Numerical Study of the Helmholtz Equation with Large Wavenumbers

by

Michelle Michelle

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

Department of Mathematical and Statistical Sciences
University of Alberta

© Michelle Michelle, 2022

Abstract

The Helmholtz equation is a fundamental wave propagation model in the time-harmonic setting, which appears in many applications such as electromagnetics, geophysics, and ocean acoustics. It is challenging and computationally expensive to solve due to (1) its highly oscillating solution and (2) the large ill-conditioned/unstable sign-indefinite linear system arising from standard discretizations, especially when a large wavenumber is present. In this thesis, we develop and extensively study high order compact finite difference methods (FDMs) and a wavelet Galerkin method for the Helmholtz equation in various settings.

In Chapter 1, we provide some background on the Helmholtz equation and wavelets.

In Chapter 2, we introduce the Dirac Assisted Tree (DAT) method coupled with an arbitrarily high order compact 1D FDM. DAT successfully overcomes the massive ill-conditioned linear system associated with the Helmholtz equation by breaking a global problem into small much better conditioned linking and local problems, as well as harnessing parallel computing resources. DAT is effective in solving 1D heterogeneous and special 2D Helmholtz equations with arbitrarily large wavenumbers. Results in this chapter have been published in *Computers and Mathematics with Applications*.

In Chapter 3, we propose a new pollution minimizing sixth order compact FDM for the 2D Helmholtz equation with interfaces and mixed boundary conditions. The new pollution minimization strategy we employ is based on the average truncation error of plane waves. Compared to existing FDMs, the errors of our method are several orders of magnitude lower. Results in this chapter have been submitted for publication in *SIAM Journal on Scientific Computing*.

In Chapter 4, we present new sharp wavenumber-explicit stability bounds for the 2D Helmholtz equation with mixed inhomogeneous boundary conditions. Such bounds are crucial in the analysis and development of numerical schemes, since they describe how the solution behaves for given data. Establishing these bounds is difficult, since they highly de-

pend on boundary conditions and the domain's geometry. These findings motivate a future development of a numerical method, which uses our stability bounds to strategically select dominant Fourier coefficients in the solution. Results in this chapter have been accepted for publication in *SIAM Journal on Numerical Analysis*.

Wavelets are sparse multiscale representation systems built from refinable functions (i.e., functions that can be expressed as dilated and shifted versions of themselves; e.g., B-splines and Hermite splines). A Riesz wavelet on a bounded domain in $\mathbb{R}^d$ (e.g., $[0,1]^d$) is obtained from the tensor product of 1D Riesz wavelets on a bounded interval. Hence, the efficacy of a wavelet method in solving multidimensional problems (e.g., image processing and numerical PDEs) relies on the optimal construction of a wavelet basis on an interval. Many available constructions suffer from shortcomings. For example, some boundary elements may have reduced vanishing moments, which adversely impact the overall sparsity of the system. Furthermore, all existing constructions in the literature are applicable only to particular examples or a very narrow family of wavelet bases. A natural question is whether a systematic construction procedure for any compactly supported wavelet basis exists. In Chapter 5, we fully answer this long-standing problem in wavelet analysis. We propose and study two systematic approaches that construct all locally supported biorthogonal multiwavelets on an interval from any compactly supported biorthogonal multiwavelets on $\mathbb{R}$. Results in this chapter have been published in *Applied and Computational Harmonic Analysis*.

In Chapter 6, we apply the direct approach in the previous chapter to construct 1D Riesz wavelets on the unit interval, and subsequently a 2D Riesz wavelet on the unit square via tensor product. The latter is used in the Galerkin scheme to solve the 2D Helmholtz equation with a non-local boundary condition, which models electromagnetic scattering from a large cavity. The implementation of our method is very efficient. Also, our numerical experiments indicate that the coefficient matrix of our wavelet Galerkin method is much better conditioned (i.e., much more stable) than that of a standard Galerkin method.

Preface

Results in Chapter 2 are based on the published journal article "Dirac assisted tree method for 1D heterogeneous Helmholtz equations with arbitrary variable wave numbers. *Computers and Mathematics with Applications* **97** (2021), 416-438." The development and composition of results in Chapter 2 are joint work with Bin Han and Yau Shu Wong.

Results in Chapter 3 are based on the paper "Sixth order compact finite difference method for 2D Helmholtz equations with singular sources and reduced pollution effect," (arXiv:2112.07154v1, 20 pages), which has been submitted for publication in *SIAM Journal on Scientific Computing* and is currently under review. The development and composition of results in Chapter 3 are joint work with Qiwei Feng and Bin Han.

Results in Chapter 4 are based on the paper "Sharp wavenumber-explicit stability bounds for 2D Helmholtz equations," (arXiv:2108.06469, 28 journal pages), which has been accepted for publication in *SIAM Journal on Numerical Analysis*. Results in Chapter 5 are based on the published journal article "Wavelets on intervals derived from arbitrary compactly supported biorthogonal multiwavelets. *Applied and Computational Harmonic Analysis* **53** (2021), 270-331." Results in Chapter 6 are currently in the manuscript preparation stage. The development and composition of results in Chapters 4 to 6 are joint work with Bin Han.

Acknowledgements

First, I would like to thank my supervisor Dr. Bin Han without whom this thesis would not have come to fruition. I am extremely grateful for his advice, generosity, guidance, and immense support at every stage of my program. Without his help, my journey as a graduate student would not have gone as smoothly as it did. His attitude towards research, deep understanding of mathematics, and hard work are truly inspiring. I have learned so much about effective communication, mathematics, and research from him. I also would like to thank my co-supervisor Dr. Yau Shu Wong for his very kind support and introducing me to the Helmholtz equation. I sincerely appreciate the insights he provided on this subject. Furthermore, I am very thankful for the opportunity to work with him on two interesting industrial projects that well complemented my academic studies. I also would like to thank Dr. Peter Minev for being a part of my thesis supervisory committee and for his valuable suggestions that are critical in the formation of this thesis. My deepest gratitude also goes to Dr. Feng Dai and Dr. Jie Shen for their willingness to be my thesis examiners and their insightful comments.

Second, I would like to thank Alberta Innovates and Alberta Advanced Education, Compute Canada Calcul Canada, the University of Alberta's Department of Mathematical and Statistical Sciences, the University of Alberta's Faculty of Graduate Studies and Research (FGSR), Natural Sciences and Engineering Research Council of Canada (NSERC), and Westgrid for the funding and resources they have given me over the course of my program.

Third, I would like to thank my mom, brother, and aunt for their love, sacrifice, and unwavering support. To all my friends, thank you for your company, encouragement, and for brightening my days; special thanks to Dasha for her constant support, witty sense of humour, and of course, for the much needed weekend food trips.

Above all, I would like to thank the Almighty God for His grace, strength, wisdom, and blessings in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

The Helmholtz equation is a fundamental time-harmonic wave propagation model, which appears in many applications such as electromagnetism [12, 108], geophysics [21, 36, 44, 55], ocean acoustics [87], and photonic crystals [56]. Its derivation from fundamental relations in physics is discussed for example in [86]. This equation is challenging, and by the same token, fascinating to study for a few reasons.

From the theoretical point of view, the analysis of the Helmholtz equation is involved because its standard weak formulation is non-coercive. Two fundamental topics to study are the stability of the underlying solution (also known as a priori bounds) and the error/convergence of a numerical scheme. A nice exposition of special techniques used for these topics can be found in [111]. The former of the two describes how the solution behaves when the boundary and source data are perturbed. Wavenumber-explicit stability bounds are of particular interest, since they describe how the solution's energy depends on the frequency and serve as a foundation for the development of numerical schemes. Several stability bounds for the Helmholtz equation in various settings are available: [21, 32, 47, 52, 103, 118] for the interior impedance Helmholtz equation, [7, 17, 61, 63, 83] for the interior Helmholtz equation with mixed boundary conditions, [12, 13, 41, 98] for an electromagnetic scattering from a large cavity problem, [107] for the Helmholtz equation with transmissions, and [60, 111, 116, 117, 118] for the exterior Helmholtz equation. More studies can be found in the references cited by the previous papers. In general, obtaining a sharp wavenumber-explicit stability bound is very challenging, since they are highly dependent on the domain and boundary configurations (e.g, [47]). This can be seen from the studies done by [47, 103, 118], which deal with a bounded Lipschitz domain. With an extra assumption that the domain is star-shaped with respect to a ball, [103] proved a stability estimate that is independent

of the wavenumber κ. Without this assumption, the stability bound in [118, Theorem 1.6] has extra factors $\kappa^{1/2}$ and κ in front of the boundary and source data respectively. We shall address a stability problem for the Helmholtz equation in Chapter 4.

From the computational point of view, the Helmholtz equation is difficult to solve due to its highly oscillating solution. To obtain a reasonable solution or observe a convergent behavior, the mesh size h used in a standard discretization has to be much smaller than the reciprocal of the wavenumber κ. Moreover, the mesh size requirement becomes exponentially demanding as the wavenumber increases. This is known as the pollution effect, which has close ties to the numerical dispersion/phase lags. The situation is worsened by the fact that many discretizations of the Helmholtz equation yield an ill-conditioned sign-indefinite coefficient matrix. Thus, when solving the Helmholtz equation, one often faces a massive ill-conditioned sign-indefinite linear system, where standard iterative schemes may fail [45].

To gain a better insight on how the mesh size requirement is related to the wavenumber, we recall some relevant findings on the finite element method (FEM) and finite difference method (FDM). The authors in [105] considered the interior impedance problem and discovered that the quasi-optimality in the hp-FEM setting can be achieved by choosing a polynomial degree p and a mesh size h such that $p \geqslant C \log(\kappa)$ (for some positive C independent of κ, h, p) and $\kappa h/p$ is small enough. The authors in [43] found that for sufficiently small $\kappa^{2p+1}h^{2p}$, the leading pollution term in an upper bound of the standard Sobolev H^1-norm is $\kappa^{2p+1}h^{2p}$. For second order FDMs, [22, 23] found that $\kappa^3 h^2 \leqslant C$ (for some positive C independent of κ, h) is required to obtain a reasonable solution. Meanwhile, for the fourth order FDM, [36] found that $\kappa^5 h^4 \leqslant C$ (for some positive C independent of κ, h) is required to obtain a reasonable solution.

A lot of research effort has been invested in developing ways to tackle these discretization-related issues. From the previous discussion, it is clear that the mesh size requirement for high order schemes is less stringent than low order ones. Hence, high order schemes are typically of interest. Various preconditioners and domain decomposition methods have been developed over the years (see the review paper [57] and references therein). Many variants of FEM/Galerkin/variational methods have been explored. For example, [53, 54] relaxed the inter-element continuity condition and imposed penalty terms on jumps across the element edges. These penalty terms can be tuned to reduce the pollution effect. Spectral methods have been used to solve the Helmholtz equation in various settings [48, 81, 82, 85, 98, 109, 115, 116, 117]. A class of Trefftz methods, where the trial and test functions consist of local solutions to the underlying (homogeneous) Helmholtz equation, were considered in [84] and references therein. A closely related method, called the generalized FEM or the partition of unity FEM, has been explored. It involves multiplying solutions to the homogeneous

Helmholtz equation (e.g. plane waves) with elements of a chosen partition of unity, which then serve as the trial and test functions. In recent years, multiscale FEM has also become an appealing alternative to deal with the pollution effect [112]. While it is widely accepted that the pollution effect in standard discretizations such as FEMs and FDMs cannot be eliminated for 2D and higher dimensions [9], we can obtain pollution free FDMs [77, 122] in 1D to solve special 2D Helmholtz equations [77, 123]. FDMs of various orders and stencil sizes have been proposed. For example, [22, 23] proposed second order FDMs, [110] proposed a third order FDM, [15, 16, 36] proposed fourth order FDMs, and [125, 126] proposed sixth order FDMs. The number of points used in these schemes varies: 9 in [22, 125], 13 in [37], and both 17 and 25 in [36]. Additionally, the schemes in [22, 23, 36, 119, 125] share a similar dispersion minimization strategy. They start with a stencil having a given accuracy order with some free parameters. Afterwards, the plane wave solution is inserted into the scheme and the ratio between the true and numerical wavenumbers is minimized by forming an overdetermined linear system with respect to a set of discretized angles and a range of $\frac{2\pi}{\kappa h}$ (i.e., the number of points per wavelength). We shall present and discuss new FDMs for the Helmholtz equation in Chapters 2 and 3.

We briefly list some notations used in the discussion of the Helmholtz equation. Throughout this thesis, we shall refer to domain as an open connected set in $\mathbb{R}^d$, where $d = 1, 2$. Suppose Ω is a bounded Lipschitz domain. Let $H^s(\Omega)$, where $s \geqslant 0$, be the classical Sobolev spaces of order s (e.g., see [62, Sections 1.3.1-1.3.2]), whose norm is denoted by $\|\cdot\|_{s,\Omega}$. Suppose Γ is the boundary of Ω. Let $H^s(\Gamma)$, where $s \geqslant 0$, be the Sobolev spaces of order s on the boundary defined in the usual sense (e.g., see [101, pages 96-99]). If $s = 0$, then $H^0(\Omega) = L_2(\Omega)$ and $H^0(\Gamma) = L_2(\Gamma)$. The standard inner product and norm in $L_2(\Omega)$ are denoted by $\langle u, v\rangle_\Omega := \int_\Omega u\overline{v}$ and $\|\cdot\|_{0,\Omega} := \langle u, u\rangle_\Omega^{1/2}$. On the boundary, the standard inner product and norm in $L_2(\Gamma)$ are denoted by $\langle u, v\rangle_\Gamma := \int_\Gamma u\overline{v}$ and $\|\cdot\|_{0,\Gamma} := \langle u, u\rangle_\Gamma^{1/2}$. In this thesis, we also use $\langle\cdot,\cdot\rangle$ (without any subscripts) to denote the standard inner product, where the domain of integration is clear from the context.

One of our proposed numerical methods employs wavelet bases as its primary approximation tool. We now recall some basic concepts and definitions related to wavelets. Wavelets are sparse multiscale representation systems, which have been successfully used in various applications such as data science, image/signal processing, and numerical analysis. They have been used to characterize various function spaces such as Sobolev and Besov spaces. Wavelets are built from refinable functions, which are functions that can be expressed as scaled and shifted versions of themselves. A good example of refinable functions is the

B-spline function of order m (i.e, B_m for $m \in \mathbb{N}$), where

$$B_1 := \chi_{(0,1]} \quad \text{and} \quad B_m := B_{m-1} * B_1 = \int_0^1 B_{m-1}(\cdot - x)dx. \tag{1.1}$$

Note that $B_m \in C^{m-2}(\mathbb{R})$, its support is $[0, m]$, and $B_m|_{(k,k+1)}$ is a nonnegative polynomial of degree $m-1$ for all $k \in \mathbb{Z}$. Another great example is the Hermite cubic splines. Next, we present several formal definitions to facilitate an in-depth discussion of the topic. Note that the Sobolev space $H^\tau(\mathbb{R})$ with $\tau \in \mathbb{R}$ consists of all tempered distributions f on $\mathbb{R}$ such that $\int_{\mathbb{R}} |\widehat{f}(\xi)|^2(1+|\xi|^2)^\tau d\xi < \infty$. Let $\phi := \{\phi_1, \dots, \phi_r\}^\mathsf{T}, \psi := \{\psi_1, \dots, \psi_s\}^\mathsf{T} \in H^\tau(\mathbb{R})$ with $\tau \in \mathbb{R}$. For $J \in \mathbb{Z}$, define *the multiwavelet affine system* in $H^\tau(\mathbb{R})$ by

$$\mathsf{AS}_J^\tau(\phi;\psi) := \left\{\phi_{J;k}^\ell : k \in \mathbb{Z}, 1 \leqslant \ell \leqslant r\right\} \cup \left\{\psi_{j;k}^\ell : j \geqslant J, k \in \mathbb{Z}, 1 \leqslant \ell \leqslant s\right\},$$

where $\phi_{J;k}^\ell := 2^{J(1/2-\tau)}\phi_\ell(2^J \cdot -k)$ and $\psi_{j;k}^\ell := 2^{j(1/2-\tau)}\psi_\ell(2^j \cdot -k)$. We say that $\{\phi;\psi\}$ is *a Riesz multiwavelet in* $H^\tau(\mathbb{R})$ if $\mathsf{AS}_J^\tau(\phi;\psi)$ is *a Riesz basis for* $H^\tau(\mathbb{R})$. I.e., (1) the linear span of $\mathsf{AS}_J^\tau(\phi;\psi)$ is dense in $H^\tau(\mathbb{R})$, and (2) there exist $C_1, C_2 > 0$ such that

$$C_1 \sum_{\eta \in \mathsf{AS}_J^\tau(\phi;\psi)} |c_\eta|^2 \leqslant \| \sum_{\eta \in \mathsf{AS}_J^\tau(\phi;\psi)} c_\eta \eta \|_{H^\tau(\mathbb{R})}^2 \leqslant C_2 \sum_{\eta \in \mathsf{AS}_J^\tau(\phi;\psi)} |c_\eta|^2$$

for all finitely supported sequences $\{c_\eta\}_{\eta \in \mathsf{AS}_J^\tau(\phi;\psi)}$. Note that if $\tau = 0$, then $H^0(\mathbb{R}) = L_2(\mathbb{R})$. Throughout this thesis, we let $\mathsf{AS}_J(\phi;\psi)$ to denote $\mathsf{AS}_J^0(\phi;\psi)$. By a simple scaling argument, it is easy to see ([70, 71]) that $\mathsf{AS}_J^\tau(\phi;\psi)$ is a Riesz basis of $H^\tau(\mathbb{R})$ for some $J \in \mathbb{Z}$ if and only if $\mathsf{AS}_J^\tau(\phi;\psi)$ is a Riesz basis of $H^\tau(\mathbb{R})$ for all $J \in \mathbb{Z}$. If $\mathsf{AS}_0(\phi;\psi)$ is an orthonormal basis of $L_2(\mathbb{R})$, then $\{\phi;\psi\}$ is called *an orthogonal multiwavelet in* $L_2(\mathbb{R})$. It immediately follows that an orthogonal multiwavelet $\{\phi;\psi\}$ is a Riesz multiwavelet in $L_2(\mathbb{R})$. If $r = 1$, the above vector refinable function becomes a scalar refinable function, and a Riesz multiwavelet in this case is often referred to as a scalar Riesz wavelet. Throughout this thesis, we often refer to scalar wavelets and multiwavelets as simply wavelets. Let $\tilde{\phi} := \{\tilde{\phi}_1, \dots, \tilde{\phi}_r\}^\mathsf{T}, \tilde{\psi} := \{\tilde{\psi}_1, \dots, \tilde{\psi}_s\}^\mathsf{T} \in H^{-\tau}(\mathbb{R})$ with $\tau \in \mathbb{R}$. We call $(\{\tilde{\phi};\tilde{\psi}\}, \{\phi;\psi\})$ *a biorthogonal multiwavelet* in $(H^{-\tau}(\mathbb{R}), H^\tau(\mathbb{R}))$ if (1) $\{\tilde{\phi};\tilde{\psi}\}$ and $\{\phi;\psi\}$ are Riesz multiwavelets in $H^{-\tau}(\mathbb{R})$ and $H^\tau(\mathbb{R})$ respectively, and (2) $\mathsf{AS}_J^{-\tau}(\tilde{\phi};\tilde{\psi})$ and $\mathsf{AS}_J^\tau(\phi;\psi)$ are biorthogonal to each other; i.e.,

$$\langle h, \tilde{h}\rangle = 1 \quad \text{and} \quad \langle h, \tilde{g}\rangle = 0, \quad \forall g \in \mathsf{AS}_J^\tau(\phi;\psi)\backslash\{h\}.$$

Moreover, the following wavelet representations hold

$$\begin{aligned} f &= \sum_{k\in\mathbb{Z}}\sum_{\ell=1}^{r}\langle f,\tilde{\phi}^{\ell}_{J;k}\rangle\phi^{\ell}_{J;k}+\sum_{j=J}^{\infty}\sum_{k\in\mathbb{Z}}\sum_{\ell=1}^{s}\langle f,\tilde{\psi}^{\ell}_{j;k}\rangle\psi^{\ell}_{j;k}, \quad f\in H^{\tau}(\mathbb{R}),\\ g &= \sum_{k\in\mathbb{Z}}\sum_{\ell=1}^{r}\langle g,\phi^{\ell}_{J;k}\rangle\tilde{\phi}^{\ell}_{J;k}+\sum_{j=J}^{\infty}\sum_{k\in\mathbb{Z}}\sum_{\ell=1}^{s}\langle g,\psi^{\ell}_{j;k}\rangle\tilde{\psi}^{\ell}_{j;k}, \quad g\in H^{-\tau}(\mathbb{R}), \end{aligned} \tag{1.2}$$

with the above series converging unconditionally in $H^{\tau}(\mathbb{R})$ and $H^{-\tau}(\mathbb{R})$ respectively. For a compactly supported (vector) function ψ, we say that ψ has m *vanishing moments* if $\int_{\mathbb{R}} x^{j}\psi(x)dx = 0$ for all $j = 0,\dots,m-1$. Furthermore, we define $\mathrm{vm}(\psi) := m$ with m being the largest of such an integer. There are a few reasons as to why it is advantageous to use wavelets in our numerical scheme. First, the coefficient matrix is often well-conditioned and has uniformly bounded condition numbers, since we use a Riesz wavelet that belongs to a certain Sobolev space. Second, the coefficient matrix can be assembled and stored efficiently due to the refinable structure and the sparsity of the system. Third, they can be customized to a certain extent to exploit the underlying structure of the problem. For some studies on wavelet based numerical schemes, see [20, 39, 74, 88, 90, 92]. A Riesz wavelet on a bounded domain can be obtained by taking the tensor product of univariate Riesz wavelets. Hence, the efficacy of a wavelet method in solving multidimensional problems relies on the optimal construction of a Riesz wavelet on a bounded interval. It turns out that many constructions in the literature can only be applied to specific examples or a small family of wavelets. In practice, this of course would limit the kinds of wavelets we can use in a numerical scheme. In the following, we list relevant studies to give more context on the current state of literature. The authors in [4, 6, 25, 31, 94, 99, 102, 113] constructed scalar orthogonal wavelets on the unit interval from Daubechies orthogonal wavelets. The authors in [18, 19, 20, 34, 35, 88, 100, 114] constructed spline scalar biorthogonal wavelets, which are built from B-splines (see [29]), on the unit interval. The semi-orthogonal spline wavelets in [24] have also been adapted to the unit interval in [25]. Even though multiwavelets generally can have higher vanishing moments and smoothness than scalar wavelets for a given support, the studies on their constructions are significantly fewer due to their technicalities [2, 5, 18, 33, 39, 71, 72, 73, 80, 96]. In addition to the previous issue, some constructions produce boundary elements are not as optimal as they can be. For example, these elements may have reduced vanishing moments, which negatively impact the system's overall sparsity, and may have a long support. Given the current state of literature and our goal of using a wavelet Galerkin method to solve the Helmholtz equation, we shall first address this critical construction issue in Chapter 5 and then present our proposed numerical scheme in Chapter 6.

1.2 Thesis structure and contributions

The main focus of this thesis is to develop and extensively study high order compact FDMs and a wavelet Galerkin method for the Helmholtz equation in various settings. In this section, we detail the major contributions of each chapter.

In Chapter 2, we introduce the Dirac Assisted Tree (DAT) method to solve the 1D heterogeneous and special 2D Helmholtz equations with arbitrarily large wavenumbers. The key idea of DAT is to break the global problem on $[0, 1]$ into many small local problems by multiplying the source term with a partition of unity (for example B-splines; here, we use hat linear functions B_2). Each local problem now has a highly localized source term with homogeneous Dirichlet boundary conditions except for those that touch the endpoints. Such local problems can be solved by any discretization method in a parallel fashion, provided that the fluxes at the endpoints of local problems can be accurately computed. As we shall see later, Dirac distributions will naturally appear in the fluxes of local solutions. Hence, to link these local probems, we want to offset these Dirac distributions by solving small linking problems also in a parallel fashion. Due to the refinability of the hat function, the local problems can be further decomposed into sub-local problems. Applying this recursively, we obtain the tree structure of DAT. DAT can solve heterogeneous Helmholtz equations that other methods have problems handling, without having any magnitude constraints on variable wavenumbers, and without dealing with a massive ill-conditioned coefficient matrix. In particular, we harness parallel computing resources to solve small and much better conditioned linear systems coming from the local and linking problems. In an extreme case, each of the local and linking problems in DAT can be solved by at most 4 linear equations. The accuracy of DAT only depends on the accuracy of the local problem solvers. If the maximum size of all local problems is bounded and independent of mesh sizes, then all coefficient matrices of the local problems have uniformly bounded condition numbers. DAT naturally brings about domain decomposition and adaptivity. The second contribution of this chapter is to present a 1D compact FDM (that handles the domain's interior and any mixed boundary conditions) with an arbitrarily high accuracy order by assuming that a, κ^2, f are piecewise smooth. Such high order compact FDMs are particularly appealing for DAT, especially when computing the fluxes and derivatives of the solutions to local and linking problems.

In Chapter 3, we propose a sixth order compact finite difference scheme with reduced pollution to solve the 2D Helmholtz equation with singular sources and mixed boundary conditions on a rectangular domain. Our proposed compact finite difference scheme attains the maximum overall accuracy order everywhere on the domain with the shortest stencil support

for the problem of interest. Similar to [49, 50], our approach is based on a critical observation regarding the inter-dependence of high order derivatives of the underlying solution. When constructing a discretization stencil, we start with a general expression that allows us to recover all possible sixth order finite difference schemes. Then, we determine the remaining free parameters in the stencil by using our new pollution minimization strategy that is based on the average truncation error of plane waves. Our method differs from existing dispersion minimization methods in the literature in several ways. First, our method does not require us to compute the numerical wavenumber. Second, we use our pollution minimization procedure in the construction of all interior, boundary, and corner stencils. This is in stark contrast to the common approach in the literature, where the dispersion is minimized only in the interior stencil. The effectiveness of our pollution minimization strategy is evident from our numerical experiments. Our proposed compact finite difference scheme with reduced pollution effect outperforms several state-of-the-art finite difference schemes in the literature, particularly in the pre-asymptotic critical region where κh is near 1. When a large wavenumber κ is present, this means that our proposed finite difference scheme is more accurate than others at a computationally feasible grid size. We also provide a comprehensive treatment of mixed inhomogeneous boundary conditions. In particular, our approach is capable of handling all possible combinations of Dirichlet, Neumann, and impedance boundary conditions for the 2D Helmholtz equation defined on a rectangular domain. For each corner, we explicitly provide a 4-point stencil with at least sixth order accuracy and reduced pollution effect. For each side, we explicitly give a 6-point stencil with at least sixth order accuracy and reduced pollution effect. To the best of our knowledge, our work is the first to comprehensively study the construction of corner and boundary finite difference stencils for all possible combinations of boundary conditions on a rectangular domain. Unlike the common technique used in the literature, no ghost or artificial points are introduced in our construction. We derive a seventh order compact finite difference scheme to handle nonzero jump functions along the interface curve (i.e., the singular source). Since our proposed finite difference scheme is compact, the linear system arising from the discretization is sparse. The stencils themselves have a nice structure in that their coefficients are symmetric and take the form of polynomials of κh. Also, the coefficients in our interior stencil are simpler compared to [28], as they are polynomials of degree 6, while those in [28] are of degree 16.

In Chapter 4, we study the stability of the 2D Helmholtz equation on a rectangular domain with inhomogeneous boundary conditions and derive several new sharp wavenumber-explicit stability bounds that hold for all positive wavenumber κ. By sharp, we mean that our stability bounds capture the leading κ-dependent term in front of the norm of a given datum, which is accurate up to a constant multiple (independent of κ and the given datum).

