\|^2_{H^1(\Gamma_E)} \\ &\le Ch^2 \sum_{j=0}^{P} \frac{1}{|a_j|} \|\gamma_j + \gamma_{j+1}\|^2_{H^2(\Gamma_E)} = Ch^2 \|\Upsilon\|^2_{\mathcal{L},2}\end{aligned}$$

and

$$\begin{aligned}\|\Upsilon - \Upsilon_h\|^2_{\mathcal{M}} &= \sum_{j=0}^{P} |a_j| \|(\gamma_j - \gamma_{j+1}) - (\gamma_{h,j} - \gamma_{h,j+1})\|^2_{L^2(\Gamma_E)} \\ &\le Ch^2 \sum_{j=0}^{P} |a_j| \|\gamma_j - \gamma_{j+1}\|^2_{H^1(\Gamma_E)} = Ch^2 \|\Upsilon\|^2_{\mathcal{M},1},\end{aligned}$$

which reveals that

$$\|(w, \Upsilon) - (w_h, \Upsilon_h)\|_{\boldsymbol{V}} \le Ch \|(w, \Upsilon)\|_{\boldsymbol{V}^2} \tag{9.5}$$

with a linear or bilinear interpolation w_h of w. The approximation property (9.5) and Lemma 8.9 show that

$$\begin{aligned}\|e\|^2_{L^2(\Omega_b)} + \|E\|^2_{\mathcal{L}} &\le C |A((e, E), (w, \Upsilon) - (w_h, \Upsilon_h))| \\ &\le Ch \|(e, E)\|_{\boldsymbol{V}} \|(w, \Upsilon)\|_{\boldsymbol{V}^2} \\ &\le C_a^2 (P+1)^2 h \|(e, E)\|_{\boldsymbol{V}} (\|e\|^2_{L^2(\Omega_b)} + \|E\|^2_{\mathcal{L}})^{1/2},\end{aligned} \tag{9.6}$$

which in turn gives

$$(\|e\|^2_{L^2(\Omega_b)} + \|E\|^2_{\mathcal{L}})^{1/2} \le C_a^2 (P+1)^2 h \|(e, E)\|_{\boldsymbol{V}}. \tag{9.7}$$

From Gårding's inequality (9.2) for (e, E),

$$\begin{aligned}C_1 \|(e, E)\|^2_{\boldsymbol{V}} - C_2 n_p^2 (\|e\|^2_{L^2(\Omega_b)} + \|E\|^2_{\mathcal{L}}) &\le |A((e, E), (e, E))| \\ &= |A((e, E), (u, \Phi))| \le C \|(e, E)\|_{\boldsymbol{V}} \|(u, \Phi)\|_{\boldsymbol{V}},\end{aligned}$$

we see that

$$C_1\|(e,E)\|_{\boldsymbol{V}} - C_2 n_p^2(\|e\|^2_{L^2(\Omega_b)} + \|E\|^2_{\mathcal{L}})^{1/2} \le C\|(u,\Phi)\|_{\boldsymbol{V}}$$

and apply (9.7) to the inequality to obtain

$$(C_1 - C_2C_a^2n_p^2(P+1)^2h)\|(e,E)\|_{\boldsymbol{V}} \le C\|(u,\Phi)\|_{\boldsymbol{V}}. \tag{9.8}$$

For unique solvability of the finite dimensional problem, suppose that $f = 0$, and so $(u,\Phi) = 0$. Then there exists h_0 such that $C_1 - C_2C_a^2n_p^2(P+1)^2h_0 > 0$. For such $0 < h < h_0$, we clearly see that $(e,E) = 0$, implying the unique solvability of finite element problem.

Also, the error estimate (9.3) in the energy norm is proved from Gårding's inequality for $0 < h < h_0$ and Theorem 8.1,

$$\begin{aligned} C\|(e,E)\|^2_{\boldsymbol{V}} &\le |A((e,E),(e,E))| = |A((e,E),(u,\Phi)-(u_h,\Phi_h))| \\ &\le Ch\|(e,E)\|_{\boldsymbol{V}}\|(u,\Phi)\|_{\boldsymbol{V}^2} \end{aligned}$$

with a linear or bilinear interpolation (u_h,Φ_h) of (u,Φ), which leads to (9.3).