To this end, we shall devise Fourier techniques, the Rellich's identity, and a lifting strategy. Furthermore, we shall give several examples to illustrate the sharpness of our stability bounds. Some of the above boundary configurations can be thought of as simplified electromagnetic scattering from a large cavity (e.g, see [3, 11, 12, 13, 98]) or waveguide (e.g., see [14, 63, 106]) problems, where we approximate the non-local inhomogeneous boundary condition with an impedance boundary condition. Even though the stability bounds in [83] hold for a rectangular domain with mixed boundary conditions, the boundary configurations in [83] are completely different from ours. As we shall see later on, our stability bounds necessarily depend on the wavenumber unlike [83]. This again highlights the sensitivity of stability bounds with respect to boundary placements and conditions. Our results complement those in [83] for a rectangular domain, thereby offering a much more complete picture of the stability behaviour of the Helmholtz equation on a rectangular domain with mixed boundary conditions. These stability bounds are indeed applicable to the model problem in Chapter 3 without any interface. While the scheme proposed in Chapter 3 can indeed be applied to the present chapter's model problem, our stability bounds motivate a future development of a numerical method, which uses them to strategically select dominant Fourier coefficients in the solution.

In Chapter 5, we do an in-depth study on the construction of wavelets on a bounded interval from arbitrary compactly supported multiwavelets. This chapter serves as the foundation for Chapter 6, where we present a wavelet Galerkin method to solve a scattering problem. In this chapter, we present classical and direct approaches to construct all possible compactly supported biorthogonal wavelets on a bounded intervals, which satisfy prescribed vanishing moments and homogeneous boundary conditions. This chapter fully answers the long-standing question in wavelet analysis on the existence of a general construction procedure that works for any compactly supported biorthogonal wavelets. Furthermore, our construction does not suffer from any shortcomings that others do. That is, we are able to maximally preserve the original desirable properties of the wavelets on the real line (e.g., short support and maximum vanishing moments). In the classical approach, we generalize the construction method from scalar wavelets to multiwavelets. The typical procedure is to first construct primal and dual refinable functions, and only after derive the corresponding primal and dual wavelets. However, this calculation is often complicated because the dual parts often have much longer support. This motivates us to propose the direct approach, which is remarkably more general and much simpler to use than the former, since the dual parts are not explicitly involved. Due to its convenience, we shall predominantly use the later method throughout this thesis.

In Chapter 6, we present our wavelet Galerkin method for solving an electromagnetic

scattering from a large cavity problem. We shall apply the direct approach in Chapter 5 to a new compactly supported biorthogonal wavelet, whose primal refinable function is interpolating [42], to form a 2D Riesz wavelet in an appropriate Sobolev space via tensor product. Our method falls in the category of high order schemes with a natural preconditioner originating from our wavelet basis. We expect that these two features help in alleviating the mesh size requirement and improving the condition/stability of the linear system. Our numerical experiments indicate that the condition number of the coefficient matrix associated with our wavelet Galerkin method is approximately 2 to 800 times smaller (i.e., more stable) than that of the standard Galerkin method.

Finally, we outline some directions of some future work in Chapter 7.

Chapter 2

Dirac Assisted Tree (DAT) Method

We have described some key challenges of the Helmholtz equation in Chapter 1. To tackle them, we present our first numerical method, which is the DAT method with an arbitrarily high order compact FDM as its solver. It efficiently solves 1D heterogeneous and special 2D Helmholtz equations by decomposing the original problem into small better conditioned local and linking problems.

In particular, we consider the following model problem. Let $\mathcal{L}$ be the linear differential operator of the 1D heterogeneous Helmholtz equations as follows

$$\mathcal{L}u := [a(x)u'(x)]' + \kappa^2(x)u(x) = f(x), \qquad x \in \Omega := (0,1) \tag{2.1}$$

with any given linear boundary conditions

$$\mathcal{B}_0 u(0) := \lambda_0^L u(0) + \lambda_1^L u'(0) = g_0, \qquad \mathcal{B}_1 u(1) := \lambda_0^R u(1) + \lambda_1^R u'(1) = g_1, \tag{2.2}$$

where $\lambda_0^L, \lambda_1^L, \lambda_0^R, \lambda_1^R \in \mathbb{C}$ satisfy $|\lambda_0^L| + |\lambda_1^L| \neq 0$ and $|\lambda_0^R| + |\lambda_1^R| \neq 0$. I.e., $\mathcal{B}_0 u(0)$ and $\mathcal{B}_1 u(1)$ can be Dirichlet, Neumann or Robin (e.g. Sommerfeld) boundary conditions.

We describe the key ingredients and algorithm for the DAT method in Section 2.1. In Section 2.2, we present a 1D compact FDM with an arbitrarily high accuracy order for 1D heterogeneous Helmholtz equation in (2.1) with piecewise smooth coefficients, wavenumbers, and source terms. We discuss the convergence of DAT in Section 2.3. Numerical experiments showcasing our method are presented in Section 2.4.

Results in this chapter are based on [77].

2.1 Main ideas and algorithm

Let $(\alpha,\beta) \subseteq (0,1)$ with $0 \leqslant \alpha < \beta \leqslant 1$. Let $f \in (H^1(\alpha,\beta))'$ be a source term. Let $u_{loc} \in H^1(\alpha,\beta)$ be the weak solution in the Sobolev space $H^1(\alpha,\beta)$ to the following local problem:

$$\mathcal{L}u_{loc}(x) = f(x), \qquad x \in (\alpha,\beta), \tag{2.3}$$

where if $\alpha,\beta \in \{0,1\}$, then we preserve the boundary conditions as in (2.2); otherwise, we use homogeneous Dirichlet boundary conditions. Putting these boundary conditions into a compact form, the boundary conditions to (2.3) are given by

$$(\mathcal{B}_0 u_{loc}(0) - g_0)\delta_{0,\alpha} + (1-\delta_{0,\alpha})u_{loc}(\alpha) = 0, \quad (\mathcal{B}_1 u_{loc}(1) - g_1)\delta_{1,\beta} + (1-\delta_{1,\beta})u_{loc}(\beta) = 0, \tag{2.4}$$

where $\delta_{c,c} = 1$ and $\delta_{c,d} = 0$ for $c \neq d$. Recall that $\psi \in H^1(\alpha,\beta)$ if $\psi \in L_2(\alpha,\beta)$ and its weak/distributional derivative $\psi' \in L_2(\alpha,\beta)$. Moreover, $\|\psi\|^2_{H^1(\alpha,\beta)} := \|\psi\|^2_{L_2(\alpha,\beta)} + \|\psi'\|^2_{L_2(\alpha,\beta)}$, where ψ' stands for the weak derivative of ψ. Due to (2.4), we can extend $u_{loc} \in H^1(\alpha,\beta)$ as an element in $H^1(0,1)$ by zero extension, which is denoted by $\tilde{u}_{loc}$. Therefore, using the definition of $\mathcal{L}$ in (2.1), we observe that

$$\tilde{f} := \mathcal{L}\tilde{u}_{loc} = \begin{cases} 0, & x \in (0,\alpha) \cup (\beta,1), \\ d_\alpha(\tilde{u}_{loc})\delta_\alpha, & x = \alpha \text{ and } \alpha \neq 0, \\ f(x), & x \in (\alpha,\beta), \\ d_\beta(\tilde{u}_{loc})\delta_\beta, & x = \beta \text{ and } \beta \neq 1, \end{cases} \tag{2.5}$$

where δ_α is the Dirac distribution at the point α and the above numbers $d_\alpha(\tilde{u}_{loc}), d_\beta(\tilde{u}_{loc}) \in \mathbb{C}$ for $\alpha \neq 0$ and $\beta \neq 1$ are given by

$$d_\alpha(\tilde{u}_{loc}) := \lim_{x\to\alpha^+} a(x)\tilde{u}'_{loc}(x), \qquad d_\beta(\tilde{u}_{loc}) := -\lim_{x\to\beta^-} a(x)\tilde{u}'_{loc}(x), \tag{2.6}$$

which, up to a sign change, are simply the fluxes of $\tilde{u}_{loc}$ at α and β. Then $\tilde{u}_{loc}$ is a global solution of

$$\mathcal{L}\tilde{u}_{loc}(x) = \tilde{f}(x), \qquad x \in \Omega = (0,1).$$

We now introduce the DAT method. Let $N_0 \in \mathbb{N}$ be a positive integer greater than one. We take a partition of unity $\{\varphi_j\}_{j=0}^{N_0}$ of piecewise smooth functions such that each function φ_j is supported on $[0,1]$ and $\sum_{j=0}^{N_0} \varphi_j(x) = 1$ for all $x \in (0,1)$. For simplicity, we use piecewise linear hat functions φ_j. Let $0 = x_0 < \cdots < x_{N_0} = 1$ be a partition of $[0,1]$. For simplicity,

we define $x_{-1} = 0$ and $x_{N_0+1} := 1$. We let φ_j be the linear hat function supported on $[x_{j-1}, x_{j+1}]$ with $\varphi_j(x_j) = 1$ and $\varphi_j(x_{j-1}) = \varphi_j(x_{j+1}) = 0$. Obviously, we define $\varphi_0(x_0) = 1$ and $\varphi_0(x_1) = 0$, while $\varphi_{N_0}(x_{N_0}) = 1$ and $\varphi_{N_0}(x_{N_0-1}) = 0$. We now partition the original source function f into small pieces as follows:

$$f_j(x) := f(x)\varphi_j(x), \qquad j = 0, \ldots, N_0.$$

Since $\sum_{j=0}^{N_0} \varphi_j(x) = 1$ for all $x \in (0,1)$, we have $f = \sum_{j=0}^{N_0} f_j$. Let $u_j \in H^1(x_{j-1}, x_{j+1})$ be the weak solution to the regular local problem:

$$\mathcal{L}u_j(x) = f_j(x), \qquad x \in (x_{j-1}, x_{j+1}) \tag{2.7}$$

with the following boundary conditions:

$$\begin{aligned} &(\mathcal{B}_0 u_j(0) - g_0\delta_{0,j})\delta_{0,x_{j-1}} + (1 - \delta_{0,x_{j-1}})u_j(x_{j-1}) = 0, \\ &(\mathcal{B}_1 u_j(1) - g_1\delta_{N_0,j})\delta_{1,x_{j+1}} + (1 - \delta_{1,x_{j+1}})u_j(x_{j+1}) = 0. \end{aligned} \tag{2.8}$$

That is, we use the homogeneous Dirichlet boundary conditions $u_j(x_{j-1}) = u_j(x_{j+1}) = 0$, except $\mathcal{B}_0 u_0(0) = g_0$, $\mathcal{B}_0 u_1(0) = 0$, $\mathcal{B}_1 u_{N_0}(1) = g_1$, and $\mathcal{B}_1 u_{N_0-1}(1) = 0$. Due to (2.8), we can extend $u_j \in H^1(x_{j-1}, x_{j+1})$ as an element in $H^1(0,1)$ by zero extension, which is denoted by $\tilde{u}_j$. Hence,

$$\tilde{f}_j(x) := \mathcal{L}\tilde{u}_j(x) = \begin{cases} 0, & x \in (0, x_{j-1}) \cup (x_{j+1}, 1), \\ d_{x_{j-1}}(\tilde{u}_j)\delta_{x_{j-1}}, & x = x_{j-1} \text{ and } x_{j-1} \neq 0, \\ f_j(x), & x \in (x_{j-1}, x_{j+1}), \\ d_{x_{j+1}}(\tilde{u}_j)\delta_{x_{j+1}}, & x = x_{j+1} \text{ and } x_{j+1} \neq 1. \end{cases} \tag{2.9}$$

Now we discuss how to link/stitch all these local solutions $\{\tilde{u}_j\}_{j=0}^{N_0}$ together. To do so, for $j = 1, \ldots, N_0 - 1$, we solve the following Dirac assisted local problem:

$$\mathcal{L}v_j(x) = \delta_{x_j}, \qquad x \in (x_{j-1}, x_{j+1}) \tag{2.10}$$

with the following boundary conditions:

$$\mathcal{B}_0 v_j(0)\delta_{0,x_{j-1}} + (1 - \delta_{0,x_{j-1}})v_j(x_{j-1}) = 0, \quad \mathcal{B}_1 v_j(1)\delta_{1,x_{j+1}} + (1 - \delta_{1,x_{j+1}})v_j(x_{j+1}) = 0. \tag{2.11}$$

That is, we use the homogeneous Dirichlet boundary condition $v_j(x_{j-1}) = v_j(x_{j+1}) = 0$, except $\mathcal{B}_0 v_1(0) = 0$ and $\mathcal{B}_1 v_{N_0-1}(1) = 0$. As explained before, due to (2.11), we can extend

$v_j \in H^1(x_{j-1}, x_{j+1})$ as an element in $H^1(0,1)$ by zero extension, which is denoted by $\tilde{v}_j$. So, we must have

$$\tilde{\delta}_{x_j}(x) := \mathcal{L}\tilde{v}_j(x) = \begin{cases} 0, & x \in (0, x_{j-1}) \cup (x_{j+1}, 1), \\ d_{x_{j-1}}(\tilde{v}_j)\delta_{x_{j-1}}, & x = x_{j-1} \text{ and } x_{j-1} \neq 0, \\ \delta_{x_j}, & x \in (x_{j-1}, x_{j+1}), \\ d_{x_{j+1}}(\tilde{v}_j)\delta_{x_{j+1}}, & x = x_{j+1} \text{ and } x_{j+1} \neq 1. \end{cases} \tag{2.12}$$

To link all the local solutions $\{\tilde{u}_j\}_{j=0}^{N_0}$ together, we need the following result.

Theorem 2.1. *The elements in $\{\tilde{v}_1, \ldots, \tilde{v}_{N_0-1}\}$ are linearly independent and for any complex numbers $\mu_j, j = 1, \ldots, N_0 - 1$, the following linear system induced by*

$$\sum_{j=1}^{N_0-1} \tilde{\mu}_j \tilde{\delta}_{x_j} = \sum_{j=1}^{N_0-1} \mu_j \delta_{x_j} \tag{2.13}$$

has a unique solution $\{\tilde{\mu}_j\}_{j=1}^{N_0-1}$. Moreover, $V = W$, where V is the linear span of $\{\tilde{v}_j\}_{j=1}^{N_0-1}$ and W is the linear span of $\{w_j\}_{j=1}^{N_0-1}$, where w_j is the weak solution to the following global problem:

$$\mathcal{L}w_j(x) = \delta_{x_j}, \qquad x \in (0,1) \quad \textit{with} \quad \mathcal{B}_0 w_j(0) = 0, \quad \mathcal{B}_1 w_j(1) = 0. \tag{2.14}$$

Proof. Since $\mathcal{L}\tilde{v}_j = \tilde{\delta}_{x_j}$ on $(0,1)$ and $\tilde{v}_j(0) = \tilde{v}_j(1) = 0$ except $\mathcal{B}_0\tilde{v}_1(0) = 0$ and $\mathcal{B}_1\tilde{v}_{N_0-1}(1) = 0$, we obviously have $\tilde{v}_j \in W$ and hence $V \subseteq W$. We now prove that $\tilde{v}_1, \ldots, \tilde{v}_{N_0-1}$ are linearly independent. To do so, we claim that it is impossible that either $\tilde{v}_j|_{(x_{j-1},x_j)}$ or $\tilde{v}_j|_{(x_j,x_{j+1})}$ can be identically zero. Without loss of generality, we assume that $\tilde{v}_j|_{(x_{j-1},x_j)}$ is identically zero. Since $\tilde{v}_j|_{(x_{j-1},x_{j+1})} = v_j \in H^1(x_{j-1}, x_{j+1})$, the function v_j is continuous on (x_{j-1}, x_{j+1}) and hence $v_j(x_j) = 0$. However, since v_j is the weak solution to the local linking problem in (2.10), we see that $\mathring{v}_j := v_j|_{(x_j,x_{j+1})}$ must be the weak solution to $\mathcal{L}\mathring{v}_j(x) = 0$ on (x_j, x_{j+1}) with the boundary conditions $\mathring{v}_j(x_j) = 0$ and $\mathring{v}_j(x_{j+1}) = 0$ (if $j = N_0 - 1$, then $x_{j+1} = 1$ and replace $\mathring{v}_j(x_{j+1}) = 0$ by $\mathcal{B}_1\mathring{v}_{N_0-1}(1) = 0$). By the uniqueness of the solution, the weak solution $\mathring{v}_j$ must be identically zero. Hence, v_j must be identically zero, which contradicts (2.10). Hence, both $v_j|_{(x_{j-1},x_j)}$ and $v_j|_{(x_j,x_{j+1})}$ cannot be identically zero; i.e., $\tilde{v}_j|_{(x_{j-1},x_j)}$ and $\tilde{v}_j|_{(x_j,x_{j+1})}$ cannot be identically zero.

Consider the linear combination $v := \sum_{j=1}^{N_0-1} \mu_j \tilde{v}_j$ such that v is identically zero. Because $\tilde{v}_j$ vanishes outside (x_{j-1}, x_{j+1}), we have $0 = v|_{(x_0,x_1)} = \mu_1\tilde{v}_1|_{(x_0,x_1)}$. Since $\tilde{v}_1|_{(x_0,x_1)}$ cannot be identically zero, we must have $\mu_1 = 0$. By induction on j, we must have $\mu_1 = \mu_2 = \cdots = \mu_{N_0-1} = 0$. This proves that the elements in $\{\tilde{v}_j\}_{j=1}^{N_0-1}$ must be linearly independent. Now

by $V \subseteq W$, we conclude that $V = W$. The uniqueness of the solution to the linear system in (2.13) follows straightforwardly, since $\mathcal{L}\tilde{v}_j = \tilde{\delta}_{x_j}$, $\mathcal{L}w_j = \delta_{x_j}$ and $V = W$. □

The following result is the main ingredient of our DAT method.

Theorem 2.2. *Define*

$$u := u_f + u_\delta \quad \textit{with} \quad u_f := \sum_{j=0}^{N_0} \tilde{u}_j, \quad u_\delta := \sum_{j=1}^{N_0-1} \mu_j \tilde{v}_j, \tag{2.15}$$

where $\tilde{u}_j$ is the weak solution to (2.7) with prescribed boundary conditions in (2.8) extended by zero, $\tilde{v}_j$ is the weak solution to (2.10) with the prescribed boundary conditions in (2.11) extended by zero, and $\{\mu_j\}_{j=1}^{N_0-1}$ is the unique solution to the following linear system for the linking problem:

$$\sum_{j=1}^{N_0-1} \mu_j \tilde{\delta}_{x_j} = -\sum_{j=1}^{N_0-1} (d_{x_j}(\tilde{u}_{j-1}) + d_{x_j}(\tilde{u}_{j+1}))\delta_{x_j}, \tag{2.16}$$

Then u must be the weak solution to the heterogeneous Helmholtz equation in (2.1) with the boundary conditions in (2.2).

Proof. Since $\sum_{j=0}^{N_0} f_j = f$, we can write

$$f(x) = \sum_{j=0}^{N_0} f_j(x) = \sum_{j=0}^{N_0} \tilde{f}_j(x) - \sum_{j=1}^{N_0-1} (d_{x_j}(\tilde{u}_{j-1}) + d_{x_j}(\tilde{u}_{j+1}))\delta_{x_j}. \tag{2.17}$$

By Theorem 2.1, there is a unique solution $\{\mu_j\}_{j=1}^{N_0-1}$ to (2.16). That is, the linking problem in (2.16) can be uniquely solved. Hence, using (2.17), we can further write

$$f(x) = \sum_{j=0}^{N_0} \tilde{f}_j(x) + \sum_{j=1}^{N_0-1} \mu_j \tilde{\delta}_{x_j}. \tag{2.18}$$

By the definition of $\tilde{u}_j$, we observe that $\mathcal{L}u_f(x) = \sum_{j=0}^{N_0} \tilde{f}_j(x)$ for $x \in (0,1)$ with the boundary conditions $\mathcal{B}_0 u_f(0) = g_0$ and $\mathcal{B}_1 u_f(1) = g_1$. On the other hand, by the definition of $\tilde{\delta}_{x_j}$, we have

$$\mathcal{L}u_\delta(x) = \sum_{j=1}^{N_0-1} \mu_j \tilde{\delta}_{x_j}, \qquad x \in (0,1)$$

and u_δ satisfies the boundary conditions $\mathcal{B}_0 u_\delta(0) = 0$ and $\mathcal{B}_1 u_\delta(1) = 0$. Thus, by (2.18) we have

$$\mathcal{L}u = \mathcal{L}u_f + \mathcal{L}u_\delta = \sum_{j=0}^{N_0} \tilde{f}_j(x) + \sum_{j=1}^{N_0-1} \mu_j \tilde{\delta}_{x_j} = f(x), \qquad x \in (0,1)$$

and u satisfies the prescribed boundary conditions in (2.2). □

Obviously, we can recursively apply the above procedure to solve each of the local problems in (2.7) with prescribed boundary conditions in (2.8) to further reduce the size of the problem. To elucidate this point, we now present the DAT algorithm below.

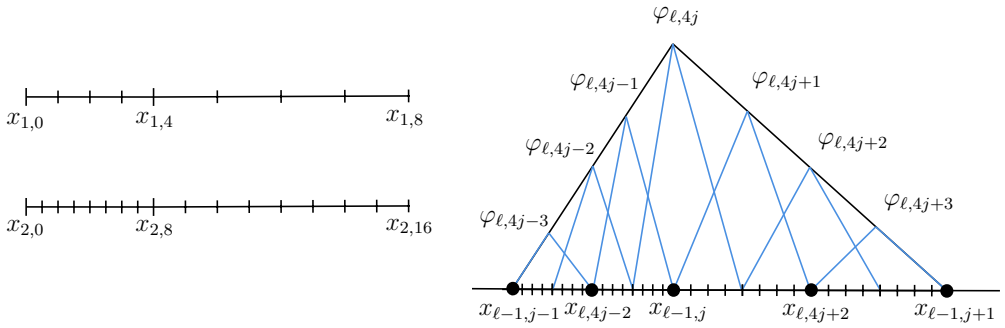


Figure 2.1: Left: An initial partition of $[0,1]$ at $\ell = 1$ with $N_0 = 8$ and its subsequent refinement at $\ell = 2$ with $s = 1$. Right: The relationship between an interior large hat function $\varphi_{\ell-1,j}$ on $[x_{\ell-1,j-1}, x_{\ell-1,j+1}]$ and smaller hat functions $\varphi_{\ell,2^s j+k}$, $-2^s+1 \leqslant k \leqslant 2^s-1$, with $s = 2$.

Algorithm 2.1. Consider the 1-level partition $\{x_{1,j}\}_{j=0}^{N_0}$ given by $0 = x_{1,-1} = x_{1,0} < x_{1,1} < \cdots < x_{1,N_0-1} < x_{1,N_0} = x_{1,N_0+1} = 1$, and let $\{\varphi_{1,j}\}_{j=0}^{N_0}$ be the associated partition of unity such that $\mathrm{supp}(\varphi_{1,j}) \subseteq [x_{1,j-1}, x_{1,j+1}]$ with $\varphi_{1,j}(x_{1,j}) = 1$. Pick $L, s \in \mathbb{N}$ such that L is the tree level and any subinterval at a level is divided equally into 2^s small subintervals at the next level. For each tree level $\ell = 2, \ldots, L$, let $\{x_{\ell,j}\}_{j=0}^{2^{\ell s}N_0}$ be a refinement partition of the grid $\{x_{\ell-1,j}\}_{j=0}^{2^{(\ell-1)s}N_0}$ such that $\{x_{\ell-1,j}\}_{j=0}^{2^{(\ell-1)s}N_0} \subset \{x_{\ell,j}\}_{j=0}^{2^{\ell s}N_0}$ with $x_{\ell-1,j} = x_{\ell,2^s j}$, $x_{\ell,-1} := x_{\ell,0} = 0$, and $x_{\ell,2^{\ell s}N_0+1} := x_{\ell,2^{\ell s}N_0} = 1$. Fix $N \in \mathbb{N}$ such that $N+1$ is the total number of points on the finest grid in the tree and $\{x_{L,j}\}_{j=0}^{2^{Ls}N_0} \subset \{x_j\}_{j=0}^{N}$. The local problems will be solved on this fixed fine grid. Note that $\mathrm{supp}(\varphi_{\ell,j}) \subseteq [x_{\ell,j-1}, x_{\ell,j+1}]$ with $\varphi_{\ell,j}(x_{\ell,j}) = 1$ for $0 \leqslant j \leqslant 2^{(\ell-1)s}N_0$ and

$$\varphi_{\ell-1,0} = \sum_{k=0}^{2^s-1} \varphi_{\ell,k}, \quad \varphi_{\ell-1,2^{(\ell-2)s}N_0} = \sum_{k=-2^s+1}^{0} \varphi_{\ell,2^{(\ell-1)s}N_0+k}, \quad \varphi_{\ell-1,j} = \sum_{k=-2^s+1}^{2^s-1} \varphi_{\ell,2^s j+k}$$

for $j = 1, \ldots, 2^{(\ell-2)s}N_0 - 1$. See Fig. 2.1 for an illustration of the setting above.

(S1) Solve the following (regular and Dirac assisted) local problems at tree level L in a parallel fashion using any chosen discretization method.

$$\begin{cases} \mathcal{L}u_{L,j} = f_j =: f\varphi_{L,j}, \qquad x \in (x_{L,j-1}, x_{L,j+1}), \qquad j = 0, \ldots, 2^{(L-1)s}N_0, \\ (\mathcal{B}_0 u_{L,j}(0) - g_0\delta_{0,j})\delta_{0,x_{L,j-1}} + (1 - \delta_{0,x_{L,j-1}})u_{L,j}(x_{L,j-1}) = 0, \\ (\mathcal{B}_1 u_{L,j}(1) - g_1\delta_{2^{Ls}N_0,j})\delta_{1,x_{L,j+1}} + (1 - \delta_{1,x_{L,j+1}})u_{L,j}(x_{L,j+1}) = 0, \end{cases} \tag{2.19}$$

$$\begin{cases} \mathcal{L}v_{L,j} = \delta_{x_{L,j}}, \qquad x \in (x_{L,j-1}, x_{L,j+1}), \qquad j = 1, \dots, 2^{(L-1)s}N_0 - 1, \\ \mathcal{B}_0 v_{L,j}(0)\delta_{0,x_{L,j-1}} + (1 - \delta_{0,x_{L,j-1}})v_{L,j}(x_{L,j-1}) = 0, \\ \mathcal{B}_1 v_{L,j}(1)\delta_{1,x_{L,j+1}} + (1 - \delta_{1,x_{L,j+1}})v_{L,j}(x_{L,j+1}) = 0. \end{cases} \tag{2.20}$$

For all $j = 0, \dots, 2^{(L-1)s}N_0$, extend $u_{L,j}$ by zero outside of $[x_{L,j-1}, x_{L,j+1}]$ and denote it by $\tilde{u}_{L,j}$. Similarly, for all $j = 1, \dots, 2^{(L-1)s}N_0 - 1$, extend $v_{L,j}$ by zero outside of $[x_{L,j-1}, x_{L,j+1}]$ and denote it by $\tilde{v}_{L,j}$.