For the L^2-error estimate, let $(w_e,\Upsilon_e) \in \boldsymbol{V}_0$ be the solution to the adjoint problem

$$A((\xi,\Psi),(w_e,\Upsilon_e)) = (\xi,e)_{\Omega_b}$$

for all $(\xi,\Psi) \in \boldsymbol{V}_0$. Then the same argument used for (9.6) with Theorem 8.1 instead of Lemma 8.9 shows again that

$$\begin{aligned} \|e\|^2_{L^2(\Omega_b)} &= A((e,E),(w_e,\Upsilon_e)) \\ &\le Ch\|(e,E)\|_{\boldsymbol{V}}\|(w_e,\Upsilon_e)\|_{\boldsymbol{V}^2} \\ &\le C_a(P+1)h\|(e,E)\|_{\boldsymbol{V}}\|e\|_{L^2(\Omega_b)}, \end{aligned}$$

which implies that

$$\|e\|_{L^2(\Omega_b)} \le C_a(P+1)h\|(e,E)\|_{\boldsymbol{V}} \le C_a(P+1)h^2\|(u,\Phi)\|_{\boldsymbol{V}^2}$$

and completes the proof. □

We note that the regularity constant in Lemma 8.9 may increase polynomially (quadratically, but linearly if cutoff modes are excluded) as P grows and so a smaller mesh h may be required for large P to retain the unique solvability and quasi-optimal convergence, though this has not been encountered in our experiments. However, when a cutoff modes is present, C_a depending on $\max_{j=0,\dots,P}\{1/|a_j|\}$ comes in the regularity constant and it is found in numerical tests that the convergence of finite element approximations is affected by the smallest parameter used for CRBCs. A discussion on the convergence with respect to C_a and h will be made in the following section.

10 Numerical experiments

In this section we provide numerical examples that confirm the well-posedness and convergence theories that were developed in the preceding sections. We specialize to $\mathbb{R}^2$ and take $\Theta = (0, W)$. Note that now

$$Y_n(y) = \sqrt{\frac{2}{W}} \cos\left(\frac{n\pi}{W} y\right)$$

are transverse eigenfunctions associated with eigenvalues $\lambda_n^2 = (n\pi/W)^2$ for $n \geq 0$. The domain $\Omega_b = (0, b) \times (0, W)$ is a rectangular region obtained by truncating the semi-infinite waveguide Ω_∞ at $x = b$ (see Fig. 2). We set $W = 1$.

In the first example, we take $k = 20$ and choose f corresponding to the analytic solution of (2.1)–(2.3):

$$u^{ex}(x, y) = \sum_{n=0}^{6} \frac{1}{7\sqrt{2}} e^{i\mu_n x} Y_n(y)$$

in Ω_b with $b = 0.2$. The exact solution u^{ex} is a superposition of seven propagating modes. In order to apply an efficient CRBC on Γ_E, the optimal parameters discussed in Sect. 7 are computed on the interval $[\mu_6, k] \approx [6.6853, 20]$ by the Remez algorithm and their distributions for $n_p = 1, 2, \ldots, 5$ are depicted in Fig. 4. Their maximal reflection coefficients for propagating modes are presented in Table 2 as well. We compute piecewise bilinear finite element approximations u_h with mesh $h = 1/800, 1/1600$ and $1/3200$ by using the finite element library `deal.II` [1]. To see the convergence of approximate solutions, we measure relative L^2- and H^1-errors and report the errors in Fig. 5. It is observed that approximate solutions obtained by CRBCs converge as the order of CRBCs increases until mesh size errors dominate. In particular, when the mesh size is small enough so that mesh error is ignorable, the relative L^2-error converges at the same convergence rate of the maximal reflection coefficients.

The second example illustrates the effect of CRBCs on evanescent modes. To do this, we take $k = 20$ and choose an analytic solution u^{ex} including seven propagating modes and ten evanescent modes

$$u^{ex}(x, y) = \sum_{n=0}^{16} \frac{1}{17\sqrt{2}} e^{i\mu_n x} Y_n(y).$$

We also assume that the source coming from the left boundary Γ_W is close to the artificial boundary Γ_E, e.g., $b = 0.1$ ($W = 1$). For this example, we use the same purely imaginary parameters as those obtained with $n_p = 4$ since the CRBC with $n_p = 4$ serves as an accurate absorbing boundary condition for propagating modes for the meshes $h = 1/800, 1/1600$ and $1/3200$. For the real parameters responsible for damping evanescent modes, we solve numerically the min–max problem (6.2) on the interval $[\tilde{\mu}_7, M_\sigma] \approx [9.1438, 147.0887]$, where M_σ is determined by $e^{-M_\sigma b} = \rho_p$. The distribution of the real parameters and the maximal reflection coefficients ρ_e for

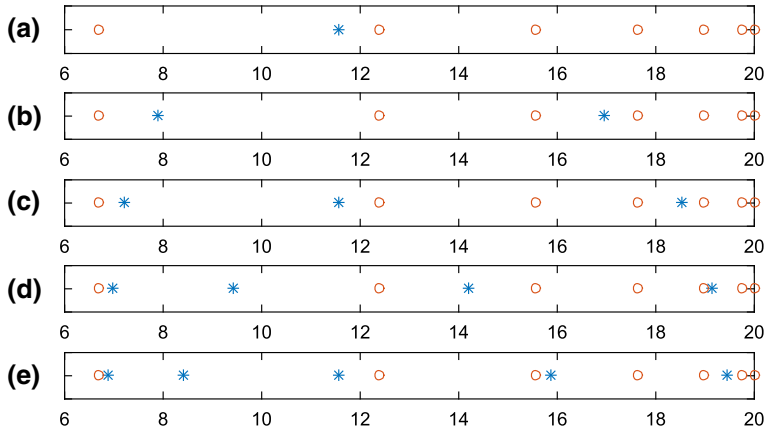


Fig. 4 Distribution of optimal parameters for $n_p = 1, 2, \ldots, 5$. The seven red circles represent the exact propagation frequencies μ_n and the blue * marks are the optimal parameters of $n_p = 1$ in (**a**) through $n_p = 5$ in (**e**)

Table 2 Maximal reflection coefficients for propagating modes and evanescent modes resulting from CRBCs with the optimal parameters for $k = 20$

n_p	ρ_p	n_e	ρ_e
1	7.1448E−2	1	3.6102E−1
2	1.2794E−3	2	3.4899E−2
3	2.2883E−5	3	3.2613E−3
4	4.0927E−7	4	3.0468E−4
5	7.3199E−9	5	2.8463E−5

each n_e are shown in Fig. 6 and Table 2, respectively. The numerical results given in Fig. 7 also illustrate the convergence of solutions with respect to increasing n_e. Also, it can be seen that the convergence rate of the relative L^2-errors coincides with the decay rate of ρ_e as long as the mesh is fine enough.

In the third example, the performance of CRBCs for the cutoff mode is examined. We set $k = 6\pi$ and choose u^{ex} such that the exact solution is composed of six propagating modes and one cutoff mode:

$$u^{ex}(x, y) = \sum_{n=0}^{6} \frac{1}{7\sqrt{2}} e^{i\mu_n x} Y_n(y)$$

defined on Ω_b with $b = 0.2$ ($W = 1$). We increase the number of purely imaginary parameters in the optimal way for propagating modes with $n_p = 1 \sim 30$. As indicated in Theorem 6.1, the error of the cutoff mode is controlled by $S_P = |b+2\sum_{j=0}^{P} a_j^{-1}|^{-1}$, which is illustrated in Fig. 8. We notice that the optimal parameters used for propagating modes do not seem to be the best choice.

In case that cutoff modes are involved, we may want to try other choices of parameters, with which CRBCs can reduce S_P to much smaller level while the reflection