(S2) Let $\ell = L, \dots, 2$ decreasingly. Consider the artificial Dirac distributions at each grid point and find the appropriate linear combination of Dirac local problems to offset it. This allows us to recover the solutions to local problems at level $\ell - 1$ from those at level ℓ. More explicitly, define an $n_2 \times n_2$ tridiagonal matrix and an n_2 column vector as follows

$$\begin{aligned} T_{\ell,n_1,n_2} &:= \mathrm{tridiag}(\{d_{x_{\ell,n_1+m}}(\tilde{v}_{\ell,n_1+m-1})\}_{m=2}^{n_2}, \{1\}_{m=1}^{n_2}, \{d_{x_{\ell,n_1+m}}(\tilde{v}_{\ell,n_1+m+1})\}_{m=1}^{n_2-1}), \\ \gamma_{\ell,n_1,n_2} &:= [-d_{x_{\ell,n_1+m}}(\tilde{u}_{\ell,n_1+m-1})(1 - \delta_{1,m}(1 - \delta_{0,n_1}(1 - \delta_{1,\ell}))) \\ &\qquad - d_{x_{\ell,n_1+m}}(\tilde{u}_{\ell,n_1+m+1})(1 - \delta_{n_2,m}(1 - \delta_{2^{(\ell-1)s}N_0-2^s,n_1}(1 - \delta_{1,\ell})))]_{1 \leqslant m \leqslant n_2}, \end{aligned}$$

where $n_1, n_2 \in \mathbb{N} \cup \{0\}$, and the first, second, and third arguments of $\mathrm{tridiag}(\cdot,\cdot,\cdot)$ correspond to the entries in the lower, main, and upper diagonals. Given a column vector μ, we denote the kth component of μ by $(\mu)_k$. For each ℓ, solve the linking problems obtained from steps (a)-(d) below in a parallel fashion.

(a) (Left-most element of partition of unity) Construct a $(2^s-1)\times(2^s-1)$ tridiagonal matrix $T_{\ell,0,2^s-1}$ and a (2^s-1) column vector $\gamma_{\ell,0,2^s-1}$. Set $\mu_{\ell,0} = (T_{\ell,0,2^s-1})^{-1}\gamma_{\ell,0,2^s-1}$.

(b) (Interior elements of partition of unity) For all $j = 1, \dots, 2^{(\ell-2)s}N_0 - 1$, construct a $(2^{s+1}-1)\times(2^{s+1}-1)$ tridiagonal matrix $T_{\ell,2^s(j-1),2^{s+1}-1}$ and a $(2^{s+1}-1)$ column vector $\gamma_{\ell,2^s(j-1),2^{s+1}-1}$. Let e_{2^s} is a $(2^{s+1}-1)$ vector with 1 in the 2^sth entry and 0 in the remaining entries. Set $\mu_{\ell,j} = (T_{\ell,2^s(j-1),2^{s+1}-1})^{-1}\gamma_{\ell,2^s(j-1),2^{s+1}-1}$ and $\eta_{\ell,j} = (T_{\ell,2^s(j-1),2^{s+1}-1})^{-1}e_{2^s}$.

(c) (Right-most element of partition of unity) Construct a $(2^s-1)\times(2^s-1)$ tridiagonal matrix $T_{\ell,2^{(\ell-1)s}N_0-2^s,2^s-1}$ and a (2^s-1) column vector $\gamma_{\ell,2^{(\ell-1)s}N_0-2^s,2^s-1}$. Set $\mu_{\ell,2^{(\ell-2)s}N_0} = (T_{\ell,2^{(\ell-1)s}N_0-2^s,2^s-1})^{-1}\gamma_{\ell,2^{(\ell-1)s}N_0-2^s,2^s-1}$.

(d) (Construct the solutions to local problems at level $\ell - 1$) For $j = 0, 2^{(\ell-2)s}N_0$

(left-most and right-most elements respectively), set

$$\tilde{u}_{\ell-1,0} = \sum_{k=0}^{2^s-1} \tilde{u}_{\ell,k} + \sum_{k=1}^{2^s-1} (\mu_{\ell,0})_k \tilde{v}_{\ell,k},$$
$$\tilde{u}_{\ell-1,2^{(\ell-2)s}N_0} = \sum_{k=-2^s+1}^{0} \tilde{u}_{\ell,2^{(\ell-1)s}N_0+k} + \sum_{k=1}^{2^s-1} (\mu_{\ell,2^{(\ell-2)s}N_0})_k \tilde{v}_{\ell,2^{(\ell-1)s}N_0-2^s+k}.$$

For $j = 1, \ldots, 2^{(\ell-2)s}N_0 - 1$ (interior elements), set

$$\tilde{u}_{\ell-1,j} = \sum_{k=-2^s+1}^{2^s-1} \tilde{u}_{\ell,2^s j+k} + \sum_{k=1}^{2^{s+1}-1} (\mu_{\ell,j})_k \tilde{v}_{\ell,2^s j-2^s+k}, \quad \tilde{v}_{\ell-1,j} = \sum_{k=1}^{2^{s+1}-1} (\eta_{\ell,j})_k \tilde{v}_{\ell,2^s j-2^s+k}.$$

(S3) Construct an $(N_0-1)\times(N_0-1)$ tridiagonal matrix $T_{1,0,N_0-1}$ and an (N_0-1) column vector $\gamma_{1,0,N_0-1}$. Set $\mu_{1,0} = (T_{1,0,N_0-1})^{-1}\gamma_{1,0,N_0-1}$. Finally, the approximated solution of the problem (2.1)-(2.2) is given by $u = \sum_{k=0}^{N_0} \tilde{u}_{1,k} + \sum_{k=1}^{N_0-1} (\mu_{1,0})_k \tilde{v}_{0,k}$.

As an illustrative example, for an equispaced grid on $[0,1]$ with $N_0 = 4$, $h = 2^{-n}$, $L = n-2$, $s = 1$ and $n \in \mathbb{N}$, the size of each linking and local problem (with the exception of those near the boundaries) is a 3×3 matrix equation. This exactly describes the situation in Examples 2.1 and 2.4 in Section 2.4.

Thus far, we have described DAT for the linear differential operator $\mathcal{L}$ defined in (2.1). The DAT method with appropriate modification can be generalized to general 1D linear differential operators $\mathcal{L}$ (e.g., the biharmonic equation involving higher order derivatives). We would need to modify Theorem 2.2 about how we patch the local problems in (2.3) by means of the Dirac assisted local problems in (2.10) equipped with suitable boundary conditions. In addition to (2.10), we may have to solve additional Dirac assisted linking problems in (2.10) using higher order distributional derivatives of δ_{x_j}.

As it currently stands, DAT can handle multidimensional problems that can be decomposed into a series of 1D problems (e.g., by the separation of variables). We shall provide a few relevant 2D numerical examples in Section 2.4.3. We shall discuss the use of DAT for solving general 2D/3D problems in Chapter 7.

2.2 Compact finite difference schemes with arbitrarily high accuracy orders

To numerically solve the heterogeneous Helmholtz equation in (2.1)–(2.2) with piecewise smooth coefficients a, κ^2 and source term f, in this section we shall study compact finite difference schemes with arbitrarily high accuracy and numerical dispersion orders. Such compact finite difference schemes are important for accurately solving local problems stemming from DAT in the foregoing section.

2.2.1 Compact stencils for interior points

We start by stating a simple observation, which is critical for proving the existence of a 1D finite difference scheme with an arbitrarily high accuracy order. The following observation uses an analyticity assumption for its theoretical analysis; however, we only require the coefficients to be differentiable up to a certain order as we shall see later in this section.

Proposition 2.3. *Let a, κ^2, f in (2.1) be analytic functions and let u be an analytic function satisfying $[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x)$ with $a(x) > 0$ for all $x \in (0,1)$. For any point $x_b \in (0,1)$, we have*

$$u^{(j)}(x_b) = E_{j,0}u(x_b) + E_{j,1}u'(x_b) + \sum_{\ell=0}^{j-2} F_{j,\ell} f^{(\ell)}(x_b), \qquad j \geqslant 2, \tag{2.21}$$

where the quantities $E_{j,0}, E_{j,1}, F_{j,\ell}$ only depend on the values $a(x_b), a'(x_b), \ldots, a^{(j-1)}(x_b)$ and $\kappa^2(x_b)$, $[\kappa^2]'(x_b), \ldots, [\kappa^2]^{(j-2)}(x_b)$ for $j \geqslant 2$ and $\ell \in \mathbb{N}_0$. Consequently, for sufficiently small h,

$$u(x_b + h) = u(x_b)E_0(h) + u'(x_b)hE_1(h) + \sum_{\ell=0}^{\infty} h^{\ell+2} f^{(\ell)}(x_b) F_\ell(h), \tag{2.22}$$

where $E_0(h), E_1(h)$ and $F_\ell(h), \ell \in \mathbb{N}_0$ are defined to be

$$E_0(h) := 1 + \sum_{j=2}^{\infty} \frac{E_{j,0}}{j!} h^j, \quad E_1(h) := 1 + \sum_{j=2}^{\infty} \frac{E_{j,1}}{j!} h^{j-1}, \quad F_\ell(h) := \sum_{j=\ell+2}^{\infty} \frac{F_{j,\ell}}{j!} h^{j-\ell-2}. \tag{2.23}$$

Proof. We prove the claim in (2.21) using mathematical induction on j. Consider the base case with $j = 2$. Since $a(x) > 0$, we deduce from $[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x)$ that

$$u^{(2)}(x) = -\tfrac{\kappa^2(x)}{a(x)}u(x) - \tfrac{a'(x)}{a(x)}u'(x) + \tfrac{f(x)}{a(x)}, \qquad x \in (0,1). \tag{2.24}$$

Hence, setting $x = x_b$ in the above identity (2.24), we conclude that (2.21) holds for $j = 2$.

Suppose that the claim in (2.21) holds for some $j \geqslant 2$. We now prove that (2.21) must hold for $j+1$. Applying the $(j-1)$th derivative to both sides of the identity in (2.24), we observe that

$$u^{(j+1)}(x) = -\left[\tfrac{\kappa^2(x)}{a(x)}u(x)\right]^{(j-1)} - \left[\tfrac{a'(x)}{a(x)}u'(x)\right]^{(j-1)} + \left[\tfrac{f(x)}{a(x)}\right]^{(j-1)}.$$

Applying the Leibniz differentiation formula to the above identity, we conclude that the quantity $u^{(j+1)}(x)$ can be written as a linear combination of $f(x), f'(x), \ldots, f^{(j-1)}(x)$ and $u(x), u'(x), \ldots, u^{(j)}(x)$ with all combination coefficients being analytic functions of x depending only on $a(x), a'(x), \ldots, a^{(j)}(x)$ and $\kappa^2(x), [\kappa^2(x)]', \ldots, [\kappa^2(x)]^{(j-1)}$. Now by induction hypothesis, we conclude that (2.21) holds for $j+1$. This proves (2.21) by mathematical induction on j.

On the other hand, since u is analytic in a neighborhood of x_b, the Taylor series of u at the base point x_b is $u(x_b+h) = u(x_b) + u'(x_b)h + \sum_{j=2}^{\infty} \frac{u^{(j)}(x_b)}{j!}h^j$. Therefore, we deduce from (2.21) that

$$\begin{aligned} u(x_b+h) &= u(x_b) + u'(x_b)h + \sum_{j=2}^{\infty} \frac{h^j}{j!}\left(E_{j,0}u(x_b) + E_{j,1}u'(x_b) + \sum_{\ell=0}^{j-2} F_{j,\ell} f^{(\ell)}(x_b)\right) \\ &= u(x_b)\left(1 + \sum_{j=2}^{\infty} \frac{E_{j,0}}{j!}h^j\right) + u'(x_b)\left(h + \sum_{j=2}^{\infty} \frac{E_{j,1}}{j!}h^j\right) + \sum_{j=2}^{\infty}\sum_{\ell=0}^{j-2} \frac{F_{j,\ell}}{j!}h^j f^{(\ell)}(x_b) \\ &= u(x_b)E_0(h) + u'(x_b)hE_1(h) + \sum_{\ell=0}^{\infty}\sum_{j=\ell+2}^{\infty} \frac{F_{j,\ell}}{j!}h^j f^{(\ell)}(x_b), \end{aligned}$$

from which we obtain (2.22). □

Let us now consider compact finite difference schemes with high accuracy and numerical dispersion orders for the heterogeneous Helmholtz equation in (2.1) with smooth coefficients a, κ^2 and source term f. Suppose the discretization stencil is centered at an interior point x_b with mesh size $0 < h < 1$. That is, we fix the base point x_b to be in $(0,1)$ such that $(x_b - h, x_b + h) \subset (0,1)$.

Theorem 2.4. *Suppose that a, κ^2, f in (2.1) are smooth functions. Let $M, \tilde{M}$ be positive integers with $M \geqslant \tilde{M}$. Let $0 < h < 1$ and $x_b \in (0,1)$ such that $(x_b - h, x_b + h) \subset (0,1)$. Consider the discretization stencil of a compact finite difference scheme for $\mathcal{L}u :=$*

$[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x)$ *(i.e.,* (2.1)*) at the base point* x_b *below*

$$\mathcal{L}_h u(x_b) := h^{-2}[c_{-1}(h)u(x_b-h) + c_0(h)u(x_b) + c_1(h)u(x_b+h)] - \sum_{\ell=0}^{\tilde{M}-1} d_\ell(h)h^\ell f^{(\ell)}(x_b), \quad (2.25)$$

where c_{-1}, c_0, c_1 *and* d_ℓ *are smooth functions of* h *for* $\ell = 0, \ldots, \tilde{M}-1$*. Suppose that*

$$\begin{aligned} c_1(h) &= \frac{\alpha(h)}{E_1(h)} + \mathscr{O}(h^{M+1}), \quad c_{-1}(h) = \frac{\alpha(h)}{E_1(-h)} + \mathscr{O}(h^{M+1}), \\ c_0(h) &= -c_1(h)E_0(h) - c_{-1}(h)E_0(-h) + \mathscr{O}(h^{M+2}) \end{aligned} \quad (2.26)$$

and

$$d_\ell(h) = -\delta_{0,\ell} + c_1(h)F_\ell(h) + (-1)^\ell c_{-1}(h)F_\ell(-h) + \mathscr{O}(h^{\tilde{M}-\ell}), \quad \ell = 0, \ldots, \tilde{M}-1, \quad (2.27)$$

as $h \to 0$*, where* α *is a smooth function of* h *with* $\alpha(0) \neq 0$ *and* $E_0(h), E_1(h)$ *and* $F_\ell(h), \ell \in \mathbb{N}_0$ *are defined uniquely in* (2.23) *of Proposition* 2.3*. Then the discretization stencil of the compact finite difference scheme has numerical dispersion order* M *at the base point* x_b*, that is,*

$$h^{-2}[c_{-1}(h)u(x_b-h) + c_0(h)u(x_b) + c_1(h)u(x_b+h)] = \mathscr{O}(h^M), \qquad h \to 0, \quad (2.28)$$

for every solution u *of* $\mathcal{L}u = 0$*, and has accuracy order* $\tilde{M}$ *at the base point* x_b*, that is,*

$$\mathcal{L}_h u(x_b) - f(x_b) = \mathscr{O}(h^{\tilde{M}}), \qquad h \to 0, \quad (2.29)$$

for every solution u *of* $\mathcal{L}u = f$*.*

Proof. By Proposition 2.3 and (2.21), all the quantities $E_{j,0}, E_{j,1}, F_{j,\ell}$ depend only on the values $a(x_b), a'(x_b), \ldots, a^{(j-1)}(x_b)$ and $\kappa^2(x_b), [\kappa^2]'(x_b), \ldots, [\kappa^2]^{(j-2)}(x_b)$ for $j \geqslant 2$ and $\ell \in \mathbb{N}_0$. For simplicity, we define $u_0 := u(x_b), u_1 := u'(x_b)$ and $f_\ell := f^{(\ell)}(x_b)$ for $\ell \in \mathbb{N}_0$. Thus, by (2.22), we deduce

$$\begin{aligned} &h^2\mathcal{L}_h u(x_b) - h^2 f_0 \\ &\quad = u_0\Big(c_{-1}(h)E_0(-h) + c_0(h) + c_1(h)E_0(h)\Big) + u_1 h\Big(-c_{-1}(h)E_1(-h) + c_1(h)E_1(h)\Big) \\ &\qquad + \sum_{\ell=0}^{\infty} h^{\ell+2} f_\ell \left(c_1(h)F_\ell(h) + (-1)^\ell c_{-1}(h)F_\ell(-h)\right) - (d_0(h)+1)h^2 f_0 - \sum_{\ell=1}^{\tilde{M}-1} d_\ell(h)h^{\ell+2} f_\ell, \end{aligned} \quad (2.30)$$

where $E_0(h)$ and $E_1(h)$ are defined in (2.23).

On the other hand, from $f(x) = [a(x)u'(x)]' + \kappa^2(x)u(x)$ we trivially observe that $f^{(\ell)}$ can be written as a linear combination of $u, u', \dots, u^{(\ell+2)}$ as well. Consequently, (2.28) holds (or equivalently, (2.29) holds with $f = 0$ and $\tilde{M} = M$) for numerical dispersion order M if and only if the coefficients of u_0 and u_1 in the above identity are $\mathscr{O}(h^{M+2})$ as $h \to 0$. That is, (2.28) is equivalent to

$$\begin{aligned} c_{-1}(h)E_0(-h) + c_0(h) + c_1(h)E_0(h) &= \mathscr{O}(h^{M+2}), \qquad h \to 0, \\ -c_{-1}(h)E_1(-h) + c_1(h)E_1(h) &= \mathscr{O}(h^{M+1}), \qquad h \to 0. \end{aligned}$$

Solving the above equation and noting that $E_0(0) = E_1(0) = 1$, we conclude that (2.26) holds if and only if (2.28) holds. Thus, (2.28) holds for numerical dispersion order M if and only if (2.26) holds.

We now prove (2.29). Since we proved (2.28) and $M \geqslant \tilde{M}$, (2.29) holds for accuracy order $\tilde{M}$ if and only if all $f_0, \dots, f_{\tilde{M}-1}$ must be $\mathscr{O}(h^{\tilde{M}+2})$ as $h \to 0$. Rearranging the last line of (2.30), we conclude

$$\begin{aligned} &\sum_{\ell=0}^{\infty} h^{\ell+2}\left(c_1(h)F_\ell(h) + (-1)^\ell c_{-1}(h)F_\ell(-h)\right) f_\ell - (d_0+1)h^2 f_0 - \sum_{\ell=1}^{\tilde{M}-1} d_\ell h^{\ell+2} f_\ell \\ &\quad = \sum_{\ell=0}^{\tilde{M}-1} h^{\ell+2}\left(-(d_\ell + \delta_{0,\ell}) + c_1(h)F_\ell(h) + (-1)^\ell c_{-1}(h)F_\ell(-h)\right) f_\ell + \mathscr{O}(h^{\tilde{M}+2}). \end{aligned}$$

Since (2.28) holds, (2.29) now is equivalent to that all the coefficients f_ℓ, $\ell = 0, \dots, \tilde{M}-1$ in the above identity must be $\mathscr{O}(h^{\tilde{M}+2})$, that is,

$$-(d_\ell + \delta_{0,\ell}) + c_1(h)F_\ell(h) + (-1)^\ell c_{-1}(h)F_\ell(-h) = \mathscr{O}(h^{\tilde{M}-\ell}), \qquad \ell = 0, \dots, \tilde{M}-1.$$

Solving the above linear equations for $d_0, \dots, d_{\tilde{M}-1}$ and using (2.26), we obtain (2.27). This proves (2.29) for accuracy order $\tilde{M}$. □

We make some remarks on Theorem 2.4. First, from the proof of Theorem 2.4, we see that Theorem 2.4 finds all the possible compact FDMs with accuracy order $\tilde{M}$ and numerical dispersion order M. Because (2.29) for accuracy order $\tilde{M}$ automatically implies (2.28) for numerical dispersion order M with $M = \tilde{M}$, we often take $M = \tilde{M}$. Also, κ^2 can be replaced by $-\kappa^2$ (i.e., κ can be complex-valued). Second, if E_0 and E_1 in (2.23) have closed forms, then we can have numerical dispersion order $M = \infty$ for a pollution free scheme by selecting $c_1(h) = -1/E_1(h), c_{-1}(h) = -1/E_1(-h)$ and $c_0(h) = E_0(h)/E_1(h) + E_0(-h)/E_1(-h)$ in

(2.26). In particular, for constant functions a and κ^2, we observe

$$E_0(h) = \cos(\tilde{h}), \quad E_1(h) = \tilde{h}^{-1}\sin(\tilde{h}) \quad \text{with} \quad \tilde{h} := h\kappa/\sqrt{a} \tag{2.31}$$

and for $\ell \in \mathbb{N}_0$,

$$F_{2\ell}(h) = \frac{\cos(\tilde{h}) - \sum_{j=0}^{\ell} \frac{(-1)^j}{(2j)!}\tilde{h}^{2j}}{(-1)^{\ell+1}\tilde{h}^{2\ell+2}a}, \quad F_{2\ell+1}(h) = \frac{\sin(\tilde{h}) - \sum_{j=0}^{\ell} \frac{(-1)^j}{(2j+1)!}\tilde{h}^{2j+1}}{(-1)^{\ell+1}\tilde{h}^{2\ell+3}a}. \tag{2.32}$$

This pollution free scheme coincides with that of [122, 124]. In the literature, a dispersion correction procedure for 1D homogeneous Helmholtz equation with (piecewise) constant wavenumbers also exists [26, 27, 46]. However, the procedure as presented uses the standard second order FDM, which itself is not a pollution free scheme. The correction solely comes from modifying the original wavenumber. More specifically, the method involves inserting the exact homogeneous solution of the 1D Helmholtz equation on the real line into the standard second order FDM to obtain the modified wavenumber.

2.2.2 Compact stencils for boundary points

We now handle the case that the base point is one of the endpoints. It is important that a compact FDM should achieve the same accuracy order and numerical dispersion order at the endpoints as it does at interior points. The following result answers this question. For simplicity, we only handle the boundary condition at a base point x_b from its right side, while the treatment for the boundary condition at x_b from its left side is similar through symmetry. Because we shall handle piecewise smooth coefficients, let us consider a general boundary condition at $x_b \in [0,1)$ from its right side. For $j \in \mathbb{N}_0$ and a function $f(x)$, $f^{(j)}(x_b+) := \lim_{x\to x_b^+} f^{(j)}(x)$ and $f^{(j)}(x_b-) := \lim_{x\to x_b^-} f^{(j)}(x)$ for one-sided derivatives.

Theorem 2.5. *Suppose that a, κ^2, f in (2.1) are smooth functions. Let $M, \tilde{M}$ be positive integers with $M \geqslant \tilde{M}$. Let $0 < h < 1$ and the boundary condition at $x_b \in [0,1)$ with $(x_b, x_b + h) \subset (0,1)$. Suppose that the boundary condition at x_b for the right side of x_b is given by*

$$\mathcal{B}^+u(x_b) := \lambda_0 u(x_b+) + \lambda_1 u'(x_b+) \quad \textit{with} \quad \lambda_0, \lambda_1 \in \mathbb{C}. \tag{2.33}$$

Consider the discretization stencil of a compact FDM for $\mathcal{L}u := [a(x)u'(x)]' + \kappa^2(x)u(x) = f(x)$ at the base point x_b, from the right side of x_b with the above boundary condition, below

$$\mathcal{L}_h^{\mathcal{B}^+}u(x_b) := h^{-1}[c_0^{\mathcal{B}^+}(h)u(x_b) + c_1^{\mathcal{B}^+}(h)u(x_b+h)] - \sum_{\ell=0}^{\tilde{M}-2} d_\ell^{\mathcal{B}^+}(h)h^{\ell+1}f^{(\ell)}(x_b+), \tag{2.34}$$

where $c_0^{\mathcal{B}^+}, c_1^{\mathcal{B}^+}$ *and* $d_\ell^{\mathcal{B}^+}$ *are smooth functions of* h *for* $\ell = 0, \dots, \tilde{M} - 2$. *Suppose that*

$$c_1^{\mathcal{B}^+}(h) = \frac{\lambda_1}{E_1(h)} + \mathscr{O}(h^M), \quad c_0^{\mathcal{B}^+}(h) = h\lambda_0 - c_1^{\mathcal{B}^+}(h)E_0(h) + \mathscr{O}(h^{M+1}) \tag{2.35}$$

and

$$d_\ell^{\mathcal{B}^+}(h) = c_1^{\mathcal{B}^+}(h)F_\ell(h) + \mathscr{O}(h^{\tilde{M}-\ell-1}), \qquad \ell = 0, \dots, \tilde{M} - 2, \tag{2.36}$$

as $h \to 0$, *where* $E_0, E_1, F_\ell, \ell \in \mathbb{N}_0$ *are given in* (2.23) *of Proposition* 2.3 *and are determined by* $a^{(n)}(x_b+)$ *and* $[\kappa^2]^{(n)}(x_b+)$ *for* $n \in \mathbb{N}_0$. *Then the discretization stencil of the compact finite difference scheme at the base point* x_b *with the boundary condition in* (2.33) *from the right side of* x_b *satisfies*

$$c_0^{\mathcal{B}^+}(h)u(x_b) + c_1^{\mathcal{B}^+}(h)u(x_b + h) = \mathscr{O}(h^M), \qquad h \to 0 \tag{2.37}$$

for every solution u *of* $\mathcal{L}u = 0$, *and*

$$\mathcal{L}_h^{\mathcal{B}^+}u(x_b) - \mathcal{B}^+u(x_b) = \mathscr{O}(h^{\tilde{M}}), \qquad h \to 0 \tag{2.38}$$

for every solution u *of* $\mathcal{L}u = f$.

Proof. The proof is similar to but easier than the proof of Theorem 2.4 by using (2.22), which implies

$$u'(x_b) = \frac{1}{hE_1(h)}u(x_b + h) - \frac{E_0(h)}{hE_1(h)}u(x_b) - \sum_{\ell=0}^{\infty} \frac{h^{\ell+1}}{E_1(h)} f^{(\ell)}(x_b+)F_\ell(h). \tag{2.39}$$

Since $\mathcal{B}^+u(x_b) = \lambda_0 u(x_b+) + \lambda_1 u'(x_b+)$, using (2.39) we obtain

$$\mathcal{B}^+u(x_b) = h^{-1}\left[\frac{\lambda_1}{E_1(h)}u(x_b + h) + \left(\lambda_0 h - \frac{\lambda_1 E_0(h)}{E_1(h)}\right)u(x_b)\right] - \sum_{\ell=0}^{\infty} \frac{\lambda_1}{E_1(h)} h^{\ell+1} f^{(\ell)}(x_b+)F_\ell(h). \tag{2.40}$$

Now the claim follows directly by using (2.40) and (2.34). □

If E_0 and E_1 in (2.23) have closed forms, then we can achieve numerical dispersion order $M = \infty$ for pollution free by selecting $c_1(h) = \frac{\lambda_1}{E_1(h)}$ and $c_0(h) = h\lambda_0 - \frac{E_0(h)}{E_1(h)}\lambda_1$ in (2.35) of Theorem 2.5. In particular, for constant functions a and κ^2, E_0, E_1, F_ℓ are given in (2.31) and (2.32). As before, this pollution free scheme coincides with that of [122, 124] when the boundary condition takes the form of $\lambda_0 = 1$ and $\lambda_1 = 0$ in (2.33) or $\lambda_0 = -\mathrm{i}\kappa$ and $\lambda_1 = -1$ in (2.33).