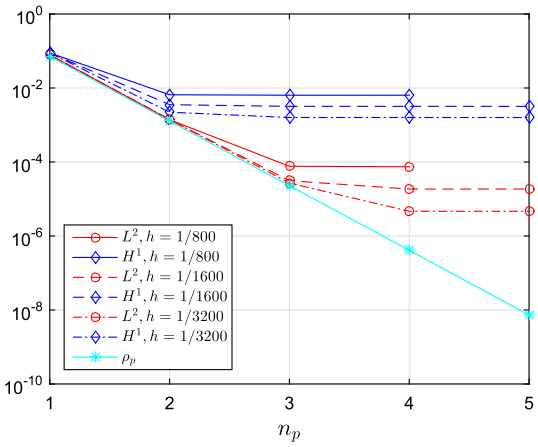


Fig. 5 Relative L^2- and H^1-errors for the exact propagating solution

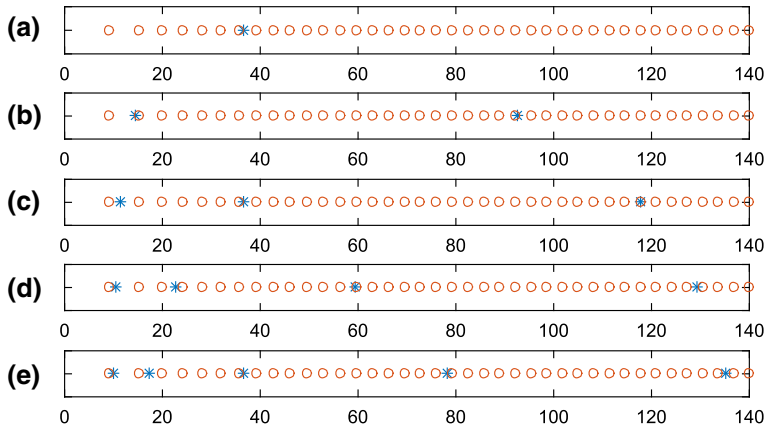


Fig. 6 Distribution of optimal parameters for $n_e = 1, 2, \ldots, 5$. The red circles represent the exact decay rate of evanescent modes $\tilde{\mu}_n$ and the blue * marks are the optimal parameters of $n_e = 1$ in (**a**) through $n_e = 5$ in (**e**)

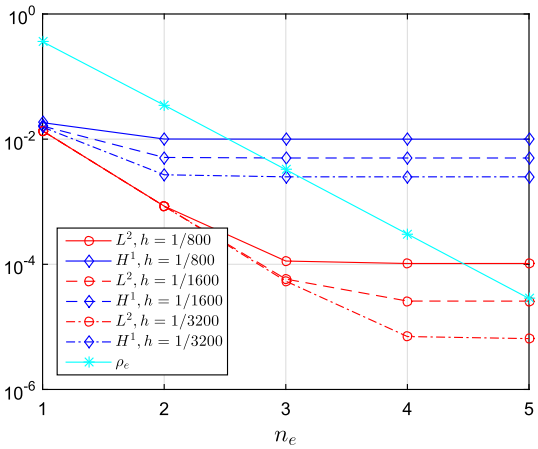


Fig. 7 Relative L^2- and H^1-errors for the solution including both of propagating modes and evanescent modes

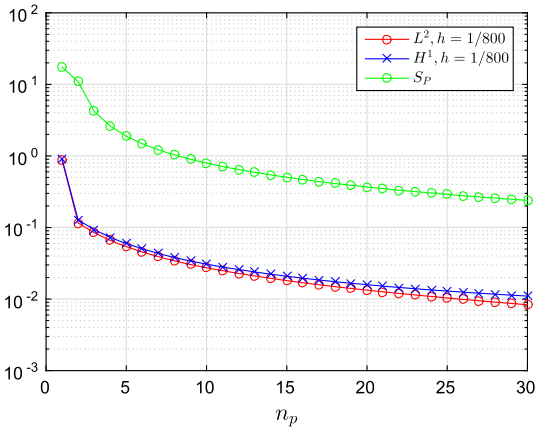


Fig. 8 Relative L^2- and H^1-errors for the solution including both of propagating modes and a cutoff mode satisfying CRBCs with optimal parameters

coefficients from propagating modes are not deteriorated too much, e.g., Newman's nodes $a_j = -ike^{j/\sqrt{P}}$ based on geometric sequences for $j = 0, \ldots, P$. As we can see Fig. 9 of relative L^2-errors, the CRBCs with geometric sequences produce improved results, though it is observed that the errors obtained from this approach have an irregular behavior for large P. It can be explained in terms of a small parameter a_P for large P. According to the formula for S_P, it seems that one might improve the accuracy of CRBCs at the continuous level by adding a small parameter such as the smallest parameter a_P of the geometric sequences, which reduces S_P to the error tolerance. However, the cutoff mode on the discrete level does not satisfy the actual equation on the continuous level

$$M\Phi^N = -2\frac{\partial u}{\partial x}\boldsymbol{e}_0$$

but solves an equation of a propagating or evanescent mode

$$(-\mu_{N,h}^2 L + M)\Phi_h^N = -2\frac{\partial u_h}{\partial x}\boldsymbol{e}_0$$

for some discrete axial frequency $\mu_{N,h} \neq 0$ since no discrete eigenvalue of the transverse Laplace operator will typically coincide with the cutoff transverse eigenvalue λ_n^2. When small parameters are used, some components of L become large but in contrast corresponding components of M become small. Therefore in case that h is not small enough that $\mu_{N,h}$ is big, $-\mu_{N,h}^2 L$ might be dominant over the actual cutoff mode system matrix M and so the resulting solution would not be accurate. The mesh size affected by the small parameter used for CRBCs can be examined in Fig. 9. We observe the minimum errors at $P = 13, 17, 22$ for $h = 1/800, 1/1600, 1/3200$, respectively, in the plot and they are shifted as h is halved. The ratios of the smallest parameter $a_P = -ike^{-\sqrt{P}}$ determining $C_a = O(a_P^{-1})$ between two minimum error

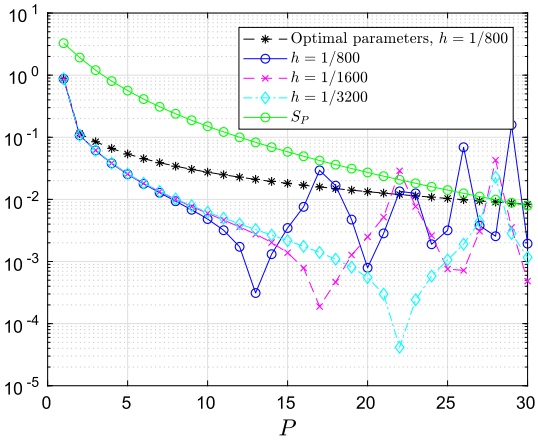
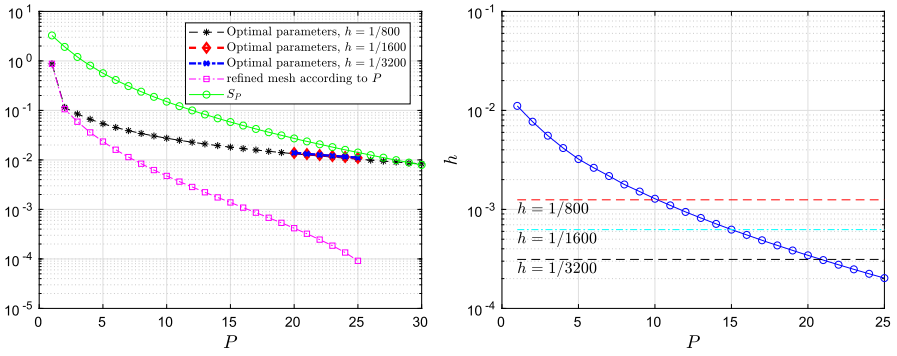


Fig. 9 Relative L^2-errors for the solution including both of propagating modes and a cutoff mode satisfying CRBCs with Newman nodes

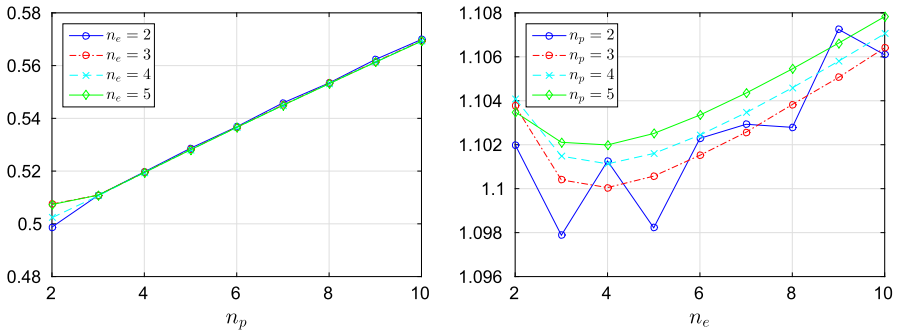


(a) Relative L^2-errors **(b)** mesh size for each P

Fig. 10 Relative L^2-errors in approximate solutions with h refined according to P for the solution including both of propagating modes and a cutoff mode satisfying CRBCs with Newman nodes

points are $e^{\sqrt{13}}/e^{\sqrt{17}} \approx 0.5960$ and $e^{\sqrt{17}}/e^{\sqrt{21}} \approx 0.5670$, which indicates that it appears that $C_a h$ in Gåding's inequality is the main factor contributing to solvability and quasi-optimality of the finite element analysis (9.8), and it is necessary to choose h small enough when cutoff modes exist and a_P is small. To see this observation in more detail, we take a mesh refinement according to P in such a way that $e^{\sqrt{P}}h$ is a constant C_{newman}. For example, C_{newman} is taken to be $e^{\sqrt{10}}/800 \approx 0.03$ and we do numerical tests with $h = C_{\text{newman}} e^{-\sqrt{P}}$ for each P. The results are given in Fig. 10a with mesh size for each P in (b). While relative L^2-errors in approximations for optimal parameters with decreasing h are not improved due to reflection errors, those for Newman's nodes decrease asymptotically at the same rate of that of S_P, without any oscillatory behavior as long as meshes are refined according to P.