2.2.3 Compact stencils for piecewise smooth coefficients

We now discuss piecewise smooth coefficients a, κ^2 and f. Assume that a, κ^2, f may have a breaking/branch point $x_c := x_b - \theta h \in (x_b - h, x_b + h)$ with $\theta \in (-1, 1)$ such that they are smooth on $(x_b - h, x_c)$ and $(x_c, x_b + h)$, but they may be discontinuous at x_c. We also assume that all the one-sided derivatives of a, κ^2, f exist at x_c and assume $\theta \in [0, 1)$ for simplicity. To solve the Dirac assisted local problems in (2.20) of Algorithm 2.1 for DAT, we also assume that $w(\delta_{x_c})$ is the weight of the Dirac distribution δ_{x_c} in the source term f. We can generalize Theorem 2.4 by considering the following discretization stencil at x_b:

$$\begin{aligned}\mathring{\mathcal{L}}_h u(x_b) := h^{-2}[c_{-1}(h)u(x_b - h) + c_0(h)u(x_b) + c_1(h)u(x_b + h)] \\ - h^{-1} d_w(h) w(\delta_{x_c}) - \sum_{\ell=0}^{\tilde{M}-1} h^\ell \left(d_\ell^+(h) f^{(\ell)}(x_c+) + d_\ell^-(h) f^{(\ell)}(x_c-) \right),\end{aligned}$$

where c_{-1}, c_0, c_1, d_w, and $d_\ell^+, d_\ell^-, \ell = 0, \ldots, \tilde{M} - 1$ are smooth functions of h satisfying

$$\begin{aligned}&c_{-1}(h)E_{0,-}((\theta-1)h) + c_0(h)E_{0,+}(\theta h) + c_1(h)E_{0,+}((1+\theta)h) = \mathscr{O}(h^{M+2}),\\ &c_{-1}(h)(\theta-1)E_{1,-}((\theta-1)h)\tfrac{a(x_c+)}{a(x_c-)} + c_0(h)\theta E_{1,+}(\theta h) + c_1(h)(\theta+1)E_{1,+}((1+\theta)h) = \mathscr{O}(h^{M+1}),\\ &d_w(h) = c_{-1}(h)(1-\theta)E_{1,-}((\theta-1)h)\tfrac{1}{a(x_c-)} + \mathscr{O}(h^{M+1}), \quad h \to 0,\end{aligned}$$

and for $\ell = 0, \ldots, \tilde{M} - 1$,

$$\begin{aligned}d_\ell^+(h) &= c_0(h)\theta^{\ell+2}F_{\ell,+}(\theta h) + c_1(h)(\theta+1)^{\ell+2}F_{\ell,+}((\theta+1)h) + \mathscr{O}(h^{\tilde{M}-\ell}),\\ d_\ell^-(h) &= c_{-1}(h)(\theta-1)^{\ell+2}F_{\ell,-}((\theta-1)h) + \mathscr{O}(h^{\tilde{M}-\ell}), \quad h \to 0,\end{aligned}$$

where $E_{0,\pm}, E_{1,\pm}$ and $F_{\ell,\pm}$ are given in Proposition 2.3 at the point x_c (instead of x_b) using $a^{(j)}(x_c\pm)$, $[\kappa^2]^{(j)}(x_c\pm)$ and $f^{(j)}(x_c\pm)$ accordingly. Then the above discretization stencil has numerical dispersion order M at x_b by satisfying (2.28) and has accuracy order $\tilde{M}$ at x_b by satisfying $\mathring{\mathcal{L}}_h u(x_b) = \mathscr{O}(h^{\tilde{M}})$ as $h \to 0$ for every solution u of $\mathcal{L}u = f$. The proof of the above equations is very similar to that of Theorem 2.4 but we expand $u(x_b - h), u(x_b)$ and $u(x_b + h)$ through Proposition 2.3 at x_c instead of x_b by noting $u(x_b - h) = u(x_c + (\theta - 1)h)$, $u(x_b) = u(x_c + \theta h)$ and $u(x_b + h) = u(x_c + (\theta + 1)h)$. Then we link the two sides of x_c through the transmission conditions $u(x_c+) = u(x_c-)$ and $a(x_c+)u'(x_c+) - a(x_c-)u'(x_c-) = w(\delta_{x_c})$. If all coefficients a, κ^2, f are smooth inside $(x_b - h, x_b + h)$, then $\mathcal{L}_h u(x_b)$ in (2.25) of Theorem 2.4 can be recovered through $\mathcal{L}_h(x_b) = \mathring{\mathcal{L}}_h u(x_b) + f(x_b)$ using $x_c = x_b$. We shall not pursue this general issue further. In the present paper, it suffices for us to only consider the special case

that $x_c = x_b$, i.e., $\theta = 0$. For $x_c = x_b$, using the following special boundary operators at x_b:

$$\mathcal{B}^+u(x_b) := u'(x_b+) \quad \text{and} \quad \mathcal{B}^-u(x_b) := u'(x_b-), \tag{2.41}$$

instead of using $\mathring{\mathcal{L}}_h u(x_b)$ we can deduce a compact stencil at the base point x_b from Theorem 2.5 for (2.41) that

$$\mathcal{L}_h u = \tfrac{2a(x_b-)}{a(x_b+)+a(x_b-)}\mathcal{L}_h^{\mathcal{B}^-}u(x_b) - \tfrac{2a(x_b+)}{a(x_b+)+a(x_b-)}\mathcal{L}_h^{\mathcal{B}^+}u(x_b) = -\tfrac{2w(\delta_{x_b})}{a(x_b+)+a(x_b-)} \tag{2.42}$$

at the base point x_b, where $w(\delta_{x_b})$ is the weight of the Dirac distribution δ_{x_b} in the source term f.

2.2.4 A concrete example of finite difference schemes for $M = 8$

For the convenience of the reader, here we provide details about how to obtain concrete compact finite difference schemes with Mth order accuracy and numerical dispersion as discussed in Sections 2.2.1 to 2.2.3. In particular, we provide details for $M = 2, 4, 6, 8$. We first discuss how to compute $E_0, E_1, F_\ell, \ell \in \mathbb{N} \cup \{0\}$ as defined in (2.23). Using (2.24) and taking derivative on both sides of (2.21), we observe that the coefficients $E_{j,0}$, $E_{j,1}$ and $F_{j,\ell}, \ell = 0, \ldots, j-2$ at a base point x_b in Proposition 2.3 can be recursively obtained by

$$E_{j+1,0} = E'_{j,0} - \frac{\kappa^2}{a}E_{j,1}, \quad E_{j+1,1} = E_{j,0} + E'_{j,1} - \frac{a'}{a}E_{j,1}, \quad F_{j+1,\ell} = F'_{j,\ell} + F_{j,\ell-1},$$
$$j \geqslant 2, \quad \ell = 0, \ldots, j-1$$

with the initial values

$$E_{2,0} := -\frac{\kappa^2}{a}, \quad E_{2,1} := -\frac{a'}{a}, \quad \text{and} \quad F_{2,0} := \frac{1}{a}, \tag{2.43}$$

where we used the convention that $F_{j,-1} := \frac{E_{j,1}}{a}$ and $F_{j,\ell} := 0$ for all $\ell > j-2$. Note that $E_0(0) = E_1(0) = 1$ and $F_\ell(0) = F_{\ell+2,\ell} = F_{2,0} = \frac{1}{a(x_b)}$ for all $\ell \in \mathbb{N}_0$.

Let $M = \tilde{M} \in 2\mathbb{N}$. At an interior point x_b we obtain from (2.25) of Theorem 2.4 that

$$c_{-1}(h)u(x_b-h) + c_0(h)u(x_b) + c_1(h)u(x_b+h) = h^2 f(x_b) + \sum_{\ell=0}^{M-2} d_\ell(h) h^{\ell+2} f^{(\ell)}(x_b) + \mathscr{O}(h^{M+2}),$$

as $h \to 0$, where one particular choice of c_{-1}, c_0, c_1 satisfying (2.26) of Theorem 2.4 is given

by

$$c_{-1}(h) := -\mathcal{E}_1^{M-1}(-h), \quad c_1(h) := -\mathcal{E}_1^{M-1}(h), \quad c_0(h) := \mathcal{E}_1^{M-1}(h)\mathcal{E}_0^{M}(h) + \mathcal{E}_1^{M-1}(-h)\mathcal{E}_0^{M}(-h), \tag{2.44}$$

and the corresponding $d_\ell, \ell = 0, \dots, M-2$ in (2.27) are given by

$$d_\ell(h) := -\delta_{0,\ell} - \mathcal{E}_1^{M-1}(h)\mathcal{F}_\ell^{M-\ell-2}(h) - (-1)^\ell \mathcal{E}_1^{M-1}(-h)\mathcal{F}_\ell^{M-\ell-2}(-h), \tag{2.45}$$

where $\mathcal{E}_0^n, \mathcal{E}_1^n, F_\ell^n, n \in \mathbb{N}$ are the unique polynomials (in terms of h) of degree n satisfying

$$\mathcal{E}_0^n(h) = E_0(h) + \mathscr{O}(h^{n+1}), \quad \mathcal{E}_1^n(h) := 1/E_1(h) + \mathscr{O}(h^{n+1}), \quad \mathcal{F}_\ell^n(h) = F_\ell(h) + \mathscr{O}(h^{n+1}), \tag{2.46}$$

as $h \to 0$. Notice that $d_{M-1}(h) = 0$ and $E_1(0) = 1$. Observing $c_1(0) + (-1)^{M-1}c_{-1}(0) = 0$ for even $M \in 2\mathbb{N}$, one can directly check that the above choice in (2.44) and (2.45) satisfies all the conditions in (2.26) and (2.27) with $\alpha(h) = 1 - \beta h^M$, where β is the coefficient of h^M in the Taylor series of $\frac{1}{E_1(h)}$ at $h = 0$. Note that c_{-1}, c_0, c_1 in (2.44) and $d_\ell, \ell = 0, \dots, M-2$ in (2.45) only depend on $a, a', \dots, a^{(M-1)}$, $\kappa^2, [\kappa^2]', \dots, [\kappa^2]^{(M-2)}$ and $f, f', \dots, f^{(M-2)}$. We can also obtain stencils for odd integers $M = \tilde{M}$; however $c_1(0) + (-1)^{M-1}c_{-1}(0) \neq 0$ and consequently, we have to use $\mathcal{E}_1^M$ instead of $\mathcal{E}_1^{M-1}$, $\mathcal{E}_0^{M+1}$ instead of $\mathcal{E}_0^M$, and $\mathcal{F}_\ell^{M-\ell-1}$ instead of $\mathcal{F}_\ell^{M-\ell-2}$ in (2.44)-(2.45). Define $a_j := a^{(j)}(x_b)$, $\kappa_j := [\kappa^2]^{(j)}(x_b)$ and $f_j := f^{(j)}(x_b)$. For $M = \tilde{M} = 8$, we explicitly have

$$\begin{aligned}
\mathcal{E}_1^7 = 1 &+ \tfrac{ha_1}{2a_0} + \left(2a_0a_2 + 2a_0\kappa_0 - a_1^2\right)\tfrac{h^2}{12a_0^2} + \left(a_3{a_0}^2 + 2\kappa_1{a_0}^2 - 2a_1a_2a_0 + a_1^3\right)\tfrac{h^3}{24a_0^3} + \left(6a_4a_0^3 + 18\kappa_2a_0^3 - 18a_1a_3a_0^2\right. \\
&-6a_1\kappa_1a_0^2 - 16a_2^2a_0^2 - 2a_2\kappa_0a_0^2 + 14\kappa_0^2a_0^2 + 46a_1^2a_2a_0 - 2a_1^2\kappa_0a_0 - 19a_1^4\big)\tfrac{h^4}{720a_0^4} + \left(2a_5a_0^4 + 8\kappa_3a_0^4 - 8a_1a_4a_0^4 - 6a_1\kappa_2a_0^3\right. \\
&-20a_2a_3a_0^3 - 8a_2\kappa_1a_0^3 - 2a_3\kappa_0a_0^3 + 28\kappa_0\kappa_1a_0^3 + 29a_1^2a_3a_0^2 + 4a_1^2\kappa_1a_0^2 + 48a_1a_2^2a_0^2 - 2a_1a_2\kappa_0a_0^2 - 14a_1\kappa_0^2a_0^2 - 78a_1^3a_2a_0 \\
&\left.+4a_1^3\kappa_0a_0 + 27a_1^5\right)\tfrac{h^5}{1440a_0^5} + \left(12a_6a_0^5 + 60\kappa_4a_0^5 - 60a_1a_5a_0^4 - 72a_1\kappa_3a_0^4 - 192a_2a_4a_0^4 - 156a_2\kappa_2a_0^4 - 135a_3^2a_0^4 - 120a_3\kappa_1a_0^4\right. \\
&-24a_4\kappa_0a_0^4 + 348\kappa_0\kappa_2a_0^4 + 300\kappa_1^2a_0^4 + 282a_1^2a_4a_0^3 + 114a_1^2\kappa_2a_0^3 + 204a_1a_2\kappa_1a_0^3 + 1296a_1a_2a_3a_0^3 - 708a_1\kappa_0\kappa_1a_0^3 + 352a_2^3a_0^3 \\
&-12a_2^2\kappa_0a_0^3 - 240a_2\kappa_0^2a_0^3 + 124\kappa_0^3a_0^3 - 1056a_1^3a_3a_0^2 - 66a_1^3\kappa_1a_0^2 - 2544a_1^2a_2^2a_0^2 + 198a_1^2a_2\kappa_0a_0^2 + 354a_1^2\kappa_0^2a_0^2 + 2910a_1^4a_2a_0 \\
&\left.-150a_1^4\kappa_0a_0 - 863a_1^6\right)\tfrac{h^6}{60480a_0^6} + \left(3a_7a_0^6 + 18\kappa_5a_0^6 - 18a_1a_6a_0^5 - 30a_1\kappa_4a_0^5 - 70a_2a_5a_0^5 - 88a_2\kappa_3a_0^5 - 126a_3a_4a_0^5\right. \\
&-108a_3\kappa_2a_0^5 - 60a_4\kappa_1a_0^5 - 10a_5\kappa_0a_0^5 + 152\kappa_0\kappa_3a_0^5 + 360\kappa_1\kappa_2a_0^5 + 104a_1^2a_5a_0^4 + 74a_1^2\kappa_3a_0^4 + 610a_1a_2a_4a_0^4 + 252a_1a_2\kappa_2a_0^4 \\
&+423a_1a_3^2a_0^4 + 156a_1a_3\kappa_1a_0^4 + 10a_1a_4\kappa_0a_0^4 - 468a_1\kappa_0\kappa_2a_0^4 - 420a_1\kappa_1^2a_0^4 + 686a_2^2a_3a_0^4 + 108a_2^2\kappa_1a_0^4 + 4a_2a_3\kappa_0a_0^4 \\
&-600a_2\kappa_0\kappa_1a_0^4 - 142a_3\kappa_0^2a_0^4 + 372\kappa_0^2\kappa_1a_0^4 - 500a_1^3a_4a_0^3 - 126a_1^3\kappa_2a_0^3 - 3316a_1^2a_2a_3a_0^3 - 240a_1^2a_2\kappa_1a_0^3 + 86a_1^2a_3\kappa_0a_0^3 \\
&+948a_1^2\kappa_0\kappa_1a_0^3 - 1760a_1a_2^3a_0^3 + 168a_1a_2^2\kappa_0a_0^3 + 600a_1a_2\kappa_0^2a_0^3 - 248a_1\kappa_0^3a_0^3 + 1899a_1^4a_3a_0^2 + 48a_1^4\kappa_1a_0^2 + 6000a_1^3a_2^2a_0^2 \\
&\left.-522a_1^3a_2\kappa_0a_0^2 - 474a_1^3\kappa_0^2a_0^2 - 5310a_1^5a_2a_0 + 264a_1^5\kappa_0a_0 + 1375a_1^7\right)\tfrac{h^7}{120960a_0^7},
\end{aligned}$$

$$\begin{aligned}
\mathcal{E}_0^8 = 1 &- \tfrac{\kappa_0h^2}{2a_0} + \left(-\kappa_1a_0 + 2\,a_1\kappa_0\right)\tfrac{h^3}{6{a_0}^2} + \left(-\kappa_2{a_0}^2 + 3\,a_1\kappa_1a_0 + 3\,a_2\kappa_0a_0 + \kappa_0{}^2a_0 - 6\,a_1{}^2\kappa_0\right)\tfrac{h^4}{24{a_0}^3} + \left(-\kappa_3{a_0}^3 + 4\,a_1\kappa_2{a_0}^2\right. \\
&\left.+6\,a_2\kappa_1{a_0}^2 + 4\,a_3\kappa_0{a_0}^2 + 4\,\kappa_0\kappa_1{a_0}^2 - 12\,a_1{}^2\kappa_1a_0 - 24\,a_1a_2\kappa_0a_0 - 6\,a_1\kappa_0{}^2a_0 + 24\,a_1{}^3\kappa_0\right)\tfrac{h^5}{120\,{a_0}^4} + \left(-\kappa_4{a_0}^4 + 5\,a_1\kappa_3{a_0}^3\right. \\
&+10\,a_2\kappa_2{a_0}^3 + 10\,a_3\kappa_1{a_0}^3 + 5\,a_4\kappa_0{a_0}^3 + 7\,\kappa_0\kappa_2{a_0}^3 + 4\,\kappa_1{}^2{a_0}^3 - 20\,a_1{}^2\kappa_2{a_0}^2 - 60\,a_1a_2\kappa_1{a_0}^2 - 40\,a_1a_3\kappa_0{a_0}^2 \\
&\left.-31\,a_1\kappa_0\kappa_1{a_0}^2 - 30\,a_2{}^2\kappa_0{a_0}^2 - 13\,a_2\kappa_0{}^2{a_0}^2 - \kappa_0{}^3{a_0}^2 + 60\,a_1{}^3\kappa_1a_0 - 120\,a_1{}^4\kappa_0 + 180\,a_1{}^2a_2\kappa_0a_0 + 36\,a_1{}^2\kappa_0{}^2a_0\right)\tfrac{h^6}{720\,{a_0}^5} \\
&+ \left(-\kappa_5a_0^5 + 6a_1\kappa_4a_0^4 + 15a_2\kappa_3a_0^4 + 20a_3\kappa_2a_0^4 + 15a_4\kappa_1a_0^4 + 6a_5\kappa_0a_0^4 + 11\kappa_0\kappa_3a_0^4 + 15\kappa_1\kappa_2a_0^4 - 30a_1^2\kappa_3a_0^3 - 120a_1a_2\kappa_2a_0^3\right. \\
&\left.-120a_1a_3\kappa_1a_0^3 - 60a_1a_4\kappa_0a_0^3 - 66a_1\kappa_0\kappa_2a_0^3 - 39a_1\kappa_1^2a_0^3 - 90a_2^2\kappa_1a_0^3 - 120a_2a_3\kappa_0a_0^3 - 81a_2\kappa_0\kappa_1a_0^3 - 24a_3\kappa_0^2a_0^3 - 9\kappa_0^2\kappa_1a_0^3\right.
\end{aligned}$$

$$+120a_1^3\kappa_2a_0^2+540a_1^2a_2\kappa_1a_0^2+360a_1^2a_3\kappa_0a_0^2+228a_1^2\kappa_0\kappa_1a_0^2+540a_1a_2^2\kappa_0a_0^2+192a_1a_2\kappa_0^2a_0^2+12a_1\kappa_0^3a_0^2-360a_1^4\kappa_1a_0$$
$$-1440a_1^3a_2\kappa_0a_0-240a_1^3\kappa_0^2a_0+720a_1^5\kappa_0)\,\tfrac{h^7}{5040a_0^6}+(-\kappa_6a_0^6+7a_1\kappa_5a_0^5+21a_2\kappa_4a_0^5+35a_3\kappa_3a_0^5+35a_4\kappa_2a_0^5+21a_5\kappa_1a_0^5$$
$$+7a_6\kappa_0a_0^5+16\kappa_0\kappa_4a_0^5+26\kappa_1\kappa_3a_0^5+15\kappa_2^2a_0^5-42a_1^2\kappa_4a_0^4-210a_1a_2\kappa_3a_0^4-280a_1a_3\kappa_2a_0^4-210a_1a_4\kappa_1a_0^4-84a_1a_5\kappa_0a_0^4$$
$$-122a_1\kappa_0\kappa_3a_0^4-174a_1\kappa_1\kappa_2a_0^4-210a_2^2\kappa_2a_0^4-420a_2a_3\kappa_1a_0^4-210a_2a_4\kappa_0a_0^4-202a_2\kappa_0\kappa_2a_0^4-120a_2\kappa_1^2a_0^4-140a_3^2\kappa_0a_0^4$$
$$-40a_4\kappa_0^2a_0^4-174a_3\kappa_0\kappa_1a_0^4-22\kappa_0^2\kappa_2a_0^4-28\kappa_0\kappa_1^2a_0^4+210a_1^3\kappa_3a_0^3+1260a_1^2a_2\kappa_2a_0^3+1260a_1^2a_3\kappa_1a_0^3+630a_1^2a_4\kappa_0a_0^3$$
$$+345a_1^2\kappa_1^2a_0^3+572a_1^2\kappa_0\kappa_2a_0^3+1890a_1a_2^2\kappa_1a_0^3+2520a_1a_2a_3\kappa_0a_0^3+1422a_1a_2\kappa_0\kappa_1a_0^3+418a_1a_3\kappa_0^2a_0^3+130a_1\kappa_0^2\kappa_1a_0^3$$
$$+630a_2^3\kappa_0a_0^3+303a_2^2\kappa_0^2a_0^3+34a_2\kappa_0^3a_0^3+\kappa_0^4a_0^3-840a_1^4\kappa_2a_0^2-5040a_1^3a_2\kappa_1a_0^2-3360a_1^3a_3\kappa_0a_0^2-1800a_1^3\kappa_0\kappa_1a_0^2$$
$$-7560a_1^2a_2^2\kappa_0a_0^2-2280a_1^2a_2\kappa_0^2a_0^2-120a_1^2\kappa_0^3a_0^2+2520a_1^5\kappa_1a_0+12600a_1^4a_2\kappa_0a_0+1800a_1^4\kappa_0^2a_0-5040a_1^6\kappa_0)\,\tfrac{h^8}{40320a_0^7},$$

$$\mathcal{F}_0^6(h)=\tfrac{1}{2a_0}-\tfrac{a_1h}{3a_0^2}+(-3a_2a_0-\kappa_0a_0+6a_1^2)\,\tfrac{h^2}{24a_0^3}+(-4a_3a_0^2-3\kappa_1a_0^2+24a_1a_2a_0+6a_1\kappa_0a_0-24a_1^3)\,\tfrac{h^3}{120a_0^4}+(-5a_4a_0^3$$
$$-6\kappa_2a_0^3+40a_1a_3a_0^2+23a_1\kappa_1a_0^2+30a_2^2a_0^2+13a_2\kappa_0a_0^2+\kappa_0^2a_0^2-180a_1^2a_2a_0-36a_1^2\kappa_0a_0+120a_1^4)\,\tfrac{h^4}{720a_0^5}+(-3a_0^4a_5-5a_0^4\kappa_3$$
$$+30a_0^3a_1a_4+28a_0^3a_1\kappa_2+60a_0^3a_2a_3+30a_0^3a_2\kappa_1+12a_0^3a_3\kappa_0+4a_0^3\kappa_0\kappa_1-180a_0^2a_1^2a_3-84a_0^2a_1^2\kappa_1-270a_0^2a_1a_2^2-6a_0^2a_1\kappa_0^2$$
$$-96a_0^2a_1a_2\kappa_0+720a_0a_1^3a_2+120a_0a_1^3\kappa_0-360a_1^5)\,\tfrac{h^5}{2520a_0^6}+(-7a_0^5a_6-15a_0^5\kappa_4+84a_0^4a_1a_5+110a_0^4a_1\kappa_3+210a_0^4a_2a_4$$
$$+171a_0^4a_2\kappa_2+140a_0^4a_3^2+129a_0^4a_3\kappa_1+40a_0^4a_4\kappa_0+21a_0^4\kappa_0\kappa_2+18a_0^4\kappa_1^2-630a_0^3a_1^2a_4-482a_0^3a_1^2\kappa_2-2520a_0^3a_1a_2a_3$$
$$-1047a_0^3a_1a_2\kappa_1-418a_0^3a_1a_3\kappa_0-115a_0^3a_1\kappa_0\kappa_1-630a_0^3a_2^3-303a_0^3a_2^2\kappa_0-34a_0^3a_2\kappa_0^2-a_0^3\kappa_0^3+3360a_0^2a_1^3a_3+1320a_0^2a_1^3\kappa_1$$
$$+7560a_0^2a_1^2a_2^2+2280a_0^2a_1^2a_2\kappa_0+120a_0^2a_1^2\kappa_0^2-12600a_0a_1^4a_2-1800a_0a_1^4\kappa_0+5040a_1^6)\,\tfrac{h^6}{40320a_0^7},$$

$$\mathcal{F}_1^5(h)=\tfrac{1}{6a_0}-\tfrac{a_1h}{8a_0^2}+(-6a_2a_0-\kappa_0a_0+12a_1^2)\,\tfrac{h^2}{120a_0^3}+(-5a_3a_0^2-2\kappa_1a_0^2+30a_1a_2a_0+4a_1\kappa_0a_0-30a_1^3)\,\tfrac{h^3}{360a_0^4}+(-15a_0^3a_4$$
$$-10a_0^3\kappa_2+120a_0^2a_1a_3+39a_0^2a_1\kappa_1+90a_0^2a_2^2+21a_0^2a_2\kappa_0+a_0^2\kappa_0^2-540a_0a_1^2a_2-60a_0a_1^2\kappa_0+360a_1^4)\,\tfrac{h^4}{5040a_0^5}$$
$$+(-21a_0^4a_5-20a_0^4\kappa_3+210a_0^3a_1a_4+115a_0^3a_1\kappa_2+420a_0^3a_2a_3+120a_0^3a_2\kappa_1+45a_0^3a_3\kappa_0+10a_0^3\kappa_0\kappa_1-1260a_0^2a_1^2a_3$$
$$-345a_0^2a_1^2\kappa_1-1890a_0^2a_1a_2^2-375a_0^2a_1a_2\kappa_0-15a_0^2a_1\kappa_0^2+5040a_0a_1^3a_2+480a_0a_1^3\kappa_0-2520a_1^5)\,\tfrac{h^5}{40320a_0^6},$$

$$\mathcal{F}_2^4(h)=\tfrac{1}{24a_0}-\tfrac{a_1h}{30a_0^2}+(-10a_2a_0-\kappa_0a_0+20a_1^2)\,\tfrac{h^2}{720a_0^3}+(-4a_0^2a_3-a_0^2\kappa_1+24a_0a_1a_2+2a_0a_1\kappa_0-24a_1^3)\,\tfrac{h^3}{1008a_0^4}+(-35a_0^3a_4$$
$$-15a_0^3\kappa_2+280a_0^2a_1a_3+59a_0^2a_1\kappa_1+210a_0^2a_2^2+31a_0^2a_2\kappa_0+a_0^2\kappa_0^2-1260a_0a_1^2a_2-90a_0a_1^2\kappa_0+840a_1^4)\,\tfrac{h^4}{40230a_0^5},$$

$$\mathcal{F}_3^3(h)=\tfrac{1}{120a_0}-\tfrac{a_1h}{144a_0^2}+(-15a_0a_2-a_0\kappa_0+30a_1^2)\,\tfrac{h^2}{5040a_0^3}+(-35a_0^2a_3-6a_0^2\kappa_1+210a_0a_1a_2+12a_0a_1\kappa_0-210a_1^3)\,\tfrac{h^3}{40320a_0^4},$$

$$\mathcal{F}_4^2(h)=\tfrac{1}{720a_0}-\tfrac{a_1h}{840a_0^2}+(-21a_0a_2-a_0\kappa_0+42a_1^2)\,\tfrac{h^2}{40320a_0^3},\qquad \mathcal{F}_5^1(h)=\tfrac{1}{5040a_0}-\tfrac{a_1h}{5760a_0^2},\qquad \mathcal{F}_6^0(h)=\tfrac{1}{40320a_0}.$$