Aside from this, it is found in Fig. 11 that the norm of auxiliary variables, $(\|\Phi\|^2_{\mathcal{L}} + \|\Phi\|^2_{\mathcal{M}})^{1/2}$, of the second and third examples increases with increasing P as in the stability analysis of Theorem 8.1 but its variance is small. The independence of the



(a) propagating and evanescent modes **(b)** propagating and cutoff modes

Fig. 11 Norm of auxiliary variables, $(\|\Phi\|^2_{\mathcal{L}} + \|\Phi\|^2_{\mathcal{M}})^{1/2}$

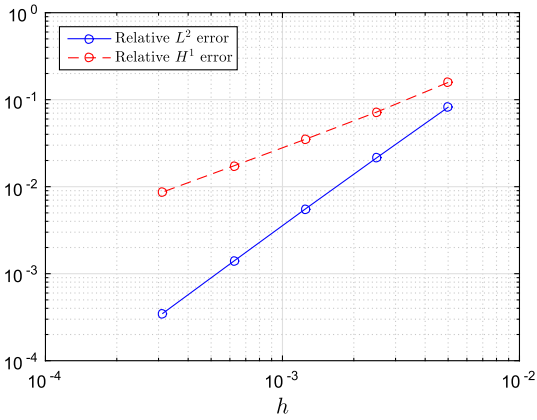


Fig. 12 Relative L^2- and H^1-errors with $h = 1/200, 1/400, 1/800, 1/1600, 1/3200$

finite element problem from P seems to be caused by the small variance of the norm with respect to P.

In the last example, we are concerned with finite element convergence as h approaches zero. To do this, we set $k = 100$ and take the computational domain to be $\Omega_b = (0, 0.1) \times (0, 1)$, i.e., $b = 0.1$ and $W = 1$, for which the number of propagating modes is 32. We choose the CRBC of order $(n_p, n_e) = (4, 3)$ for which $\rho_p = 3.9590 \times 10^{-6}$ and $\rho_e = 5.3492 \times 10^{-5}$ and so reflection errors are negligible compared with mesh errors. The wave source f on Γ_W is given so that the exact solution is defined by

$$u(x, y) = \sum_{n=0}^{31} \frac{1}{64\sqrt{2}} e^{i\mu_n x} Y_n(y) + \sum_{n=32}^{63} \frac{1}{64\sqrt{2}} e^{-\tilde{\mu}_n x} Y_n(y)$$

having 32 propagating modes and 32 evanescent modes. The plot in Fig. 12 shows the quasi-optimal convergence of relative L^2- and H^1-errors in finite element approximations with $(n_p, n_e) = (4, 3)$.

11 Appendix

Proof of Lemma 8.7 $\mathrm{t}_{\ell,j}$ is given by

$$\begin{cases} \dfrac{(1-e^{2i\mu_n b}(Q^n_{0,\ell-1})^2)Q^n_{\ell,j-1}(1-Q^n_{j,j}(Q^n_{j+1,P})^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}(1+Q^n_{j,j}) & \text{if } \ell \le j, \\ \dfrac{(1-e^{2i\mu_n b}(Q^n_{0,j-1})^2Q^n_{j,j})Q^n_{j+1,\ell-1}(1-(Q^n_{\ell,P})^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}(1+Q^n_{j,j}) & \text{if } \ell > j. \end{cases} \tag{11.1}$$

We first consider $\Delta^{\pm}_{\ell,j}$ for $\ell \neq j$. A simple computation shows

$$\Delta^{+}_{\ell,j} = \left[\frac{(1-e^{2i\mu_n b}(Q^n_{0,\ell-1})^2Q^n_{\ell,\ell})Q^n_{\ell+1,j-1}(1-Q^n_{j,j}(Q^n_{j+1,P})^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}\right] (1+Q^n_{\ell,\ell})(1+Q^n_{j,j})$$

for $\ell < j$, and $\Delta^{+}_{\ell,j} = \Delta^{+}_{j,\ell}$ by the symmetry of $t_{\ell,j}$. Analogously, it can be shown that $\Delta^{-}_{\ell,j}$ is given by

$$\left[\frac{(1+e^{2i\mu_n b}(Q^n_{0,\ell-1})^2Q^n_{\ell,\ell})Q^n_{\ell+1,j-1}(1-Q^n_{j,j}(Q^n_{j+1,P})^2)}{4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}\right](1-Q^n_{\ell,\ell})(1+Q^n_{j,j})$$

for $\ell < j$ and

$$\left[\frac{(1-e^{2i\mu_n b}(Q^n_{0,j-1})^2Q^n_{j,j})Q^n_{j+1,\ell-1}(1+Q^n_{\ell,\ell}(Q^n_{\ell+1,P})^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}\right](1-Q^n_{\ell,\ell})(1+Q^n_{j,j})$$

for $\ell > j$. Thus, by (8.18) we have

$$\frac{1}{\sqrt{|a_\ell|}}|\Delta^{+}_{\ell,j}|\frac{1}{\sqrt{|a_j|}} \le \frac{C}{|i\mu_n|}\left(\frac{|1+Q^n_{\ell,\ell}|}{\sqrt{|a_\ell|}}\right)\left(\frac{|1+Q^n_{j,j}|}{\sqrt{|a_j|}}\right) \le \frac{C}{|\mu_n|^2}, \tag{11.2}$$

$$\sqrt{|a_\ell|}|\Delta^{-}_{\ell,j}|\frac{1}{\sqrt{|a_j|}} \le \frac{C}{|i\mu_n|}\left(\sqrt{|a_\ell|}|1-Q^n_{\ell,\ell}|\right)\left(\frac{|1+Q^n_{j,j}|}{\sqrt{|a_j|}}\right) \le \frac{C}{|\mu_n|}. \tag{11.3}$$

In case of $\ell = j$, we see that

$$\Delta^{+}_{\ell,\ell} = \left[\frac{2(1+e^{2i\mu_n b}(Q^n_{0,P})^2/Q^n_{\ell,\ell}) - (1+Q^n_{\ell,\ell})((Q^n_{\ell+1,P})^2 + e^{2i\mu_n b}(Q^n_{0,\ell-1})^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}\right](1+Q^n_{\ell,\ell}), \tag{11.4}$$

$$\Delta^{-}_{\ell,\ell} = \left[\frac{((Q^n_{\ell+1,P})^2 - e^{2i\mu_n b}(Q^n_{0,\ell-1})^2)}{-4i\mu_n(1-e^{2i\mu_n b}(Q^n_{0,P})^2)}\right](1-Q^n_{\ell,\ell})(1+Q^n_{\ell,\ell}). \tag{11.5}$$

Now by the fact that $|1 + Q^n_{\ell,\ell}|/|a_\ell| = 2/|a_\ell - i\mu_n| \le C/|\mu_n|$, (8.18), (11.4) and (11.5), it is easy to show that

$$\frac{1}{\sqrt{|a_\ell|}}|\Delta^+_{\ell,\ell}|\frac{1}{\sqrt{|a_\ell|}} \le C\frac{|1+Q^n_{\ell,\ell}|}{|i\mu_n||a_\ell|} \le \frac{C}{|\mu_n|^2},$$
$$\sqrt{|a_\ell|}|\Delta^-_{\ell,\ell}|\frac{1}{\sqrt{|a_\ell|}} \le \frac{C}{|\mu_n|},$$

which completes the proof. □

References

1. Bangerth, W., Hartmann, R., Kanschat, G.: deal. II—a general-purpose object-oriented finite element library. ACM Trans. Math. Softw. **33**(4), 24 (2007)
2. Bécache, E., Dhia, A.-S.B.-B., Legendre, G.: Perfectly matched layers for the convected Helmholtz equation. SIAM J. Numer. Anal. **42**(1), 409–433 (2004)
3. Bendali, A., Guillaume, P.: Non-reflecting boundary conditions for waveguides. Math. Comput. **68**(225), 123–144 (1999)
4. Bramble, J.H., Pasciak, J.E.: Analysis of a finite PML approximation for the three dimensional time-harmonic Maxwell and acoustic scattering problems. Math. Comput. **76**(258), 597–614 (2007)
5. Chen, Z., Zheng, W.: Convergence of the uniaxial perfectly matched layer method for time-harmonic scattering problems in two-layered media. SIAM J. Numer. Anal. **48**(6), 2158–2185 (2010)
6. Courant, R., Hilbert, D.: Methods of Mathematical Physics, vol. 1. Wiley, New York (1953)
7. Druskin, V., Güttel, S., Knizhnerman, L.: Near-optimal perfectly matched layers for indefinite Helmholtz problems. SIAM Rev. **58**(1), 90–116 (2016)
8. Druskin, V., Lieberman, C., Zaslavsky, M.: On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. SIAM J. Sci. Comput. **32**(5), 2485–2496 (2010)
9. Givoli, D.: Nonreflecting boundary conditions. J. Comput. Phys. **94**(1), 1–29 (1991)
10. Givoli, D., Hagstrom, T., Patlashenko, I.: Finite element formulation with high-order absorbing boundary conditions for time-dependent waves. Comput. Methods Appl. Mech. Eng. **195**(29–32), 3666–3690 (2006)
11. Givoli, D., Neta, B.: High-order non-reflecting boundary scheme for time-dependent waves. J. Comput. Phys. **186**(1), 24–46 (2003)
12. Givoli, D., Neta, B., Patlashenko, I.: Finite element analysis of time-dependent semi-infinite waveguides with high-order boundary treatment. Internat J. Numer. Methods Eng. **58**(13), 1955–1983 (2003)
13. Goldstein, C.I.: A finite element method for solving Helmholtz type equations in waveguides and other unbounded domains. Math. Comput. **39**(160), 309–324 (1982)
14. Güttel, S., Knizhnerman, L.: A black-box rational Arnoldi variant for Cauchy–Stieltjes matrix functions. BIT **53**(3), 595–616 (2013)
15. Hagstrom, T.: Radiation boundary conditions for the numerical simulation of waves. Acta Numer. **8**, 47–106 (1999)
16. Hagstrom, T., Mar-Or, A., Givoli, D.: High-order local absorbing conditions for the wave equation: extensions and improvements. J. Comput. Phys. **227**(6), 3322–3357 (2008)
17. Hagstrom, T., Warburton, T.: Complete radiation boundary conditions: minimizing the long time error growth of local methods. SIAM J. Numer. Anal. **47**(5), 3678–3704 (2009)
18. Hagstrom, T., Warburton, T., Givoli, D.: Radiation boundary conditions for time-dependent waves based on complete plane wave expansions. J. Comput. Appl. Math. **234**(6), 1988–1995 (2010)
19. Harari, I., Patlashenko, I., Givoli, D.: Dirichlet-to-Neumann maps for unbounded wave guides. J. Comput. Phys. **143**(1), 200–223 (1998)
20. Higdon, R.L.: Absorbing boundary conditions for difference approximations to the multidimensional wave equation. Math. Comput. **47**(176), 437–459 (1986)