Define $\mathcal{E}^n_{0,\pm},\mathcal{E}^n_{1,\pm},\mathcal{F}^n_{\ell,\pm}$ for $n\in\mathbb{N}_0$ to be just $\mathcal{E}^n_0,\mathcal{E}^n_1,\mathcal{F}^n_\ell$ as in (2.46), respectively but using $a_j=a^{(j)}(x_b\pm)$, $\kappa_j=[\kappa^2]^{(j)}(x_b\pm)$ and $f_j=f^{(j)}(x_b\pm)$. Let $M=\tilde{M}\in 2\mathbb{N}$. For the left boundary condition $\mathcal{B}^+u(x_b)=\lambda_0u(x_b+)+\lambda_1u'(x_b+)$, we deduce from (2.34) of Theorem 2.5 that

$$c_0^{\mathcal{B}^+}(h)u(x_b)+c_1^{\mathcal{B}^+}(h)u(x_b+h)=h\mathcal{B}^+u(x_b)+\sum_{\ell=0}^{M-2}d_\ell^{\mathcal{B}^+}(h)h^{\ell+2}f^{(\ell)}(x_b+)+\mathscr{O}(h^{M+1}),$$

as $h\to 0$, where one particular choice of $c_1^{\mathcal{B}^+}$, $c_0^{\mathcal{B}^+}$, and $d_\ell^{\mathcal{B}^+}$ for $\ell=0,\dots,M-2$ satisfying (2.35) and (2.36) of Theorem 2.5 are given by

$$\begin{aligned}&c_1^{\mathcal{B}^+}(h):=\lambda_1\mathcal{E}^{M-1}_{1,+}(h),\qquad c_0^{\mathcal{B}^+}(h):=h\lambda_0-\lambda_1\mathcal{E}^{M-1}_{1,+}(h)\mathcal{E}^M_{0,+}(h),\\&d_\ell^{\mathcal{B}^+}(h):=\lambda_1\mathcal{E}^{M-1}_{1,+}(h)\mathcal{F}^{M-\ell-2}_{\ell,+}(h),\qquad \ell=0,\dots,M-2.\end{aligned}\tag{2.47}$$

Similarly, let the boundary condition at x_b for the left side of x_b be given by

$$\mathcal{B}^- u(x_b) := \lambda_0 u(x_b-) + \lambda_1 u'(x_b-) \quad \text{with} \quad \lambda_0, \lambda_1 \in \mathbb{C}. \tag{2.48}$$

By symmetry, the discretization at the base point x_b from the left side of x_b is

$$c_{-1}^{\mathcal{B}^-}(h)u(x_b-h) + c_0^{\mathcal{B}^-}(h)u(x_b) = h\mathcal{B}^- u(x_b) + \sum_{\ell=0}^{M-2} d_\ell^{\mathcal{B}^-}(h) h^{\ell+2} f^{(\ell)}(x_b-) + \mathscr{O}(h^{M+1}), \tag{2.49}$$

as $h \to 0$, which satisfies the corresponding relations in (2.37) and (2.38) if

$$\begin{aligned} &c_{-1}^{\mathcal{B}^-}(h) := -\lambda_1 \mathcal{E}_{1,-}^{M-1}(-h), \qquad c_0^{\mathcal{B}^-}(h) := h\lambda_0 + \lambda_1 \mathcal{E}_{1,-}^{M-1}(-h)\mathcal{E}_{0,-}^{M}(-h), \\ &d_\ell^{\mathcal{B}^-}(h) := (-1)^{\ell+1}\lambda_1 \mathcal{E}_{1,-}^{M-1}(-h)\mathcal{F}_{\ell,-}^{M-\ell-2}(-h), \qquad \ell = 0, \ldots, M-2. \end{aligned} \tag{2.50}$$

For the stencil used at the breaking/branch point x_b such that $w(\delta_{x_b})$ is the weight of the Dirac distribution δ_{x_b} of the source term f, we deduce from (2.47) and (2.50) with $\lambda_0 = 0$ and $\lambda_1 = 1$ that

$$\begin{aligned} &-\alpha \mathcal{E}_{1,-}^{M-1}(-h)u(x_b-h) + \left[\alpha \mathcal{E}_{1,-}^{M-1}(-h)\mathcal{E}_{0,-}^{M}(-h) + \beta \mathcal{E}_{1,+}^{M-1}(h)\mathcal{E}_{0,+}^{M}(h)\right] u(x_b) \\ &\quad - \beta \mathcal{E}_{1,+}^{M-1}(h)u(x_b+h) = -h\gamma w(\delta_{x_b}) \\ &\quad - \sum_{\ell=0}^{M-2} h^{\ell+2}[\beta \mathcal{E}_{1,+}^{M-1}(h)\mathcal{F}_{\ell,+}^{M-\ell-2}(h) f^{(j)}(x_b+) + \alpha(-1)^\ell \mathcal{E}_{1,-}^{M-1}(-h)\mathcal{F}_{\ell,-}^{M-\ell-2}(-h) f^{(j)}(x_b-)] \\ &\quad + \mathscr{O}(h^{M+1}), \end{aligned} \tag{2.51}$$

as $h \to 0$, where $\alpha := \frac{2a(x_b-)}{a(x_b+)+a(x_b-)}$, $\beta := \frac{2a(x_b+)}{a(x_b+)+a(x_b-)}$ and $\gamma := \frac{2}{a(x_b+)+a(x_b-)}$.

Finite difference schemes with lower accuracy orders $M = 2, 4, 6$ can be easily obtained by truncating the above given $\mathcal{E}_0^8, \mathcal{E}_1^7, \mathcal{F}_\ell^{6-\ell}$ accordingly. In the above compact FDM with accuracy order M with $M = \tilde{M} \in 2\mathbb{N}$, we only need $a, a', \ldots, a^{(M-1)}$, $\kappa^2, [\kappa^2]', \ldots, [\kappa^2]^{(M-2)}$, and $f, f', \ldots, f^{(M-2)}$.

2.3 Convergence of DAT using compact FDMs

In this section, we discuss the convergence of DAT in Section 2.1 using compact FDMs described in Theorems 2.4 and 2.5 of Section 2.2. Let us first outline the notations and assumptions for our discussion in this section and for our numerical experiments in the next section. Let $0 = b_0 < b_1 < \cdots < b_p < b_{p+1} = 1$ with $p \in \mathbb{N} \cup \{0\}$. The coefficients a, κ^2 and f in (2.1) are piecewise smooth in the sense that they have uniformly continuous derivatives

of orders up to n_M on (b_j, b_{j+1}) for all $j = 0, \ldots, p$ for certain given integer $n_M \in \mathbb{N}$ (see Section 2.2 for details). Note that these coefficients may be discontinuous on $(0,1)$ and we call the points $b_1, \ldots, b_p$ breaking/branch points. For simplicity of discussion, we assume a Dirichlet boundary condition at 0, while a Dirichlet, Neumann, or Robin boundary condition at 1. Let u_e be the exact weak solution to (2.1) with the boundary conditions in (2.2). Let $N \in \mathbb{N}$ and $0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$ for the computational mesh with the average mesh size $h := N^{-1}$. Let $\{u_N(x_j)\}_{j=0}^N$ be the approximated solution on knot points $\{x_j\}_{j=0}^N$. To study theoretical convergence rates and to evaluate the performance of DAT, we define

$$\|u_N - u_e\|_\infty := \max_{0 \leqslant j \leqslant N} |u_N(x_j) - u_e(x_j)|, \quad \|u_N - u_e\|_2^2 := \sum_{j=0}^{N} h_j |u_N(x_j) - u_e(x_j)|^2 \tag{2.52}$$

with $h_j := x_{j+1} - x_j$ and $x_{N+1} := 1$. Because $\|u_N - u_e\|_2 \leqslant \|u_N - u_e\|_\infty$ and $\|u_N' - u_e'\|_2 \leqslant \|u_N' - u_e'\|_\infty$ always hold, we shall only discuss the convergence in ∞-norm instead of 2-norm. Throughout this section, positive constants C, C_1, C_2 are always independent of both matrix size N and mesh size h. The computational mesh is assumed to be quasi-uniform, i.e., there exists $C > 0$ independent of h such that $C^{-1}h \leqslant h_j \leqslant Ch$ for all $j = 0, \ldots, N$. Note that the weak solution $u_e \in H^1(0,1)$ but u_e' may be discontinuous at branch points on $(0,1)$, because the coefficients a, κ^2, f are only piecewise smooth. For convenience, every branch point b_j is assumed to be a grid/knot point and the mesh on each piece (b_j, b_{j+1}) is uniform for all $j = 0, \ldots, p$. These restrictions could be dropped as we already discussed in Section 2.2.3 but they make our discussion here and implementation in Section 2.4 much simpler.

For an Mth order compact FDM in Section 2.2 with $\tilde{M} = M \in \mathbb{N}$, the stencil at x_j:

$$c_{j,-1}(h)u(x_{j-1}) + c_{j,0}(h)u(x_j) + c_{j,1}(h)u(x_{j+1}) = F_j(h), \qquad j = 1, \ldots, N \tag{2.53}$$

is given in Theorems 2.4 and 2.5 as follows:

(1) If x_j is neither a branch point nor a boundary point, then (2.53) is given by (2.25) with (2.26) and (2.27) under the normalization condition $\alpha(0) = -1$. Because we impose a Dirichlet boundary condition at 0, the known term $c_{1,-1}(h)u(x_0)$ is moved to and combined with F_j.

(2) If x_j is a branch point, then (2.53) is given by (2.42) or more explicitly (2.51).

(3) For x_N (right boundary point), (2.53) is given by (2.49) with (2.50). Note that $c_{N,N+1}(h) = 0$. If we impose a Dirichlet boundary condition at 1, the known term

$c_{N-1,1}(h)u(x_N)$ is moved to and combined with F_j such that the last term in (2.53) is $N-1$ (instead of N).

The approximated numerical solution $u_N = \{u_N(x_j)\}_{j=0}^N$ is obtained by solving the linear system in (2.53) with $u = u_N$. By Theorems 2.4 and 2.5, the exact solution u_e must satisfy

$$c_{j,-1}(h)u_e(x_{j-1}) + c_{j,0}(h)u_e(x_j) + c_{j,1}(h)u_e(x_{j+1}) = F_j(h) + R_j(h), \qquad j = 1, \ldots, N, \tag{2.54}$$

where the local truncation error functions $R_j(h)$ resulted from Taylor approximation satisfy

$$|R_j(h)| \leqslant \begin{cases} Ch^{M+2}, & \text{if } x_j \notin \{b_0, \ldots, b_{p+1}\}, \text{ i.e., } x_j \text{ is an interior point,} \\ Ch^{M+1}, & \text{if } x_j \in \{b_0, \ldots, b_{p+1}\}, \text{ i.e., } x_j \text{ is a branch point or a boundary point,} \end{cases} \tag{2.55}$$

where the constant C is independent of h and only depends on derivatives of u_e, a, κ^2 and f. The error is then defined by $Q_j := u_N(x_j) - u_e(x_j)$ at the knot point x_j for $j = 1, \ldots, N$. By (2.54) and (2.55),

$$c_{j,-1}(h)Q_{j-1} + c_{j,0}(h)Q_j + c_{j,1}(h)Q_{j+1} = R_j(h), \qquad j = 1, \ldots, N, \tag{2.56}$$

which can be put together into the following matrix form

$$A(h)\vec{Q} = \vec{R}(h) \quad \text{with} \quad \vec{Q} := [Q_1, \ldots, Q_N]^\mathsf{T}, \vec{R}(h) := [R_1, \ldots, R_N]^\mathsf{T}, \tag{2.57}$$

where $A(h)$ is an $N \times N$ tridiagonal matrix defined by

$$A(h) = \text{tridiag}(\{c_{j,-1}(h)\}_{j=2}^N, \{c_{j,0}(h)\}_{j=1}^N, \{c_{j,1}(h)\}_{j=1}^{N-1}). \tag{2.58}$$

Setting $h = 0$, we have a related $N \times N$ constant tridiagonal matrix $A(0)$ given by

$$A(0) = \text{tridiag}(\{c_{j,-1}(0)\}_{j=2}^N, \{c_{j,0}(0)\}_{j=1}^N, \{c_{j,1}(0)\}_{j=1}^{N-1}), \tag{2.59}$$

where tridiag$(\cdot,\cdot,\cdot)$ is defined in (S2) of Algorithm 2.1 and the entries of $A(0)$ are given as follows:

$$\begin{cases} c_{j,-1}(0) = c_{j,1}(0) = -1, \quad c_{j,0}(0) = 2 & \text{if } x_j \text{ is an interior point,} \\ c_{j,0}(0) = 2, \quad c_{j,-1}(0) = -\frac{2a(b_j-)}{a(b_j+)+a(b_j-)}, \quad c_{j,1}(0) = -2 - c_{j,-1}(0), & \text{if } x_j \text{ is a branch point,} \\ c_{0,0}(0) = 2, \quad c_{0,1}(0) = -1 & \text{if } \lambda_1^L = 0,\ \lambda_0^L = 1 \text{ in (2.2)}, \\ c_{N,-1}(0) = -1, \quad c_{N,0}(0) = 1, & \text{if } \lambda_1^R \neq 0 \text{ in (2.2)}. \end{cases}$$

If we impose a Dirichlet boundary condition at 1, then we replace N with $N-1$ in (2.54), (2.56), (2.57), (2.58), and (2.59). Furthermore, $c_{N-1,-1}(0) = -1$ and $c_{N-1,0}(0) = 2$ in (2.59).

Next, we highlight some key issues as to why the theoretical convergence of compact FDMs in Section 2.2 for 1D heterogeneous Helmholtz equations with various boundary conditions requires a separate comprehensive treatment and will be addressed elsewhere. First, the solution stability of such Helmholtz equations is far from trivial and warrants further investigation. There are some cases in which the stability constant may exponentially rise; i.e., the solution is close to being 'unstable' in some sense. In fact, the solution may become highly unstable under perturbation or even with fairly accurate approximation of boundary and source data. See [61, Section 5.2]. In these situations, the convergence of FDM (and any other discretization methods) is severely affected. For illustration purposes, we mention two such cases by considering the simplest Helmholtz equation:

$$u'' + \kappa^2 u = f \quad \text{on} \quad [0,1] \quad \text{with} \quad u(0) = u(1) = 0, \quad \text{a constant wavenumber } \kappa > 0. \tag{2.60}$$

First, it is well known that solving the simplest Helmholtz equation in (2.60) with large wavenumbers κ is challenging, because the huge stability constant grows quickly with κ^2 and causes the pollution effect. This requires the mesh size h to be extremely small for any numerical schemes to start effectively approximating the true solution and exhibiting convergence behavior. Second, if $\kappa = n\pi$ with $n \in \mathbb{N}$, then the solution u to (2.60) is obviously not unique since $u(x) + \alpha \sin(n\pi x)$ are also solutions to (2.60) for all $\alpha \in \mathbb{C}$. Now consider (2.60) with $\kappa = n\pi \pm \epsilon$ with $n \in \mathbb{N}$ and a very small $\epsilon > 0$. Though the solution to (2.60) is now unique and the wavenumber κ is quite small, as we shall explain later, its true solution is highly unstable in some sense. One has to use a small mesh size h in proportion to ϵ (which may be smaller than machine precision) for any numerical scheme to start effectively approximating the true solution and exhibiting convergence behavior. These phenomena and difficulties call for further investigation of the stability of Helmholtz equations and its relations to convergence properties of FDMs. Because DAT can break any large problem into very small ones, the above discussion in fact shows the advantages and contributions of DAT for numerical solutions of Helmholtz equations.

Recall that for an $m \times n$ matrix A, the ∞-norm of A is $\|A\|_\infty := \sup_{1\leqslant j\leqslant m} \sum_{k=1}^n |A_{j,k}|$, which is the operator norm mapping ℓ_∞^n to ℓ_∞^m. In the convergence analysis of FDM, one can deduce from the identity (2.57) that

$$\|\vec{Q}\|_\infty := \sup_{1\leqslant j\leqslant N} |Q_j| \leqslant \|A(h)^{-1}\vec{R}(h)\|_\infty \leqslant \|A(h)^{-1}\|_\infty \|\vec{R}(h)\|_\infty. \tag{2.61}$$

Hence, how $\|A(h)^{-1}\|_\infty$ behaves for small h is a key issue. Even though $A(h)$ in (2.57) converges entrywise to $A(0)$ in (2.59), the invertibility of $A(h)$ and the norm estimates of $\|A(h)^{-1}\|_\infty$ are not immediately guaranteed by the properties of $A(0)$ in (2.59), since the size N of $A(h)$ goes to ∞ as $h \to 0$. Furthermore, the structure of $A(h)^{-1}$ may be unknown. In stark contrast to elliptic equations, $A(h)$ may be singular or highly ill-conditioned not only for large wavenumbers but also for small wavenumbers. Let us consider the simplest Helmholtz equation in (2.60) again and use the standard second order FDM. Then at mesh size $h = N^{-1}$, our coefficient matrix is $A(h) = \text{tridiag}(\{-1\}_{j=2}^{N-1}, \{2 - \kappa^2 h^2\}_{j=1}^{N-1}, \{-1\}_{j=1}^{N-2})$, whose eigenvalues are known to take the following form

$$\sigma_n := (2 - \kappa^2 h^2) - 2\cos(nh\pi), \quad \forall n = 1, \ldots, N-1.$$

Note that the nth eigenvalue, σ_n, vanishes and hence $\det(A(h)) = 0$ if

$$\kappa = \kappa_*(h, n), \quad \text{where} \quad \kappa_*(h, n) := h^{-1}\sqrt{2(1 - \cos(nh\pi))}. \tag{2.62}$$

This situation is not encountered in the elliptic case, since all its eigenvalues $(2 + \kappa^2 h^2) - 2\cos(nh\pi) > 0$ for all $n \in \mathbb{N}$. Consider $\kappa = \kappa_*(2^{-7}, 3) \approx 9.4226$, that is, $\kappa = 3\pi - \epsilon$ for some $0 < \epsilon < 0.0022$. Then the standard second order FDM fails to produce any solution at $h = 2^{-7}$ because $\det(A(h)) = 0$. For $\kappa > 0$, we define the distance $\rho_\kappa := \min_{n\in\mathbb{N}} |\kappa - n\pi|$, which can be arbitrarily small for any mesh size h. For example, for the mesh size $h = 2^{-19}$, we see that $\rho_\kappa \approx 10^{-10}$ with $\kappa := \kappa_*(2^{-19}, 3) \approx 9.4248$ but $\det(A(h)) = 0$ at $h = 2^{-19}$. Note that the commonly used criterion $\kappa^2 h = \mathscr{O}(1)$ is satisfied because $\kappa^2 h \approx 2 \times 10^{-4}$ with $\kappa = \kappa_*(2^{-19}, 3)$ and $h = 2^{-19}$. However, we need to employ an impractically small grid size for the FDM before any convergence is perceived. The situation is exacerbated if κ is very large and ρ_κ is very small. The foregoing point first demonstrates how we need to carefully quantify and elaborate on what 'sufficiently small h' means for some form of convergence in the pre-asymptotic (computationally feasible) range to take place, which theoretically may be challenging (much harder than elliptic equations); and second, it refers back to an earlier key issue regarding the significance of understanding the solution's stability. For the example presented above, one can check by a direct calculation that the energy norm of the true solution is large. The theoretical convergence for 1D heterogeneous Helmholtz equations with piecewise smooth coefficients demands more sophisticated analysis due to its underlying intricacies.

Before we turn to the convergence of DAT, we discuss how to estimate $u'(x_b)$ for $x_b = x_j$ for some $0 \leqslant j \leqslant N$ from $u(x_k) := u_N(x_k), k = 0, \ldots, N$, since u'_N is used in the linking problems of DAT and in the error $\|u'_N - u'_e\|_\infty$ for measuring performance. Assume that the

numerical u (i.e., u_N) is computed with accuracy order M, that is, $|u(x_k) - u_e(x_k)| \leqslant Ch^M$ for all $k = 0, \ldots, N$ for some $C > 0$ independent of N and h. We can estimate one-sided derivatives $u'(x_b+)$ and $u'(x_b-)$ with the same accuracy order as well. Basically, let $\mathcal{L}_h^{\mathcal{B}^+} u(x_b)$ and $\mathcal{L}_h^{\mathcal{B}^-} u(x_b)$ be the stencils with accuracy order M for boundary conditions in (2.41) through Theorem 2.5. Then

$$u_e'(x_b+) = \mathcal{L}_h^{\mathcal{B}^+} u(x_b) + \mathscr{O}(h^M), \qquad u_e'(x_b-) = \mathcal{L}_h^{\mathcal{B}^-} u(x_b) + \mathscr{O}(h^M), \qquad h \to 0, \tag{2.63}$$

which can be also derived from (2.39) easily. Higher order one-sided derivatives at x_b can also be estimated with accuracy order M thanks to Proposition 2.3. Moreover, since we can obtain the one-sided derivatives of u_e at all knot points with accuracy order M, using interpolation we can obtain a function $u(x), 0 \leqslant x \leqslant 1$ from the computed data $\{u(x_j)\}_{j=0}^N$ such that u accurately approximates the exact solution u_e in the function setting. The identities in (2.63) play a critical role in DAT to accurately estimate artificial Dirac distributions in (2.9) and (2.12) for DAT.

Now we are ready to discuss the convergence of DAT. Recall that the average mesh size $h := N^{-1}$. Assume that the Helmholtz equation in (2.1)–(2.2) has a unique solution. Let N_0 be a given integer independent of N and h. Now we claim that

If all local problems in DAT are at most $N_0 \times N_0$ in size, then all the condition numbers of all local problems in DAT must be uniformly bounded and DAT using the Mth order compact FDM exhibits $\mathscr{O}(h^M)$ convergence for sufficiently small h. (2.64)

The argument is as follows. According to the theory of DAT in Section 2.1, the accuracy of DAT only depends on the accuracy of the local problem solver and the error accumulated from the tree depth and the linking problems. So, let us look at one typical local problem with grid points $\alpha = x_L < x_{L+1} < \cdots < x_{H-1} < x_H = \beta$ on (α, β). For small h, as explained in Section 2.1 on DAT, the boundary conditions for this typical local problem are either Dirichlet boundary conditions at both α and β with at most one branch point inside (α, β), or Dirichlet boundary condition at 0 and the prescribed boundary condition as in (2.2) at 1. Let m be the size of this local problem. Then the relation in (2.57) still holds with $N = m$, $\vec{U} := [U_{L+1}, \ldots, U_{L+m}]^\mathsf{T}$ and $\vec{R}(h) = [R_L, \ldots, R_{L+m}]^\mathsf{T}$. Because the size $m \leqslant N_0$, we have $\lim_{h\to 0} \|A(h) - A(0)\|_\infty = 0$, where the $m \times m$ matrix $A(0)$ is given in (2.59). If the local mesh $\{x_L, \ldots, x_R\}$ does not contain any branch point, then $A(0)$ must be the standard $m \times m$ tridiagonal matrix generated by $[-1, 2, -1]$, probably with $[A(0)]_{m,m} = 1$ instead of 2 depending on the boundary condition at β. The later matrix $A(0)$ is known to be invertible with

$\det(A(0)) = m+1$, or 1 if $[A(0)]_{m,m} = 1$. Suppose now that the local mesh contains a branch point b_j and the kth row of $A(0)$ corresponds to this branch point b_j. Then the kth row of the standard tridiagonal matrix $A(0)$ with $[-1, 2, -1]$ is replaced by $[\frac{-2a(b_j-)}{a(b_j+)+a(b_j-)}, 2, \frac{-2a(b_j+)}{a(b_j+)+a(b_j-)}]$. Then $\det(A(0)) = 2((1+m-k)a(b_j-)+ka(b_j+))(a(b_j-)+a(b_j+))^{-1}$ if a Dirichlet boundary condition is imposed at β, or $\det(A(0)) = 2a(b_j-)(a(b_j-)+a(b_j+))^{-1}$ if a Neumann/Robin boundary condition is imposed at β. In all cases, the determinant of $A(0)$ is nonzero; thus, $A(0)$ must be an invertible matrix. Because $A(h)$ is at most $N_0 \times N_0$, we conclude that $A(h)$ is invertible for all sufficiently small h, $\lim_{h\to 0} \|A(h)^{-1} - A(0)^{-1}\|_\infty = 0$, and there exists $C_1 > 0$ independent of h such that $\|A(h)^{-1}\|_\infty \leqslant C_1$ for all small $h > 0$. Hence, the condition number of $A(h)$ is uniformly bounded for all local problems and we deduce from (2.57) and (2.61) that $\|\vec{Q}\|_\infty \leqslant C_1\|\vec{R}(h)\|_\infty$. If we use the Mth order compact FDM, then (2.55) must hold and hence $\|\vec{R}(h)\|_\infty \leqslant Ch^{M+1}$ for all sufficiently small h. Putting everything together, we proved that $\|\vec{Q}\|_\infty \leqslant C_1 C h^{M+1}$ for convergence of all local problems in (S2) of Algorithm 2.1.

For the linking problems, we have to estimate one-sided derivatives u' for approximated solutions u of all local problems. As we discussed before, this can be done by using (2.63) with M being replaced by $M+1$, because the local problems are solved with accuracy order $M+1$ as we discussed a moment ago. However, we cannot expect from (2.63) to achieve $\|u'_e - u'\|_\infty \leqslant C_2 h^{M+1}$ with a positive constant C_2 independent of h, where u_e and u stand for the exact solution and approximated solution of a local problem. Note that the constant C_2 only depends on a, κ and the partitioned source term $f_j = f\varphi_j$, where φ_j is the hat function supported on $[\alpha, \beta]$ with $\varphi_j(\gamma) = 1$ for some $\gamma \in [\alpha, \beta]$. However, $\beta - \alpha = \mathscr{O}(h)$ due to $m \leqslant N_0$ and hence, $\|\varphi'_j\|_\infty = \mathscr{O}(h^{-1})$. Consequently, one can observe that $\|f_j^{(n)}\|_\infty \leqslant C_3 h^{-1}$ for all $n = 0, \ldots, M$, where the positive constant C_3 only depends on f and is independent of h. That is, we can only expect $C_2 \leqslant C_3 h^{-1}$ and consequently, $\|u'_e - u'\|_\infty \leqslant C_2 h^{M+1} \leqslant C_3 h^M$. It is hard to exactly quantify how the error propagates from the deepest tree level to the surface tree level through the linking problems. Our numerical experiments seem to indicate that the linking problems do not further reduce accuracy. Because the one-sided derivatives u' can be estimated with accuracy $\mathscr{O}(h^M)$, the solution $\vec{Q}$ is expected to behave like $\|\vec{Q}\|_\infty \leqslant CC_1C_2h^{M+1} \leqslant CC_1C_3h^M$ for sufficiently small h. This leads to the claim in (2.64).

2.4 Numerical experiments

In this section, we present several numerical experiments to illustrate the performance of DAT in Section 2.1 and the developed compact FDMs in Section 2.2. Let u_e and $\{u_N(x_j)\}_{j=0}^N$ be

the exact (if its analytic expression is known) and approximated solutions on knot points $\{x_j\}_{j=0}^N$ with $0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$, respectively. Because 2-norm is controlled by ∞-norm in (2.52), we shall measure the performance in ∞-norm using relative errors $\frac{\|u_N - u_e\|_\infty}{\|u_e\|_\infty}$ and $\frac{\|u'_N - u'_e\|_\infty}{\|u'_e\|_\infty}$, where $\{u'_N(x_j)\}_{j=0}^N$ are estimated from $\{u_N(x_j)\}_{j=0}^N$ through (2.63). When the analytic expression of the exact solution u_e is unknown, we calculate the relative error between two consecutive levels. Due to the pollution effect, we know that our grid size has to be at least smaller than $\|\kappa\|_\infty^{-1}$. When we perform our experiments, we initially set our grid size to be approximately $\|\kappa\|_\infty^{-1}$, refine dyadically, and only record the numerical results where a convergent behaviour is present (either with respect to the exact solution or the solution at the subsequent grid refinement). All condition numbers are approximated by using `condest` in MATLAB, after renormalizing all the diagonal entries to be one in the coefficient matrices. The columns "Local CN" and "Link CN" in all tables in this section list the maximum condition numbers associated with local and link problems in DAT. The tree level and split parameter used are denoted by ℓ and s. The default choice is $s = 1$. Also, $\ell = 0$ means we use FDM without DAT. All linear systems are solved by using MATLAB's backslash command. For all examples below, we use the Mth order compact FDMs in Section 2.2.4 with $M = 6$ or $M = 8$. To visualize the numerical performance, the vertical axis in each convergence plot uses a base-10 log scale and the horizontal axis uses a base-2 log scale.