21. Higdon, R.L.: Numerical absorbing boundary conditions for the wave equation. Math. Comput. **49**(179), 65–90 (1987)
22. Ingerman, D., Druskin, V., Knizherman, L.: Optimal finite difference grids and rational approximationsof the square root. I. Elliptic functions. Commun. Pure Appl. Math. **53**, 1039–1066 (2000)
23. Kim, S.: Analysis of complete radiation boundary conditions for the Helmholtz equation in perturbed waveguides. (Manuscript)
24. Kim, S.: Analysis of the convected Helmholtz equation with a uniform mean flow in a waveguide with complete radiation boundary conditions. J. Math. Anal. Appl. **410**(1), 275–291 (2014)
25. Kim, S., Pasciak, J.E.: Analysis of a Cartesian PML approximation to acoustic scattering problems in $\mathbb{R}^2$. J. Math. Anal. Appl. **370**(1), 168–186 (2010)
26. Kim, S., Zhang, H.: Optimized Schwarz method with complete radiation transmission conditions for the Helmholtz equation in waveguides. SIAM J. Numer. Anal. **53**(3), 1537–1558 (2015)
27. Kim, S., Zhang, H.: Optimized double sweep Schwarz method with complete radiation boundary conditions for the Helmholtz equation in waveguides. Comput. Math. Appl. **72**(6), 1573–1589 (2016)
28. Knizherman, L., Druskin, V., Zaslavsky, M.: On optimal convergence rate of the rational Krylov-subspace reduction for electromagnetic problems in unbounded domains. SIAM J. Numer. Anal. **47**, 953–971 (2009)
29. Koshiba, M., Tsuji, Y., Sasaki, S.: High-performance absorbing boundary conditions for photonic crystal waveguide simulations. Microw. Wirel. Compon. Lett. IEEE **11**(4), 152–154 (2001)
30. Petrushev, P., Popov, V.: Rational Approximation of Real Functions, Volume 28 of Encyclopedia of Mathematics. Cambridge University Press, Cambridge (1987)
31. Rabinovich, D., Givoli, D., Bécache, E.: Comparison of high-order absorbing boundary conditions and perfectly matched layers in the frequency domain. Int. J. Numer. Methods Biomed. Eng. **26**(10), 1351–1369 (2010)
32. Saad, Y.: Numerical Methods for Large Eigen Value Problems, vol. 66. Society for Industrial and Applied Mathematics, Philadelphia, PA (2011)
33. Schatz, A.H.: An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. Math. Comput. **28**, 959–962 (1974)
34. Strang, G.: The discrete cosine transform. SIAM Rev. **41**(1), 135–147 (1999)
35. Tsynkov, S.V.: Numerical solution of problems on unbounded domains. A review. Appl. Numer. Math. **27**(4), 465–532 (1998)