2.4.1 A comparison with PUFEM

DAT differs from PUFEM (see [8, 104]) in the way the partition of unity is applied. The former multiplies the partition of unity with the source term f, while the latter multiplies the partition of unity with local approximation spaces. In the presence of a large wavenumber, the trial functions in PUFEM are highly oscillatory. Hence, finding an appropriate quadrature becomes a major concern and challenge for PUFEM. Moreover, numerical experiments in [121] indicate that the coefficient matrix of PUFEM has an extremely large condition number, which may produce extra stability issues. For the heterogeneous Helmholtz equation with piecewise smooth coefficients and wavenumber, finding suitable local approximation spaces can in fact be challenging and computationally expensive.

Example 2.1. Consider the model problem (2.1)-(2.2) given by $[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x), x \in (0,1)$ with the coefficients $a = 1$, $\kappa = 10^6$, $f = \kappa^2 \cosh(x)$, and the boundary conditions $u(0) = 0$ and $u'(1) - \mathrm{i}\kappa u(1) = 0$. The exact solution has the following analytic expression

$$u_e = \frac{-\kappa \sin(\kappa x)}{\kappa^2 + 1}(\sinh(1) - \mathrm{i}\cosh(1)\kappa)e^{\mathrm{i}\kappa} + \frac{\kappa^2(\cosh(x) - e^{\mathrm{i}\kappa x})}{\kappa^2 + 1}.$$

See Table 2.1 for the numerical performance measured by $\frac{\|u_N-u_e\|_\infty}{\|u_e\|_\infty}$ and $\frac{\|u'_N-u'_e\|_\infty}{\|u'_e\|_\infty}$. The errors for PUFEM are evaluated at nodal points. Because the wavenumber $\kappa = 10^6$ is large, to fairly compare DAT with PUFEM, all inner products in PUFEM are calculated exactly via symbolic computation to minimize possible errors due to numerical quadrature. "Local CN" for PUFEM lists the condition number of its coefficient matrix. All local and linking problems in DAT in Table 2.1 solve at most 4×4 linear systems with uniformly bounded small condition numbers. Table 2.1 demonstrates that DAT can handle very small mesh size and the maximum condition numbers of coefficient matrices coming from all local and linking problems are much smaller than those in FDM and PUFEM by several orders of magnitude.

N	ℓ	$\frac{\|u_N-u_e\|_\infty}{\|u_e\|_\infty}$	$\frac{\|u'_N-u'_e\|_\infty}{\|u'_e\|_\infty}$	Local CN	Link CN	$\frac{\|u_N-u_e\|_\infty}{\|u_e\|_\infty}$	$\frac{\|u'_N-u'_e\|_\infty}{\|u'_e\|_\infty}$	Local CN	Link CN
		DAT using the compact FDM with order $M = 6$				DAT using the compact FDM with order $M = 8$			
2^{21}	0	1.7276×10^{-1}	4.0087×10^{-1}	1.61×10^{7}	–	4.3849×10^{-4}	1.0173×10^{-3}	2.07×10^{7}	–
2^{21}	19	1.7276×10^{-1}	4.0087×10^{-1}	3.23×10^{1}	4.18×10^{2}	4.3849×10^{-4}	1.0173×10^{-3}	3.23×10^{1}	6.27×10^{1}
2^{22}	0	2.6379×10^{-3}	6.1212×10^{-3}	7.15×10^{7}	–	1.6674×10^{-6}	3.8683×10^{-6}	8.65×10^{7}	–
2^{22}	20	2.6379×10^{-3}	6.1212×10^{-3}	4.15×10^{1}	6.37×10^{1}	1.6671×10^{-6}	3.8677×10^{-6}	4.15×10^{1}	6.28×10^{1}
2^{23}	0	4.0945×10^{-5}	9.5012×10^{-5}	3.51×10^{8}	–	8.1795×10^{-9}	1.8977×10^{-8}	3.51×10^{8}	–
2^{23}	21	4.0946×10^{-5}	9.5014×10^{-5}	4.41×10^{1}	6.29×10^{1}	1.1594×10^{-8}	2.6901×10^{-8}	4.41×10^{1}	6.28×10^{1}
		PUFEM in [9]							
2^{21}	–	1.2806×10^{-1}	5.7930×10^{-1}	2.09×10^{7}	–				
2^{22}	–	3.2473×10^{-2}	2.1123×10^{-1}	8.69×10^{7}	–				
2^{23}	–	8.1473×10^{-3}	8.9740×10^{-2}	2.38×10^{8}	–				

Table 2.1: Relative errors for Example 2.1 using DAT with $N_0 = 4$ and $s = 1$ in Algorithm 2.1, and PUFEM. The grid increment used in $[0, 1]$ is N^{-1}.

2.4.2 Numerical experiments on 1D heterogeneous Helmholtz equations

Example 2.2. Consider the model problem (2.1)-(2.2) given by $[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x), x \in (0, 1)$ with the following piecewise smooth jumping coefficients having a large variation:

$$
\begin{aligned}
a &= \chi_{[0,\frac{1}{8})} + 10^{-1}\chi_{[\frac{1}{8},\frac{2}{8})} + \chi_{[\frac{2}{8},\frac{3}{8})} + 10^{-2}\chi_{[\frac{3}{8},\frac{4}{8})} + \chi_{[\frac{4}{8},\frac{5}{8})} + 10^{-3}\chi_{[\frac{5}{8},\frac{6}{8})} + \chi_{[\frac{6}{8},\frac{7}{8})} + 10^{-4}\chi_{[\frac{7}{8},1]},\\
\kappa &= 10^{4}(\chi_{[0,\frac{1}{8})\cup[\frac{2}{8},\frac{3}{8})\cup[\frac{4}{8},\frac{5}{8})\cup[\frac{6}{8},\frac{7}{8})}) + 500(\chi_{[\frac{1}{8},\frac{2}{8})\cup[\frac{3}{8},\frac{4}{8})\cup[\frac{5}{8},\frac{6}{8})\cup[\frac{7}{8},1]}),\\
f &= 10^{7}e^{x}(\chi_{[0,\frac{1}{8})\cup[\frac{2}{8},\frac{3}{8})\cup[\frac{4}{8},\frac{5}{8})\cup[\frac{6}{8},\frac{7}{8})}) - e^{-2x}(\chi_{[\frac{1}{8},\frac{2}{8})\cup[\frac{3}{8},\frac{4}{8})\cup[\frac{5}{8},\frac{6}{8})\cup[\frac{7}{8},1]}),
\end{aligned}
$$

and the boundary conditions $u(0) = 0$ and $10^{-2}u'(1) - \mathrm{i}500u(1) = 0$. The exact solution u_e has an analytic expression which is given on each interval $(2^{-3}(j-1), 2^{-3}j)$ for $j = 1, \ldots, 8$

by

$$u_e(x) = A_j \exp\left(\mathsf{i}\frac{\kappa(x)}{\sqrt{a(x)}}x\right) + B_j \exp\left(-\mathsf{i}\frac{\kappa(x)}{\sqrt{a(x)}}x\right) + \frac{\exp\left(\mathsf{i}\frac{\kappa(x)}{\sqrt{a(x)}}x\right)}{2\mathsf{i}\kappa(x)\sqrt{a(x)}} \int_{2^{-3}(j-1)}^{x} f(t)\exp\left(-\mathsf{i}\frac{\kappa(t)}{\sqrt{a(t)}}t\right)dt$$
$$- \frac{\exp\left(-\mathsf{i}\frac{\kappa(x)}{\sqrt{a(x)}}x\right)}{2\mathsf{i}\kappa(x)\sqrt{a(x)}} \int_{2^{-3}(j-1)}^{x} f(t)\exp\left(\mathsf{i}\frac{\kappa(t)}{\sqrt{a(t)}}t\right)dt, \qquad x \in (2^{-3}(j-1), 2^{-3}j),$$

where all the coefficients A_j, B_j for $j = 1, \ldots, 8$ are uniquely determined by solving a system of linear equations that arises from imposing the boundary conditions and the following transmission conditions

$$u_e(2^{-3}j-) = u_e(2^{-3}j+), \quad a(2^{-3}j-)u_e'(2^{-3}j-) = a(2^{-3}j+)u_e'(2^{-3}j+), \qquad j = 1, \ldots, 8.$$

See Table 2.2 for the numerical performance measured by $\frac{\|u_N - u_e\|_\infty}{\|u_e\|_\infty}$ and $\frac{\|u_N' - u_e'\|_\infty}{\|u_e'\|_\infty}$, and Fig. 2.2 for the convergence plot and approximated solution u_N. As can be seen from Table 2.2, the convergence rates agree with the theoretical discussion in Sections 2.2 and 2.3.

		DAT using the compact FDM with order $M = 6$				DAT using the compact FDM with order $M = 8$			
N	ℓ	$\frac{\|u_N - u_e\|_\infty}{\|u_e\|_\infty}$	$\frac{\|u_N' - u_e'\|_\infty}{\|u_e'\|_\infty}$	Local CN	Link CN	$\frac{\|u_N - u_e\|_\infty}{\|u_e\|_\infty}$	$\frac{\|u_N' - u_e'\|_\infty}{\|u_e'\|_\infty}$	Local CN	Link CN
2^{15}	0	2.0833×10^{-1}	1.2491	1.62×10^{6}	–	8.0406×10^{-3}	4.8823×10^{-2}	1.59×10^{7}	–
	7	2.0833×10^{-1}	1.2491	2.27×10^{4}	2.32×10^{3}	8.0406×10^{-3}	4.8823×10^{-2}	2.22×10^{4}	2.32×10^{3}
	10	2.0833×10^{-1}	1.2491	3.98×10^{2}	2.32×10^{3}	8.0406×10^{-3}	4.8823×10^{-2}	3.87×10^{2}	2.32×10^{3}
2^{16}	0	3.5328×10^{-3}	2.1422×10^{-2}	7.55×10^{6}	–	2.3512×10^{-5}	1.4404×10^{-4}	7.56×10^{6}	–
	8	3.5328×10^{-3}	2.1422×10^{-2}	1.66×10^{3}	2.32×10^{3}	2.3512×10^{-5}	1.4404×10^{-4}	1.66×10^{3}	2.32×10^{3}
	11	3.5328×10^{-3}	2.1422×10^{-2}	1.32×10^{2}	2.32×10^{3}	2.3512×10^{-5}	1.4404×10^{-4}	1.32×10^{2}	2.32×10^{3}
2^{17}	0	5.1547×10^{-5}	3.1264×10^{-4}	2.87×10^{7}	–	8.5834×10^{-8}	5.2706×10^{-7}	2.87×10^{7}	–
	9	5.1547×10^{-5}	3.1264×10^{-4}	1.18×10^{4}	2.32×10^{3}	8.5834×10^{-8}	5.2705×10^{-7}	1.18×10^{4}	2.32×10^{3}
	12	5.1547×10^{-5}	3.1264×10^{-4}	3.65×10^{1}	2.32×10^{3}	8.5834×10^{-8}	5.2706×10^{-7}	3.65×10^{1}	2.36×10^{3}
2^{18}	0	7.9194×10^{-7}	4.8033×10^{-6}	1.14×10^{8}	–	3.2902×10^{-10}	2.0239×10^{-9}	1.14×10^{8}	–
	10	7.9194×10^{-7}	4.8033×10^{-6}	1.09×10^{4}	2.32×10^{3}	9.3775×10^{-10}	2.0278×10^{-9}	1.09×10^{4}	2.32×10^{3}
	13	7.9194×10^{-7}	4.8033×10^{-6}	4.28×10^{1}	2.32×10^{3}	3.2825×10^{-10}	2.0194×10^{-9}	4.28×10^{1}	2.32×10^{3}

Table 2.2: Relative errors for Example 2.2 using DAT with $N_0 = 32$ and $s = 1$ in Algorithm 2.1. The grid increment used in each sub-interval $[(k-1)2^{-3}, k2^{-3}]$ with $1 \leqslant k \leqslant 2^3$ is N^{-1}.

Example 2.3. Consider $[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x), x \in (0,1)$ with the following coefficients

$$a = e^{-x}\chi_{[0,\frac{31}{100})\cup[\frac{69}{100},\frac{81}{100})} + (e^x+1)\chi_{[\frac{31}{100},\frac{69}{100})\cup[\frac{81}{100},1]},$$
$$\kappa = 10^4 e^{2x}\chi_{[0,\frac{31}{100})} + 10^5 x^4 \chi_{[\frac{31}{100},\frac{69}{100})} + 10^4(1+x^4)\chi_{[\frac{69}{100},\frac{81}{100})} + 10^5 e^{-3x}\chi_{[\frac{81}{100},1]},$$

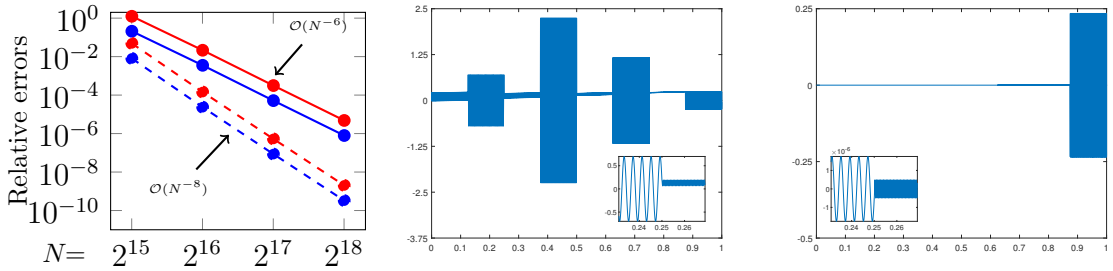


Figure 2.2: Example 2.2: Convergence plot (left) of DAT using the compact FDM with order $M = 6$ (solid) and $M = 8$ (dashed) for errors $\frac{\|u_N - u_e\|_\infty}{\|u_e\|_\infty}$ (blue) and $\frac{\|u'_N - u'_e\|_\infty}{\|u'_e\|_\infty}$ (red). The displayed convergence rates are obtained by calculating $\log_2\left(\frac{\|u_N - u_e\|_\infty}{\|u_{2N} - u_e\|_\infty}\right)$ and $\log_2\left(\frac{\|u'_N - u'_e\|_\infty}{\|u'_{2N} - u'_e\|_\infty}\right)$. The real (middle) and imaginary (right) parts of u_N with $N = 2^{18}$, $\ell = 13$ and $M = 8$.

$$f = 10^7\left(\chi_{[0,\frac{31}{100})} + x^2\chi_{[\frac{31}{100},\frac{69}{100})} + x^3\chi_{[\frac{69}{100},\frac{81}{100})} + x^5\chi_{[\frac{81}{100},1]}\right),$$

and the boundary conditions $u(0) = 1$ and $(e+1)^{1/2}u'(1) - \mathrm{i}10^5 e^{-3}u(1) = 0$. The exact solution's analytic expression is unknown. See Table 2.3 for the numerical performance measured by $\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$ and $\frac{\|u'_N - u'_{2N}\|_\infty}{\|u'_{2N}\|_\infty}$, and Fig. 2.3 for the convergence plot and approximated solution u_N. As can be seen from Table 2.3, the convergence rates agree with the theoretical discussion in Sections 2.2 and 2.3. This example shows how DAT is stable with respect to splits and tree levels. For simplicity, we consider a tree level that is not high. Hence, it is to be expected that the maximum condition numbers of the local and linking problems are still relatively large, but are nonetheless smaller than the condition numbers of FDM. In fact, if we look at these condition numbers in granular detail, a large proportion of them are significantly smaller than those of FDM for any given N. We also note that the maximum condition numbers listed in the column "Local CN" are the same for $(\ell, s) = (5, 1)$ and $(\ell, s) = (3, 2)$. The reason is because these two rows share the same local problems as defined in (2.19). The only difference lies in the size of the linking problems: 3×3 for $(\ell, s) = (5, 1)$ and 7×7 for $(\ell, s) = (3, 2)$.

		DAT using the compact FDM with order $M = 6$				DAT using the compact FD with order $M = 8$			
N	(ℓ, s)	$\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$	$\frac{\|u'_N - u'_{2N}\|_\infty}{\|u'_{2N}\|_\infty}$	Local CN	Link CN	$\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$	$\frac{\|u'_N - u'_{2N}\|_\infty}{\|u'_{2N}\|_\infty}$	Local CN	Link CN
2^{14}	$(0,0)$	5.9033×10^{-1}	8.6408×10^{-1}	2.63×10^{9}	–	9.4394×10^{-2}	1.2948×10^{-1}	4.59×10^{9}	–
	$(5,1)$	5.9033×10^{-1}	8.6408×10^{-1}	6.29×10^{4}	7.40×10^{4}	9.4394×10^{-2}	1.2948×10^{-1}	2.68×10^{5}	8.17×10^{4}
	$(3,2)$	5.9033×10^{-1}	8.6408×10^{-1}	6.29×10^{4}	1.95×10^{4}	9.4394×10^{-2}	1.2948×10^{-1}	2.68×10^{5}	2.94×10^{4}
2^{15}	$(0,0)$	4.7473×10^{-2}	6.7611×10^{-2}	5.19×10^{6}	–	2.4465×10^{-4}	3.3087×10^{-4}	5.56×10^{6}	–
	$(5,1)$	4.7473×10^{-2}	6.7611×10^{-2}	2.07×10^{5}	8.41×10^{3}	2.4465×10^{-4}	3.3087×10^{-4}	2.07×10^{5}	8.41×10^{3}
	$(3,2)$	4.7473×10^{-2}	6.7611×10^{-2}	2.07×10^{5}	2.90×10^{4}	2.4465×10^{-4}	3.3087×10^{-4}	2.07×10^{5}	2.92×10^{4}
2^{16}	$(0,0)$	7.2618×10^{-4}	1.0353×10^{-3}	2.22×10^{7}	–	8.7748×10^{-7}	1.1834×10^{-6}	2.22×10^{7}	–
	$(5,1)$	7.2618×10^{-4}	1.0353×10^{-3}	8.29×10^{5}	8.41×10^{3}	8.8029×10^{-7}	1.1867×10^{-6}	8.29×10^{5}	8.41×10^{3}
	$(3,2)$	7.2618×10^{-4}	1.0353×10^{-3}	8.29×10^{5}	2.92×10^{4}	8.7595×10^{-7}	1.1813×10^{-6}	8.29×10^{5}	2.92×10^{4}
2^{17}	$(0,0)$	1.1182×10^{-5}	1.5959×10^{-5}	8.89×10^{7}	–	3.7755×10^{-9}	5.0250×10^{-9}	8.89×10^{7}	–
	$(5,1)$	1.1188×10^{-5}	1.5967×10^{-5}	3.32×10^{6}	8.41×10^{3}	3.2508×10^{-9}	6.3200×10^{-9}	3.32×10^{6}	8.41×10^{3}
	$(3,2)$	1.1179×10^{-5}	1.5958×10^{-5}	3.32×10^{6}	2.92×10^{4}	6.9405×10^{-9}	9.3269×10^{-9}	3.32×10^{6}	2.92×10^{4}

Table 2.3: Relative errors for Example 2.3 using DAT with $N_0 = 16$ and $s = 1, 2$ in Algorithm 2.1. The grid increments used in $[0, \frac{31}{100}]$, $[\frac{31}{100}, \frac{69}{100}]$, $[\frac{69}{100}, \frac{81}{100}]$, and $[\frac{81}{100}, 1]$ are respectively $\frac{31}{25N}$, $\frac{38}{25N}$, $\frac{12}{25N}$, and $\frac{19}{25N}$.

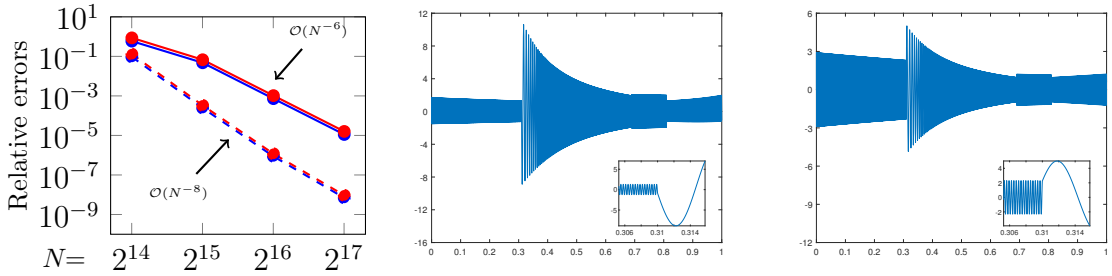


Figure 2.3: Example 2.3: Convergence plot (left) of DAT using the compact FDM with order $M = 6$ (solid) and $M = 8$ (dashed) for errors $\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$ (blue) and $\frac{\|u'_N - u'_{2N}\|_\infty}{\|u'_{2N}\|_\infty}$ (red). The displayed convergence rates are obtained by calculating $\log_2\left(\frac{\|u_N - u_{2N}\|_\infty / \|u_{2N}\|_\infty}{\|u_{2N} - u_{4N}\|_\infty / \|u_{4N}\|_\infty}\right)$ and $\log_2\left(\frac{\|u'_N - u'_{2N}\|_\infty / \|u'_{2N}\|_\infty}{\|u'_{2N} - u'_{4N}\|_\infty / \|u'_{4N}\|_\infty}\right)$. The real (middle) and imaginary (right) parts of u_N with $N = 2^{17}$, $(\ell, s) = (5, 1)$ and $M = 8$.

Example 2.4. Consider $[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x), x \in (0, 1)$ with the following coefficients

$$a = 1.1 + \sin(40\pi x), \quad \kappa = 10^5\left(1 - (x - 0.5)^2\right), \quad f = 10^9(x^7 + 1),$$

and the boundary conditions $\sqrt{1.1}u'(0) + \mathrm{i}75000u(0) = -1$ and $\sqrt{1.1}u'(1) - \mathrm{i}75000u(1) = 0$. The exact solution's analytic expression is unknown. See Table 2.4 for the numerical performance measured by $\frac{\|u_N - u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$ and $\frac{\|u'_N - u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$, and Fig. 2.4 for the convergence plot and approximated solution u_N. As can be seen from Table 2.4, the convergence rates agree with the theoretical discussion in Sections 2.2 and 2.3. As studied in [61], having a and κ that are oscillating and/or possess a large variation leads to an ill-conditioned coefficient

matrix. This example explores DAT's potential in handling the Helmholtz problem with an oscillatory coefficient a and a large wavenumber κ.

		DAT using the compact FDM with order $M = 6$				DAT using the compact FDM with order $M = 8$			
N	ℓ	$\frac{\|u_N - u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$	$\frac{\|u'_N - u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$	Local CN	Link CN	$\frac{\|u_N - u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$	$\frac{\|u'_N - u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$	Local CN	Link CN
$2^{18}+1$	0	4.8707×10^{-1}	7.6812×10^{-1}	5.53×10^{6}	–	6.7606×10^{-3}	1.0602×10^{-2}	5.66×10^{6}	–
	16	4.8707×10^{-1}	7.6812×10^{-1}	2.74×10^{5}	7.71×10^{6}	6.7606×10^{-3}	1.0602×10^{-2}	5.25×10^{5}	1.20×10^{7}
$2^{19}+1$	0	7.1208×10^{-3}	1.1068×10^{-2}	1.47×10^{7}	–	2.2709×10^{-5}	3.5577×10^{-5}	1.42×10^{7}	–
	17	7.1208×10^{-3}	1.1068×10^{-2}	4.39×10^{1}	1.16×10^{7}	2.2709×10^{-5}	3.5577×10^{-5}	4.39×10^{1}	1.19×10^{7}
$2^{20}+1$	0	1.0754×10^{-4}	1.6712×10^{-4}	6.21×10^{7}	–	8.6824×10^{-8}	1.3591×10^{-7}	5.98×10^{7}	–
	18	1.0754×10^{-4}	1.6712×10^{-4}	4.47×10^{1}	1.19×10^{7}	8.6814×10^{-8}	1.3589×10^{-7}	4.47×10^{1}	1.19×10^{7}
$2^{21}+1$	0	1.6652×10^{-6}	2.5876×10^{-6}	2.23×10^{8}	–	1.9291×10^{-9}	2.9177×10^{-9}	2.23×10^{8}	–
	19	1.6652×10^{-6}	2.5877×10^{-6}	4.49×10^{1}	1.19×10^{7}	1.8549×10^{-9}	2.8085×10^{-9}	4.49×10^{1}	1.19×10^{7}

Table 2.4: Relative errors for Example 2.4 using DAT with $N_0 = 4$ and $s = 1$ in Algorithm 2.1. The grid increment used in $[0, 1]$ is $(N-1)^{-1}$.

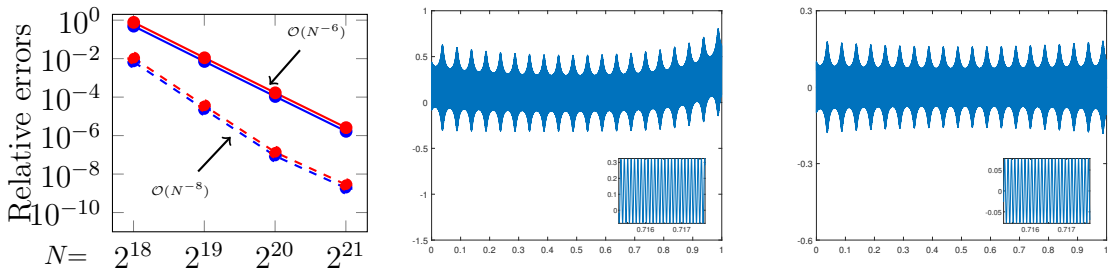


Figure 2.4: Example 2.4: Convergence plot (left) of DAT using the compact FDM with order $M = 6$ (solid) and $M = 8$ (dashed) for errors $\frac{\|u_N - u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$ (blue) and $\frac{\|u'_N - u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$ (red). The displayed convergence rates are obtained by calculating $\log_2\left(\frac{\|u_N - u_{2N-1}\|_\infty / \|u_{2N-1}\|_\infty}{\|u_{2N-1} - u_{4N-3}\|_\infty / \|u_{4N-3}\|_\infty}\right)$ and $\log_2\left(\frac{\|u'_N - u'_{2N-1}\|_\infty / \|u'_{2N-1}\|_\infty}{\|u'_{2N-1} - u'_{4N-3}\|_\infty / \|u'_{4N-3}\|_\infty}\right)$. The real (middle) and imaginary (right) parts of u_N with $N = 2^{21} + 1$, $\ell = 19$ and $M = 8$.

Example 2.5. Consider $[a(x)u'(x)]' + \kappa^2(x)u(x) = f(x), x \in (0, 1)$ with the following coefficients

$$
\begin{aligned}
a &= (5 + \sin(10\pi x))\chi_{[0,\frac{23}{100})\cup[\frac{83}{100},1]} + (2 + \sin(10\pi x))\chi_{[\frac{23}{100},\frac{53}{100})} + (9 + \sin(10\pi x))\chi_{[\frac{53}{100},\frac{83}{100})}, \\
\kappa &= 2000\left(e^{-x}\chi_{[0,\frac{23}{100})} + \chi_{[\frac{23}{100},\frac{53}{100})\cup[\frac{83}{100},1]} + 0.5e^{x}\chi_{[\frac{53}{100},\frac{83}{100})}\right), \\
f &= 2^{21}\left(\cosh(x)\chi_{[0,\frac{23}{100})} + \sinh(x)\chi_{[\frac{23}{100},\frac{53}{100})} - \cosh(x)\chi_{[\frac{53}{100},\frac{83}{100})} - \sinh(x)\chi_{[\frac{83}{100},1]}\right),
\end{aligned}
$$

and the boundary conditions $u'(0) = 1$ and $\sqrt{5}u'(1) - 2000iu(1) = 0$. The exact solution's analytic expression is unknown. See Table 2.5 for the numerical performance measured

by $\frac{\|u_N-u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$ and $\frac{\|u'_N-u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$, and Fig. 2.5 for the convergence plot and approximated solution u_N. As can be seen from Table 2.5, the convergence rates agree with the theoretical discussion in Sections 2.2 and 2.3.

		DAT using the compact FDM with order $M=6$				DAT using the compact FDM with order $M=8$			
N	ℓ	$\frac{\|u_N-u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$	$\frac{\|u'_N-u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$	Local CN	Link CN	$\frac{\|u_N-u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$	$\frac{\|u'_N-u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$	Local CN	Link CN
$2^{10}+1$	0	2.3932	1.8969	1.25×10^{6}	–	4.8820×10^{-2}	1.1942×10^{-1}	4.51×10^{5}	–
	5	2.3932	1.8969	2.66×10^{4}	3.44×10^{3}	4.8820×10^{-2}	1.1942×10^{-1}	6.47×10^{3}	2.01×10^{3}
$2^{11}+1$	0	6.9794×10^{-3}	6.4621×10^{-3}	3.13×10^{4}	–	1.6095×10^{-4}	5.3061×10^{-4}	3.11×10^{4}	–
	5	6.9794×10^{-3}	6.4621×10^{-3}	9.00×10^{3}	6.92×10^{3}	1.6095×10^{-4}	5.3061×10^{-4}	9.00×10^{3}	2.48×10^{3}
$2^{12}+1$	0	1.0216×10^{-4}	9.0880×10^{-5}	9.04×10^{4}	–	6.6055×10^{-7}	2.1900×10^{-6}	9.04×10^{4}	–
	5	1.0216×10^{-4}	9.0880×10^{-5}	3.66×10^{4}	2.53×10^{3}	6.6055×10^{-7}	2.1900×10^{-6}	3.66×10^{4}	2.51×10^{3}
$2^{13}+1$	0	1.5468×10^{-6}	1.4036×10^{-6}	3.60×10^{5}	–	2.6471×10^{-9}	8.6758×10^{-9}	3.60×10^{5}	–
	5	1.5468×10^{-6}	1.4036×10^{-6}	1.47×10^{5}	2.51×10^{3}	2.6448×10^{-9}	8.6783×10^{-9}	1.47×10^{5}	2.51×10^{3}

Table 2.5: Relative errors for Example 2.5 using DAT with $N_0=16$ and $s=1$ in Algorithm 2.1. The grid increments used in $[0,\frac{23}{100}]$, $[\frac{23}{100},\frac{53}{100}]$, $[\frac{53}{100},\frac{83}{100}]$, and $[\frac{83}{100},1]$ are respectively $\frac{23}{25(N-1)}$, $\frac{6}{5(N-1)}$, $\frac{6}{5(N-1)}$, and $\frac{17}{25(N-1)}$.

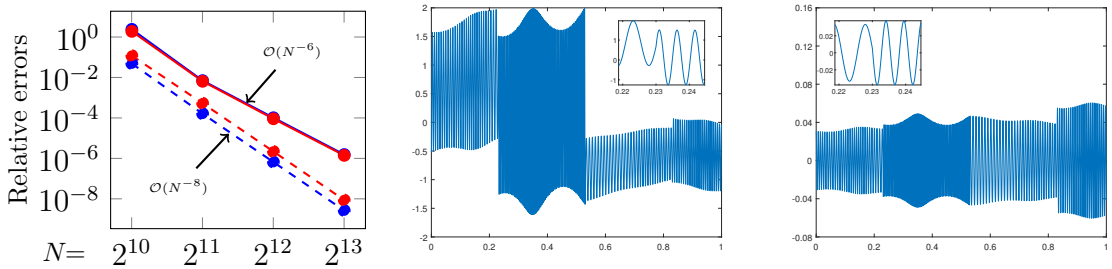


Figure 2.5: Example 2.5: Convergence plot (left) of DAT using the compact FDM with order $M=6$ (solid) and $M=8$ (dashed) for errors $\frac{\|u_N-u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$ (blue) and $\frac{\|u'_N-u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$ (red). The displayed convergence rates are obtained by calculating $\log_2\left(\frac{\|u_N-u_{2N-1}\|_\infty/\|u_{2N-1}\|_\infty}{\|u_{2N-1}-u_{4N-3}\|_\infty/\|u_{4N-3}\|_\infty}\right)$ and $\log_2\left(\frac{\|u'_N-u'_{2N-1}\|_\infty/\|u'_{2N-1}\|_\infty}{\|u'_{2N-1}-u'_{4N-3}\|_\infty/\|u'_{4N-3}\|_\infty}\right)$. The real (middle) and imaginary (right) parts of u_N with $N=2^{13}+1$, $\ell=5$ and $M=8$.

2.4.3 Numerical experiments on 2D Helmholtz equations

Separable 2D Helmholtz equations can be converted into a sequence of 1D Helmholtz problems, to which we may apply DAT as demonstrated below.

Example 2.6. Let $D_1:=\{(r,\theta):1\leqslant r<2,\theta\in[0,2\pi)\}$, $D_2:=\{(r,\theta):2\leqslant r\leqslant 4,\theta\in[0,2\pi)\}$, and $D:=D_1\cup D_2$. Consider the following 2D Helmholtz equation $\nabla\cdot(\nabla u)+\kappa^2u=0$ on the domain D, which can be rewritten in the polar coordinate system as follows:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2}+\kappa^2u=0\quad\text{on}\quad D,$$

$$\frac{\partial u}{\partial r}|_{r=1} = 0, \quad \left(\frac{\partial u}{\partial r} + \left(\frac{1}{2r} - 100\mathrm{i}\right) u\right)|_{r=4} = \left(\frac{\partial u_I}{\partial r} + \left(\frac{1}{2r} - 100\mathrm{i}\right) u_I\right)|_{r=4},$$

where $\kappa = 50\chi_{D_1} + 100\chi_{D_2} = \kappa_0(r)$ with $\kappa_0 := 50\chi_{[1,2)} + 100\chi_{[2,4]}$, $u_I := \sum_{m=0}^{\infty} \mathrm{i}^m(\delta_{0,m} + 2(1-\delta_{0,m}))J_m(100r)\cos(m\theta)$, and $J_m(\cdot)$ is the Bessel function of the first kind of order m. Using the method outlined in [97, Section 7.1], the exact solution u_e is given by the series $u_e = \sum_{m=0}^{\infty} r^{-1/2} v_m(r)\cos(m\theta)$, where $v_m, m \in \mathbb{N}_0$ satisfy the following 1D Helmholtz equations:

$$\begin{aligned} &v_m'' + \left(\kappa_0^2 - r^{-2}\left(m^2 - \tfrac{1}{4}\right)\right) v_m = 0, \quad r \in (1,4), \quad \text{with} \quad \left(v_m' - \tfrac{1}{2} v_m\right)|_{r=1} = 0, \\ &\left(v_m' - 100\mathrm{i} v_m\right)|_{r=4} = 2\mathrm{i}^m(\delta_{0,m} + 2(1-\delta_{0,m}))\left((J_m(100r))'|_{r=4} + \left(\tfrac{1}{8} - 100\mathrm{i}\right) J_m(400)\right). \end{aligned} \tag{2.65}$$

In particular, for each $m \in \mathbb{N}_0$, v_m has the following analytic expression

$$v_m = r^{1/2}\Big(A_m J_m(50r) + B_m Y_m(50r)\Big)\chi_{[1,2)} + r^{1/2}\Big(C_m J_m(100r) + D_m Y_m(100r)\Big)\chi_{[2,4]},$$

where $Y_m(\cdot)$ is the Bessel function of the second kind of order m and all the coefficients A_m, B_m, C_m, D_m are uniquely determined by solving a system of linear equations that arises from imposing the boundary conditions and the transmission conditions: $v_m(2-) = v_m(2+)$ and $v_m'(2-) = v_m'(2+)$.

Our approximated solution then takes the form $u_N = \sum_{m=0}^{640} r^{-1/2} v_{m,N}\cos(m\theta)$, where $v_{m,N}$ is the approximated solution to v_m in (2.65) using N points. In all cases, we use 2049 points to discretize the angle θ in our exact and approximated solutions. Also note that the following "Local CN" and "Link CN" record the maximum condition number of all local problems and all $m = 0, \ldots, 640$. See Table 2.6 for the numerical performance measured by both $\frac{\|u_N - u_e\|_\infty}{\|u_e\|_\infty}$ and $\frac{\|u_N - u_e\|_2}{\|u_e\|_2}$, where we use the first 641 terms of u_e (i.e., $u_e \approx \sum_{m=0}^{640} r^{-1/2} v_m(r)\cos(m\theta)$), and Fig. 2.6 for the convergence plot and approximated solution u_N. Due to the separation of variables, the convergence rates observed in the plot are solely driven by the convergence rates that take place in each 1D problem. As can be seen, the convergence rates agree with the theoretical discussion in Sections 2.2 and 2.3.

		DAT using the compact FDM with order $M = 6$				DAT using the compact FD with order $M = 8$			
N	ℓ	$\frac{\|u_N-u_e\|_\infty}{\|u_e\|_\infty}$	$\frac{\|u_N-u_e\|_2}{\|u_e\|_2}$	Local CN	Link CN	$\frac{\|u_N-u_e\|_\infty}{\|u_e\|_\infty}$	$\frac{\|u_N-u_e\|_2}{\|u_e\|_2}$	Local CN	Link CN
2^8+1	0	1.0461×10^{-1}	6.6021×10^{-2}	1.84×10^{5}	–	6.8220×10^{-3}	3.4958×10^{-3}	1.79×10^{5}	–
	5	1.0461×10^{-1}	6.6021×10^{-2}	4.79×10^{4}	3.05×10^{5}	6.8220×10^{-3}	3.4958×10^{-3}	9.13×10^{4}	2.81×10^{5}
2^9+1	0	1.2885×10^{-3}	7.6950×10^{-4}	4.04×10^{4}	–	2.7208×10^{-5}	1.4102×10^{-5}	4.04×10^{4}	–
	6	1.2885×10^{-3}	7.6950×10^{-4}	7.06×10^{2}	2.78×10^{5}	2.7208×10^{-5}	1.4102×10^{-5}	7.05×10^{2}	2.78×10^{5}
$2^{10}+1$	0	1.9204×10^{-5}	1.1290×10^{-5}	1.52×10^{5}	–	1.0825×10^{-7}	5.5999×10^{-8}	1.52×10^{5}	–
	7	1.9204×10^{-5}	1.1290×10^{-5}	4.78×10^{1}	2.78×10^{5}	1.0825×10^{-7}	5.5999×10^{-8}	4.78×10^{1}	2.78×10^{5}
$2^{11}+1$	0	2.9671×10^{-7}	1.7375×10^{-7}	6.01×10^{5}	–	4.2441×10^{-10}	2.1932×10^{-10}	6.01×10^{5}	–
	8	2.9671×10^{-7}	1.7375×10^{-7}	4.56×10^{1}	2.78×10^{5}	4.2435×10^{-10}	2.1934×10^{-10}	4.56×10^{1}	2.78×10^{5}

Table 2.6: Relative errors for Example 2.6 using DAT with $N_0 = 8$ and $s = 1$ in Algorithm 2.1. The grid increments used in each $[1, 2]$ and $[2, 4]$ are respectively $2(N-1)^{-1}$ and $4(N-1)^{-1}$.

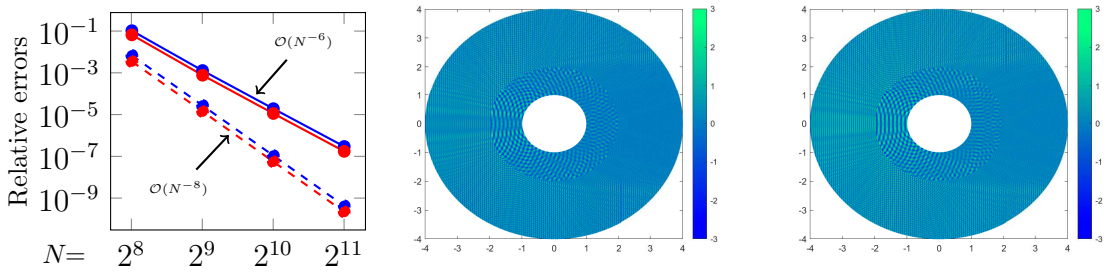


Figure 2.6: Example 2.6: Convergence plot (left) of DAT using the compact FDM with order $M = 6$ (solid) and $M = 8$ (dashed) for relative errors $\frac{\|u_N - u_e\|_\infty}{\|u_e\|_\infty}$ (red) and $\frac{\|u_N - u_e\|_2}{\|u_e\|_2}$ (blue). The displayed convergence rates are obtained by calculating $\log_2\left(\frac{\|u_N - u_e\|_\infty}{\|u_{2N} - u_e\|_\infty}\right)$ and $\log_2\left(\frac{\|u_N - u_e\|_2}{\|u_{2N} - u_e\|_2}\right)$. The real (middle) and imaginary (right) parts of u_N with $N = 2^{11} + 1$, $\ell = 8$ and $M = 8$.

Example 2.7. Let $D_1 := \{(r, \theta) : 1 \leqslant r < 3, \theta \in [0, 2\pi)\}$, $D_2 := \{(r, \theta) : 3 \leqslant r \leqslant 4, \theta \in [0, 2\pi)\}$, and $D := D_1 \cup D_2$. Consider the following 2D Helmholtz equation $\nabla \cdot (\nabla u) + \kappa^2 u = f$ on the domain D, which can be rewritten in the polar coordinate system as follows:

$$\tfrac{1}{r}\tfrac{\partial}{\partial r}\left(r\tfrac{\partial u}{\partial r}\right) + \tfrac{1}{r^2}\tfrac{\partial^2 u}{\partial \theta^2} + \kappa^2 u = f \quad \text{on} \quad D \quad \text{with} \quad u|_{r=1} = \tfrac{\sin(4\theta)}{\sqrt{\pi}} \quad \text{and} \quad \left(\tfrac{\partial u}{\partial r} - 50\mathrm{i}u\right)|_{r=4} = 0,$$

where $\kappa = 400\chi_{D_1} + 50\chi_{D_2} = \kappa_0(r)$ with $\kappa_0(r) := 400\chi_{[1,3)} + 50\chi_{[3,4]}$ and $f = (10^5 J_1(r)\chi_{D_1} + 10^4 J_0(r)\chi_{D_2})$. By applying the separation of variables twice, the exact solution u_e in the polar coordinate system to the above 2D Helmholtz equation is given by the series $u_e = \sum_{m\in\mathbb{Z}} (2\pi r)^{-1/2} v_m(r) e^{\mathrm{i}m\theta}$, where for each $m \in \mathbb{Z}$, v_m satisfies

$$\begin{aligned} &v_m'' + \left(\kappa_0^2 - r^{-2}\left(m^2 - \tfrac{1}{4}\right)\right) v_m = r^{1/2} f_m(r), \quad r \in (1, 4), \quad \text{with} \\ &v_m|_{r=1} = \tfrac{i}{\sqrt{2}}(-\delta_{4,m} + \delta_{-4,m}), \quad \left(v_m' - \left(\tfrac{1}{8} + 50\mathrm{i}\right) v_m\right)|_{r=4} = 0, \end{aligned} \tag{2.66}$$

and $f_m(r) := (2\pi)^{-1/2} \int_0^{2\pi} f(r, \theta) e^{im\theta} d\theta$ can be efficiently computed by FFT. Note that f_m are zero except for $m = 0$. Since v_m are zero for $m \in \mathbb{Z}\backslash\{0, \pm 4\}$, our approximated solution is of the form $u_N = (2\pi r)^{-1/2} v_{0,N} + (2\pi r)^{-1/2}\left(v_{4,N} e^{\mathrm{i}4\theta} + v_{-4,N} e^{-\mathrm{i}4\theta}\right)$, where $v_{m,N}$ is the approximated solution to v_m in (2.66) using N points. We use 2049 points to discretize the angle θ in our approximated solutions. Note that the following "Local CN" and "Link CN" record the maximum condition number of all local and linking problems, and all $m = 0, \pm 4$. See Table 2.7 for the numerical performance measured by both $\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$ and $\frac{\|u_N - u_{2N}\|_2}{\|u_{2N}\|_2}$, and Fig. 2.7 for the convergence plot and approximated solution u_N. Due to the separation

of variables, the convergence rates observed in the plot are solely driven by the convergence rates that take place in each 1D problem. As can be seen from Table 2.7, the convergence rates agree with the theoretical discussion in Sections 2.2 and 2.3.

		DAT using the compact FDM with order $M = 6$				DAT using the compact FD with order $M = 8$			
N	ℓ	$\frac{\|u_N-u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$	$\frac{\|u_N-u_{2N}\|_2}{\|u_{2N}\|_2}$	Local CN	Link CN	$\frac{\|u_N-u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$	$\frac{\|u_N-u_{2N}\|_2}{\|u_{2N}\|_2}$	Local CN	Link CN
2^{10}	0	2.8213×10^{-1}	2.0128×10^{-1}	1.54×10^{6}	–	1.4812×10^{-2}	1.1534×10^{-2}	1.47×10^{6}	–
	7	2.8213×10^{-1}	2.0128×10^{-1}	1.61×10^{2}	5.26×10^{3}	1.4812×10^{-2}	1.1534×10^{-2}	1.43×10^{2}	4.95×10^{3}
2^{11}	0	5.6999×10^{-3}	4.3636×10^{-3}	1.21×10^{5}	–	4.1235×10^{-5}	3.2294×10^{-5}	1.22×10^{5}	–
	8	5.6999×10^{-3}	4.3636×10^{-3}	7.06×10^{2}	4.98×10^{3}	4.1235×10^{-5}	3.2294×10^{-5}	7.05×10^{2}	4.98×10^{3}
2^{12}	0	8.3447×10^{-5}	6.3909×10^{-5}	4.85×10^{5}	–	1.5005×10^{-7}	1.1778×10^{-7}	4.86×10^{5}	–
	9	8.3447×10^{-5}	6.3909×10^{-5}	4.50×10^{1}	4.98×10^{3}	1.5005×10^{-7}	1.1778×10^{-7}	4.50×10^{1}	4.98×10^{3}
2^{13}	0	1.2805×10^{-6}	9.8055×10^{-7}	1.95×10^{6}	–	5.8547×10^{-10}	4.5911×10^{-10}	1.95×10^{6}	–
	10	1.2805×10^{-6}	9.8055×10^{-7}	4.50×10^{1}	4.98×10^{3}	5.8767×10^{-10}	4.6687×10^{-10}	4.50×10^{1}	4.98×10^{3}

Table 2.7: Relative errors for Example 2.7 using DAT with $N_0 = 8$ and $s = 1$ in Algorithm 2.1. The grid increments used in each $[1, 3]$ and $[3, 4]$ are respectively $4N^{-1}$ and $2N^{-1}$.

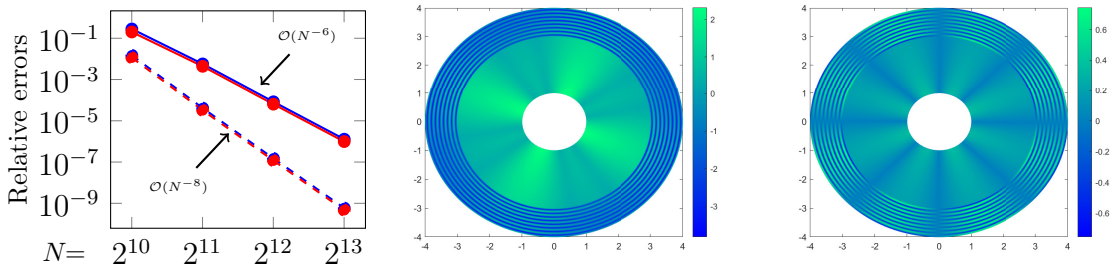


Figure 2.7: Example 2.7: Convergence plot (left) of DAT using the compact FDM with order $M = 6$ (solid) and $M = 8$ (dashed) for errors $\frac{\|u_N-u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$ (red) and $\frac{\|u_N-u_{2N}\|_2}{\|u_{2N}\|_2}$ (blue). The displayed convergence rates are obtained by calculating $\log_2\left(\frac{\|u_N-u_{2N}\|_\infty/\|u_{2N}\|_\infty}{\|u_{2N}-u_{4N}\|_\infty/\|u_{4N}\|_\infty}\right)$ and $\log_2\left(\frac{\|u_N-u_{2N}\|_2/\|u_{2N}\|_2}{\|u_{2N}-u_{4N}\|_2/\|u_{4N}\|_2}\right)$. The real (middle) and imaginary (right) parts of u_N with $N = 2^{13}$, $\ell = 10$ and $M = 8$.

Example 2.8. Consider the following 2D Helmholtz equation $\nabla \cdot (\nabla u) + \kappa^2 u = f$ on $\Omega = (0, 1)^2$, where

$$\begin{aligned}\kappa_0(x) &:= 2^{\frac{5}{2}}(900000x^5 - 2250000x^4 + 2130000x^3 - 945000x^2 + 198450x - 15445)^{\frac{1}{2}}\chi_{[\frac{3}{10},\frac{7}{10})} \\ &\quad + 2^7(\chi_{[0,\frac{3}{10})} + 2\chi_{[\frac{7}{10},1]}), \\ f(x,y) &:= 5000\sqrt{2}\left((6x+1)\cos(6.5\pi y)\chi_{[0,\frac{3}{10})\times[0,1]} + 2\cos(25.5\pi y)\chi_{[0,\frac{3}{10})\times[0,1]}\right. \\ &\quad \left.+(-6x+7)\cos(6.5\pi y)\chi_{[\frac{7}{10},1]\times[0,1]}\right),\end{aligned}$$

and $\kappa(x,y) := \kappa_0(x)$ with the following boundary conditions

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on} \quad (0,1)\times\{0\}, \qquad \frac{\partial u}{\partial \nu} - \mathrm{i}2^8 u = 0 \quad \text{on} \quad \{1\}\times(0,1), \qquad u = 0 \quad \text{on} \quad (0,1)\times\{1\},$$
$$\text{and} \qquad u = \sqrt{2}(\cos(14.5\pi y) + \cos(30.5\pi y)) \quad \text{on} \quad \{0\}\times(0,1).$$

By the separation of variables, the exact solution u_e to the above 2D Helmholtz equation is given by the series $u_e(x,y) = \sum_{m=0}^{\infty} \sqrt{2} v_m(x) \cos((m+1/2)\pi y)$, where for each $m \in \mathbb{N}_0$, v_m satisfies

$$v_m'' + (\kappa_0^2 - (m+\tfrac{1}{2})^2\pi^2) v_m = f_m(x), \quad x \in (0,1) \quad \text{with}$$
$$v_m(0) = \delta_{14,m} + \delta_{30,m}, \quad v_m'(1) - 2^8 \mathrm{i} v_m(1) = 0,$$

and $f_m(x) := \sqrt{2}\int_0^1 f(x,y)\cos((m+\frac{1}{2})\pi y)dy$ can be efficiently computed through FFT. Note that f_m are zero except $m = 6, 25$. Since v_m are zero for all $m \in \mathbb{N}_0\backslash\{6,14,25,30\}$, our approximated solution is of the form $u_N = \sqrt{2}(v_{6,N}\cos(6.5\pi y) + v_{14,N}\cos(14.5\pi y) + v_{25,N}\cos(25.5\pi y) + v_{30,N}\cos(30.5\pi y))$, where $v_{m,N}$ with $m = 6, 14, 25, 30$ are the approximated solutions to v_m in (2.66) using N points. We use 2049 points to discretize $\cos((m+1/2)\pi y)$ for $m = 6, 14, 25, 30$ in our approximated solutions. Also note that the following "Local CN" and "Link CN" record the maximum condition number of all local problems and all $m = 6, 14, 25, 30$. See Table 2.8 for the numerical performance measured by both $\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$ and $\frac{\|u_N - u_{2N}\|_2}{\|u_{2N}\|_2}$, and Fig. 2.8 for the convergence plot and approximated solution u_N. Due to the separation of variables, the convergence rates observed in the plot are solely driven by the convergence rates that take place in each 1D problem. As can be seen from Table 2.8, the convergence rates agree with the theoretical discussion in Sections 2.2 and 2.3.

		DAT using the compact FDM with order $M = 6$				DAT using the compact FD with order $M = 8$			
N	ℓ	$\frac{\|u_N-u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$	$\frac{\|u_N-u_{2N}\|_2}{\|u_{2N}\|_2}$	Local CN	Link CN	$\frac{\|u_N-u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$	$\frac{\|u_N-u_{2N}\|_2}{\|u_{2N}\|_2}$	Local CN	Link CN
$3(2^6)$	0	1.2184×10^{-2}	1.4965×10^{-2}	1.52×10^{5}	–	2.5802×10^{-3}	3.4368×10^{-3}	4.52×10^{5}	–
	2	1.2184×10^{-2}	1.4965×10^{-2}	1.54×10^{3}	5.18×10^{2}	2.5802×10^{-3}	3.4368×10^{-3}	3.22×10^{3}	5.19×10^{2}
$3(2^7)$	0	1.9809×10^{-4}	2.3348×10^{-4}	7.48×10^{3}	–	1.1417×10^{-5}	1.5130×10^{-5}	7.48×10^{3}	–
	3	1.9809×10^{-4}	2.3348×10^{-4}	2.60×10^{3}	5.19×10^{2}	1.1417×10^{-5}	1.5130×10^{-5}	2.60×10^{3}	5.19×10^{2}
$3(2^8)$	0	3.1521×10^{-6}	3.7059×10^{-6}	2.96×10^{4}	–	4.6092×10^{-8}	6.0943×10^{-8}	2.96×10^{4}	–
	4	3.1521×10^{-6}	3.7059×10^{-6}	3.85×10^{4}	8.49×10^{2}	4.6094×10^{-8}	6.0946×10^{-8}	3.85×10^{4}	8.49×10^{2}
$3(2^9)$	0	4.9446×10^{-8}	5.8172×10^{-8}	1.18×10^{5}	–	1.8328×10^{-10}	2.3989×10^{-10}	1.18×10^{5}	–
	5	4.9447×10^{-8}	5.8172×10^{-8}	1.54×10^{2}	8.87×10^{3}	1.8344×10^{-10}	2.4014×10^{-10}	1.54×10^{2}	8.87×10^{3}

Table 2.8: Relative errors for Example 2.8 using DAT with $N_0 = 12$ and $s = 1$ in Algorithm 2.1. The grid increments used in each $[0, \frac{3}{10}]$, $[\frac{3}{10}, \frac{7}{10}]$, and $[\frac{7}{10}, 1]$ are respectively $\frac{9}{10N}$, $\frac{6}{5N}$, and $\frac{9}{10N}$.

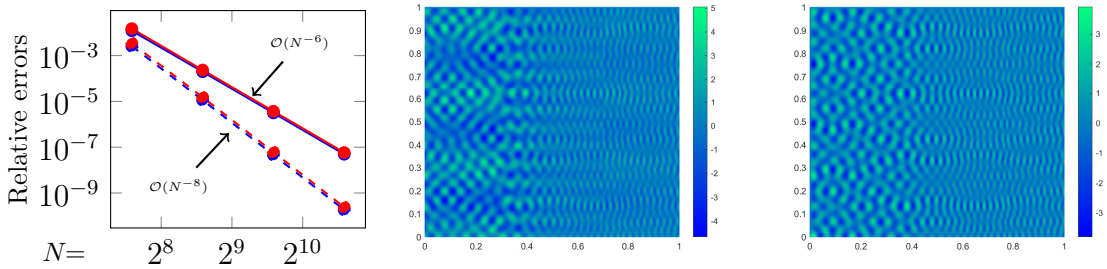


Figure 2.8: Example 2.8: Convergence plot (left) of DAT using the compact FDM with order $M = 6$ (solid) and $M = 8$ (dashed) for errors $\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$ (red) and $\frac{\|u_N - u_{2N}\|_2}{\|u_{2N}\|_2}$ (blue). The displayed convergence rates are obtained by calculating $\log_2\left(\frac{\|u_N - u_{2N}\|_\infty / \|u_{2N}\|_\infty}{\|u_{2N} - u_{4N}\|_\infty / \|u_{4N}\|_\infty}\right)$ and $\log_2\left(\frac{\|u_N - u_{2N}\|_2 / \|u_{2N}\|_2}{\|u_{2N} - u_{4N}\|_2 / \|u_{4N}\|_2}\right)$. The real (middle) and imaginary (right) parts of u_N with $N = 3(2^9)$, $\ell = 5$ and $M = 8$.

2.4.4 DAT and compact FDMs using only function values

The direct usage of derivatives of a, κ^2, f in Theorems 2.4 and 2.5 may not be computationally efficient. These derivatives can in fact be estimated by using only function values of a, κ^2, f through a local polynomial approximation. For example, if we consider an interior stencil of the form (2.44) and (2.45), we know from (2.3) that all stencil coefficients depend on $a(x_b), a'(x_b), \ldots, a^{(M-1)}(x_b)$, $\kappa^2(x_b), [\kappa^2]'(x_b), \ldots, [\kappa^2]^{(M-2)}(x_b)$, and $f(x_b)$, $f'(x_b)$, $\ldots$, $f^{(M-2)}(x_b)$. Consider $a(x_b), a'(x_b), \ldots, a^{(M-1)}(x_b)$. Let $J \geqslant M$ and take J points $\{x_j\}_{j=1}^J$ near the base point x_b such that all the points fall into one piece of the piecewise smooth functions a, κ^2 and f. Find the unique polynomial p of degree $J-1$ satisfying $p(x_j) = a(x_j)$ for all $j = 1, \ldots, J$. Then $a^{(n)}(x_b) \approx p^{(n)}(x_b)$ for $n = 0, \ldots, M-1$. We often use $J = M$ and $x_b \in \{x_j\}_{j=1}^J$ such that $\{x_j\}_{j=1}^J$ is evenly spaced with mesh size $h/2$. Using only function values, we re-calculate numerical experiments in Examples 2.1–2.8, which yield virtually same results as those using derivatives explicitly. It demonstrates the convenience of using a local polynomial approximation in lieu of true derivatives, which may have complicated expressions. For the sake of conciseness, we only provide re-calculated Examples 2.3 and 2.5.

		DAT using the compact FDM with order $M = 6$ for Example 2.3						DAT using the compact FDM with order $M = 6$ for Example 2.5			
N	(ℓ, s)	$\frac{\|u_N - u_{2N}\|_\infty}{\|u_{2N}\|_\infty}$	$\frac{\|u'_N - u'_{2N}\|_\infty}{\|u'_{2N}\|_\infty}$	Local CN	Link CN	N	ℓ	$\frac{\|u_N - u_{2N-1}\|_\infty}{\|u_{2N-1}\|_\infty}$	$\frac{\|u'_N - u'_{2N-1}\|_\infty}{\|u'_{2N-1}\|_\infty}$	Local CN	Link CN
2^{15}	$(0,0)$	4.7473×10^{-2}	6.7611×10^{-2}	5.19×10^{6}	–	$2^{10}+1$	0	2.3932	1.8969	1.25×10^{6}	–
	$(5,1)$	4.7473×10^{-2}	6.7611×10^{-2}	2.07×10^{5}	8.41×10^{3}		5	2.3932	3.3990	2.66×10^{4}	3.44×10^{3}
	$(3,2)$	4.7473×10^{-2}	6.7611×10^{-2}	2.07×10^{5}	2.90×10^{4}						
2^{16}	$(0,0)$	7.2618×10^{-4}	1.0353×10^{-3}	2.22×10^{7}	–	$2^{11}+1$	0	6.9794×10^{-3}	6.4621×10^{-3}	3.13×10^{4}	–
	$(5,1)$	7.2618×10^{-4}	1.0353×10^{-3}	8.29×10^{5}	8.41×10^{3}		5	6.9794×10^{-3}	6.4621×10^{-3}	9.00×10^{3}	6.92×10^{3}
	$(3,2)$	7.2618×10^{-4}	1.0353×10^{-3}	8.29×10^{5}	2.92×10^{4}						
2^{17}	$(0,0)$	1.1182×10^{-5}	1.5959×10^{-5}	8.89×10^{7}	–	$2^{12}+1$	0	1.0216×10^{-4}	9.0880×10^{-5}	9.03×10^{4}	–
	$(5,1)$	1.1188×10^{-5}	1.5967×10^{-5}	3.32×10^{6}	8.41×10^{3}		5	1.0216×10^{-4}	9.0880×10^{-5}	3.66×10^{4}	2.53×10^{3}
	$(3,2)$	1.1179×10^{-5}	1.5955×10^{-5}	3.32×10^{6}	2.92×10^{4}						
2^{18}	$(0,0)$	1.7274×10^{-7}	2.4663×10^{-7}	3.56×10^{8}	–	$2^{13}+1$	0	1.5468×10^{-6}	1.4036×10^{-6}	3.60×10^{5}	–
	$(5,1)$	1.6132×10^{-7}	2.3056×10^{-7}	1.33×10^{7}	8.41×10^{3}		5	1.5468×10^{-6}	1.4036×10^{-6}	1.47×10^{5}	2.51×10^{3}
	$(3,2)$	1.7914×10^{-7}	2.5548×10^{-7}	1.33×10^{7}	2.92×10^{4}						

Table 2.9: Relative errors for Examples 2.3 and 2.5 using only point values (without explicitly computing derivatives) in DAT with the compact FDM with order $M = 6$.

Chapter 3

Sixth Order Compact FDM for 2D Helmholtz Equations with Singular Sources and Reduced Pollution Effect

Now that we have presented an arbitrarily high order compact 1D FDM in Chapter 2, we are ready to shift our attention to a 2D FDM. In this chapter, we present a sixth order compact FDM for the 2D Helmholtz equation with reduced pollution effect.

We start by introducing the model problem. Let $\Omega = (l_1, l_2) \times (l_3, l_4)$ and ψ be a smooth two-dimensional function. Consider a smooth curve $\Gamma_I := \{(x, y) \in \Omega : \psi(x, y) = 0\}$, which partitions Ω into two subregions: $\Omega^+ := \{(x, y) \in \Omega \ : \ \psi(x, y) > 0\}$ and $\Omega^- := \{(x, y) \in \Omega \ : \ \psi(x, y) < 0\}$. The model problem is explicitly defined as follows:

$$\begin{cases} \Delta u + \kappa^2 u = f & \text{in } \Omega \setminus \Gamma_I, \\ [u] = g_D, \quad [\nabla u \cdot \boldsymbol{\nu}] = g_N & \text{on } \Gamma_I, \\ \mathcal{B}_1 u = g_1 \text{ on } \Gamma_1 := \{l_1\} \times (l_3, l_4), & \mathcal{B}_2 u = g_2 \text{ on } \Gamma_2 := \{l_2\} \times (l_3, l_4), \\ \mathcal{B}_3 u = g_3 \text{ on } \Gamma_3 := (l_1, l_2) \times \{l_3\}, & \mathcal{B}_4 u = g_4 \text{ on } \Gamma_4 := (l_1, l_2) \times \{l_4\}, \end{cases} \tag{3.1}$$

where κ is the wavenumber, f is the source term, and for any point $(x_0, y_0) \in \Gamma_I$,

$$[u](x_0, y_0) := \lim_{(x,y)\in\Omega^+, (x,y)\to(x_0,y_0)} u(x, y) - \lim_{(x,y)\in\Omega^-, (x,y)\to(x_0,y_0)} u(x, y), \tag{3.2}$$

$$[\nabla u \cdot \boldsymbol{\nu}](x_0, y_0) := \lim_{(x,y)\in\Omega^+, (x,y)\to(x_0,y_0)} \nabla u(x, y) \cdot \boldsymbol{\nu} - \lim_{(x,y)\in\Omega^-, (x,y)\to(x_0,y_0)} \nabla u(x, y) \cdot \boldsymbol{\nu}, \tag{3.3}$$

where $\boldsymbol{\nu}$ is the unit normal vector of Γ_I pointing towards Ω^+. In (3.1), the boundary operators $\mathcal{B}_1, \ldots, \mathcal{B}_4 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}, \frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa \mathbf{I}_d\}$, where $\mathbf{I}_d$ corresponds to the Dirichlet boundary

condition (sound soft boundary condition for the identical zero boundary datum), $\frac{\partial}{\partial \boldsymbol{\nu}}$ corresponds to the Neumann boundary condition (sound hard boundary condition for the identical zero boundary datum), and $\frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa \mathbf{I}_d$ (with i being the imaginary unit) corresponds to the impedance boundary condition. Moreover, the Helmholtz equation of (3.1) with $g_D = 0$ is equivalent to finding the weak solution $u \in H^1(\Omega)$ of $\Delta u + \kappa^2 u = f + g_N \delta_{\Gamma_I}$ in Ω, where δ_{Γ_I} is the Dirac distribution along the interface curve Γ_I.

In particular, we assume that

(A1) The solution u and the source term f have uniformly continuous partial derivatives of (total) orders up to seven and six respectively in each of the subregions Ω^+ and Ω^-. However, both u and f may be discontinuous across the interface Γ_I.

(A2) The interface curve Γ_I is smooth in the sense that for each $(x^*, y^*) \in \Gamma_I$, there exists a local parametric equation: $\gamma : (-\epsilon, \epsilon) \to \Gamma_I$ with $\epsilon > 0$ such that $\gamma(0) = (x^*, y^*)$ and $\|\gamma'(0)\|_2 \neq 0$.

(A3) The one-dimensional functions $g_D \circ \gamma$ and $g_N \circ \gamma$ have uniformly continuous derivatives of (total) orders up to eight and seven respectively on the interface Γ_I, where γ is given in (A2).

(A4) Each of the functions $g_1, \ldots, g_4$ has uniformly continuous derivatives of (total) order up to seven on the boundary Γ_j.

We explain how our proposed sixth order compact finite difference scheme with reduced pollution effect is developed in Section 3.1. We first construct the interior finite difference stencil with reduced pollution. Then, we construct the sixth order boundary and corner finite difference stencils with reduced pollution. Finally, we construct the compact interface finite difference stencil. Numerical experiments to demonstrate the superiority of our proposed method to several state-of-the-art FDMs are presented in Section 3.2. In Section 3.3, we present the proofs of several theorems stated in Section 3.1.

Results in this chapter are based on [51].

3.1 Stencils for sixth order compact finite difference schemes with reduced pollution effect using uniform cartesian grids

Let $\Omega = (l_1, l_2) \times (l_3, l_4)$. Without loss of generality, we assume $l_4 - l_3 = N_0(l_2 - l_1)$ for some $N_0 \in \mathbb{N}$. For any positive integer $N_1 \in \mathbb{N}$, we define $N_2 := N_0 N_1$ and so the grid size is

$h := (l_2 - l_1)/N_1 = (l_4 - l_3)/N_2$. Let

$$x_i = l_1 + ih, \quad i = 0, \ldots, N_1, \quad \text{and} \quad y_j = l_3 + jh, \quad j = 0, \ldots, N_2. \tag{3.4}$$

Our focus of this section is to develop sixth order compact finite difference schemes with reduced pollution effect on uniform Cartesian grids. Recall that a compact stencil centered at (x_i, y_j) contains nine points $(x_i + kh, y_j + lh)$ for $k, l \in \{-1, 0, 1\}$. Define

$$\begin{aligned} d_{i,j}^+ &:= \{(k, \ell) \ : \ k, \ell \in \{-1, 0, 1\}, \psi(x_i + kh, y_j + \ell h) \geqslant 0\}, \quad \text{and} \\ d_{i,j}^- &:= \{(k, \ell) \ : \ k, \ell \in \{-1, 0, 1\}, \psi(x_i + kh, y_j + \ell h) < 0\}. \end{aligned}$$

Thus, the interface curve $\Gamma_I := \{(x, y) \in \Omega \ : \ \psi(x, y) = 0\}$ splits the nine points in our compact stencil into two disjoint sets $\{(x_{i+k}, y_{j+\ell}) \ : \ (k, \ell) \in d_{i,j}^+\} \subseteq \Omega^+ \cup \Gamma_I$ and $\{(x_{i+k}, y_{j+\ell}) \ : \ (k, \ell) \in d_{i,j}^-\} \subseteq \Omega^-$. We refer to a grid/center point (x_i, y_j) as *a regular point* if $d_{i,j}^+ = \emptyset$ or $d_{i,j}^- = \emptyset$. The center point (x_i, y_j) of a stencil is *regular* if all its nine points are completely in $\Omega^+ \cup \Gamma_I$ (hence $d_{i,j}^- = \emptyset$) or in Ω^- (i.e., $d_{i,j}^+ = \emptyset$). Otherwise, the center point (x_i, y_j) of a stencil is referred to as *an irregular point* if both $d_{i,j}^+$ and $d_{i,j}^-$ are nonempty.

Now, let us pick and fix a base point (x_i^*, y_j^*) inside the open square $(x_i - h, x_i + h) \times (y_j - h, y_j + h)$, which can be written as

$$x_i^* = x_i - v_0 h \quad \text{and} \quad y_j^* = y_j - w_0 h \quad \text{with} \quad -1 < v_0, w_0 < 1. \tag{3.5}$$

We shall use the following notations:

$$u^{(m,n)} := \frac{\partial^{m+n} u}{\partial^m x \partial^n y}(x_i^*, y_j^*) \quad \text{and} \quad f^{(m,n)} := \frac{\partial^{m+n} f}{\partial^m x \partial^n y}(x_i^*, y_j^*), \tag{3.6}$$

which are used to represent their (m, n)th partial derivatives at the base point (x_i^*, y_j^*). Define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, the set of all nonnegative integers. Given $L \in \mathbb{N}_0$, we define

$$\Lambda_L := \{(m, n - m) \ : \ n = 0, \ldots, L \ \text{ and } \ m = 0, \ldots, n\}, \qquad L \in \mathbb{N}_0. \tag{3.7}$$

For a smooth function u and small x, y, the values $u(x + x_i^*, y + y_j^*)$ are well approximated by its Taylor polynomial as follows:

$$u(x + x_i^*, y + y_j^*) = \sum_{(m,n) \in \Lambda_{M+1}} \frac{u^{(m,n)}}{m! n!} x^m y^n + \mathscr{O}(h^{M+2}), \qquad x, y \in (-2h, 2h). \tag{3.8}$$

To put differently, in a neighborhood of the base point (x_i^*, y_j^*), the function u is well ap-

proximated and completely determined by the partial derivatives of u of total degree less than $M+2$ at the base point (x_i^*, y_j^*), i.e., by the unknown quantities $u^{(m,n)}, (m,n) \in \Lambda_{M+1}$. The same holds for $f(x+x_i^*, y+y_j^*)$. For $x \in \mathbb{R}$, the floor function $\lfloor x \rfloor$ is defined to be the largest integer less than or equal to x. For an integer m, we define

$$\operatorname{odd}(m) := \frac{1-(-1)^m}{2} = \begin{cases} 0, & \text{if } m \text{ is even,} \\ 1, & \text{if } m \text{ is odd.} \end{cases}$$

Since the function u is a solution to the partial differential equation in (3.1), all quantities $u^{(m,n)}, (m,n) \in \Lambda_{M+1}$ are not independent of each other. The next lemma describes this dependence.

Lemma 3.1. *Let u be a smooth function satisfying $\Delta u + \kappa^2 u = f$ in $\Omega \setminus \Gamma_I$. If a point $(x_i^*, y_j^*) \in \Omega \setminus \Gamma_I$, then*

$$u^{(m,n)} = (-1)^{\lfloor \frac{m}{2} \rfloor} \sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} \binom{\lfloor \frac{m}{2} \rfloor}{i} \kappa^{2i} u^{(\operatorname{odd}(m), 2\lfloor \frac{m}{2} \rfloor + n - 2i)} + \sum_{i=1}^{\lfloor \frac{m}{2} \rfloor} \sum_{j=0}^{i-1} (-1)^{i-1} \binom{i-1}{j} \kappa^{2(i-j-1)} f^{(m-2i, n+2j)} \tag{3.9}$$

for all $(m,n) \in \Lambda_{M+1}^{V,2}$, where

$$\Lambda_{M+1}^{V,2} := \Lambda_{M+1} \setminus \Lambda_{M+1}^{V,1} \quad \text{with} \quad \Lambda_{M+1}^{V,1} := \{(\ell, k-\ell) \ : k = \ell, \ldots, M+1-\ell \ \text{ and } \ \ell = 0, 1\}. \tag{3.10}$$

Define

$$\Lambda_{M+1}^{H,j} := \{(n,m) : (m,n) \in \Lambda_{M+1}^{V,j}, j = 1, 2\}.$$

If a point $(x_i^*, y_j^*) \in \Omega \setminus \Gamma_I$, then

$$u^{(m,n)} = (-1)^{\lfloor \frac{n}{2} \rfloor} \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{\lfloor \frac{n}{2} \rfloor}{i} \kappa^{2i} u^{(2\lfloor \frac{n}{2} \rfloor + m - 2i, \operatorname{odd}(n))} + \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sum_{j=0}^{i-1} (-1)^{i-1} \binom{i-1}{j} \kappa^{2(i-j-1)} f^{(m+2j, n-2i)} \tag{3.11}$$

for all $(m,n) \in \Lambda_{M+1}^{H,2}$.

Proof. The proof is similar to the proof of [49, Lemma 2.1] and [50, Lemma 2.1]. □

See [49, Figure 6] for an illustration of how each $u^{(m,n)}$ with $(m,n) \in \Lambda_7$ is categorized

based on $\Lambda_7^{V,j}$ with $j \in \{1,2\}$. From (3.9), we have

$$\sum_{(m,n)\in\Lambda_{M+1}^{V,2}} \frac{x^m y^n}{m!n!} u^{(m,n)} = \overbrace{\sum_{(m,n)\in\Lambda_{M+1}^{V,2}} \frac{x^m y^n}{m!n!} \Big\{ (-1)^{\lfloor \frac{m}{2} \rfloor} \sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} \binom{\lfloor \frac{m}{2} \rfloor}{i} \kappa^{2i} u^{(\mathrm{odd}(m), 2\lfloor \frac{m}{2} \rfloor + n - 2i)} \Big\}}^{=:I_1}$$
$$+ \underbrace{\sum_{(m,n)\in\Lambda_{M+1}^{V,2}} \frac{x^m y^n}{m!n!} \Big\{ \sum_{i=1}^{\lfloor \frac{m}{2} \rfloor} \sum_{j=0}^{i-1} (-1)^{i-1} \binom{i-1}{j} \kappa^{2i-2j-2} f^{(m-2i, n+2j)} \Big\}}_{=:I_2}, \tag{3.12}$$

where the first summation I_1 above can be expressed as

$$\begin{aligned}
I_1 &= \sum_{\substack{(m,n)\in\Lambda_{M+1}^{V,2} \\ \ell = \frac{m}{2},\ \text{even } m}} \frac{(-1)^\ell x^{2\ell} y^n}{(2\ell)!n!} \sum_{i=0}^{\ell} \binom{\ell}{i} \kappa^{2i} u^{(0, 2\ell+n-2i)} \\
&\quad + \sum_{\substack{(m,n)\in\Lambda_{M+1}^{V,2} \\ \ell = \frac{m-1}{2},\ \text{odd } m}} \frac{(-1)^\ell x^{2\ell+1} y^n}{(2\ell+1)!n!} \sum_{i=0}^{\ell} \binom{\ell}{i} \kappa^{2i} u^{(1, 2\ell+n-2i)} \\
&= \sum_{n=2}^{M+1} \sum_{\ell=1}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^\ell x^{2\ell} y^{n-2\ell}}{(2\ell)!(n-2\ell)!} \sum_{i=0}^{\ell} \binom{\ell}{i} \kappa^{2i} u^{(0, n-2i)} + \sum_{n=2}^{M} \sum_{\ell=1}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^\ell x^{2\ell+1} y^{n-2\ell}}{(2\ell+1)!(n-2\ell)!} \sum_{i=0}^{\ell} \binom{\ell}{i} \kappa^{2i} u^{(1, n-2i)} \\
&= \sum_{\substack{(m,n)\in\Lambda_{M+1}^{V,1} \\ n \geqslant 2}} \sum_{\ell=1}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^\ell x^{m+2\ell} y^{n-2\ell}}{(m+2\ell)!(n-2\ell)!} \sum_{i=0}^{\ell} \binom{\ell}{i} \kappa^{2i} u^{(m, n-2i)},
\end{aligned}$$

and the second summation I_2 above can be expressed as

$$\begin{aligned}
I_2 &= \sum_{(m,n)\in\Lambda_{M-1}} \sum_{\ell=1}^{1+\lfloor \frac{n}{2} \rfloor} \sum_{p=0}^{\ell-1} (-1)^{\ell-1} \binom{\ell-1}{p} \kappa^{2(\ell-p-1)} f^{(m, n+2(p+1-\ell))} \frac{x^{m+2\ell} y^{n-2\ell+2}}{(m+2\ell)!(n-2\ell+2)!} \\
&= \sum_{(m,n)\in\Lambda_{M-1}} \sum_{\substack{j\in\{n+2p \mid p\in\mathbb{N}_0, \\ n+2p\leqslant M+1-m\}}} \sum_{\ell=1+\frac{j-n}{2}}^{1+\lfloor \frac{j}{2} \rfloor} (-1)^{\ell-1} \binom{\ell-1}{\frac{j-n}{2}} \kappa^{j-n} \frac{x^{m+2\ell} y^{j-2\ell+2}}{(m+2\ell)!(j-2\ell+2)!} f^{(m,n)} \\
&= \sum_{(m,n)\in\Lambda_{M-1}} \underbrace{\sum_{p=0}^{\lfloor \frac{M+1-m-n}{2} \rfloor} \sum_{\ell=1+p}^{1+\lfloor p+\frac{n}{2} \rfloor} (-1)^{\ell-1} \binom{\ell-1}{p} \kappa^{2p} \frac{x^{m+2\ell} y^{2p+n+2-2\ell}}{(m+2\ell)!(2p+n+2-2\ell)!}}_{=:Q_{M,m,n}^V(x,y)} f^{(m,n)}.
\end{aligned} \tag{3.13}$$

Hence, using the right-hand side of (3.8) and the definitions of $\Lambda_{M+1}^{V,1}, \Lambda_{M+1}^{V,2}$ in (3.10), we

have

$$
\begin{aligned}
I_1+\sum_{(m,n)\in\Lambda^{V,1}_{M+1}}\frac{x^m y^n}{m!n!}u^{(m,n)} &= \sum_{(m,n)\in\Lambda^{V,1}_{M+1}}\sum_{i=0}^{\lfloor\frac{n}{2}\rfloor}\sum_{\ell=i}^{\lfloor\frac{n}{2}\rfloor}\frac{(-1)^\ell x^{m+2\ell}y^{n-2\ell}}{(m+2\ell)!(n-2\ell)!}\binom{\ell}{i}\kappa^{2i}u^{(m,n-2i)}\\
&= \sum_{(m,n)\in\Lambda^{V,1}_{M+1}}\sum_{\substack{i\in\{n+2p|p\in\mathbb{N}_0,\\ n+2p\leqslant M+1-m\}}}\sum_{\ell=\frac{i-n}{2}}^{\lfloor\frac{i}{2}\rfloor}\frac{(-1)^\ell x^{m+2\ell}y^{i-2\ell}}{(m+2\ell)!(i-2\ell)!}\binom{\ell}{\frac{i-n}{2}}\kappa^{i-n}u^{(m,n)}\\
&= \sum_{(m,n)\in\Lambda^{V,1}_{M+1}}\underbrace{\sum_{p=0}^{\lfloor\frac{M+1-m-n}{2}\rfloor}\sum_{\ell=p}^{p+\lfloor\frac{n}{2}\rfloor}\frac{(-1)^\ell x^{m+2\ell}y^{n+2p-2\ell}}{(m+2\ell)!(n+2p-2\ell)!}\binom{\ell}{p}\kappa^{2p}}_{=:G^V_{M,m,n}(x,y)}u^{(m,n)}.
\end{aligned}
\tag{3.14}
$$

Suppose $x,y\in(-2h,2h)$. The lowest degree of h for each polynomial $G^V_{M,m,n}(x,y)$ with $(m,n)\in\Lambda^{V,1}_{M+1}$ in (3.14) is $m+n$. The lowest degree of h for each polynomial $Q^V_{M,m,n}(x,y)$ with $(m,n)\in\Lambda^V_{M-1}$ in (3.13) is $m+n+2$. Therefore, by (3.12)-(3.13), we can rewrite the approximation of $u(x+x_i^*,y+y_j^*)$ with $(x,y)\in(-2h,2h)$ in (3.8) as follows:

$$
u(x+x_i^*,y+y_j^*)=\sum_{(m,n)\in\Lambda^{V,1}_{M+1}}u^{(m,n)}G^V_{M,m,n}(x,y)+\sum_{(m,n)\in\Lambda_{M_f-1}}f^{(m,n)}Q^V_{M_f,m,n}(x,y)+\mathcal{O}(h^{M+2}),
\tag{3.15}
$$

where $M,M_f\in\mathbb{N}_0$ and $M_f\geqslant M$. By a similar calculation, for $(x,y)\in(-2h,2h)$, we also have

$$
u(x+x_i^*,y+y_j^*)=\sum_{(m,n)\in\Lambda^{H,1}_{M+1}}u^{(m,n)}G^H_{M,m,n}(x,y)+\sum_{(m,n)\in\Lambda_{M_f-1}}f^{(m,n)}Q^H_{M_f,m,n}(x,y)+\mathcal{O}(h^{M+2}),
\tag{3.16}
$$

where $M,M_f\in\mathbb{N}_0$, $M_f\geqslant M$ and

$$
\begin{aligned}
G^H_{M,m,n}(x,y)&:=G^V_{M,n,m}(y,x),\quad \text{for all } n\in\{0,1\},m\in\mathbb{N}_0\\
Q^H_{M,m,n}(x,y)&:=Q^V_{M,n,m}(y,x),\quad \text{for all } m,n\in\mathbb{N}_0.
\end{aligned}
\tag{3.17}
$$

Identities (3.15)-(3.16) are critical in finding compact stencils achieving a desired accuracy order.

In the following subsections, we shall explicitly present our stencils having at least accuracy order six with reduced pollution effect for interior, boundary and corner points. As we shall explain in details in Section 3.3, we construct such stencils by first finding a general expression for all possible discretization stencils achieving the maximum order. Then we minimize the average truncation error of plane waves to determine the remaining free

parameters in each stencil to reduce pollution effect.

3.1.1 Regular points (interior)

In this subsection, we state one of our main results on a sixth order (which is the highest possible order) compact finite difference scheme (with reduced pollution effect) centered at a regular point (x_i, y_j) and $(x_i, y_j) \notin \partial\Omega$. We let (x_i, y_j) be the base point (x_i^*, y_j^*) by setting $v_0 = w_0 = 0$ in (3.5). The proof of the following theorem is deferred to Section 3.3.

Theorem 3.2. *Let a grid point (x_i, y_j) be a regular point, i.e., either $d_{i,j}^+ = \emptyset$ or $d_{i,j}^- = \emptyset$ and $(x_i, y_j) \notin \partial\Omega$. Let $(u_h)_{i,j}$ be the numerical approximated solution of the exact solution u of the Helmholtz equation (3.1) at an interior regular point (x_i, y_j). Then the following discretization stencil centered at (x_i, y_j)*

$$\begin{aligned}\mathcal{L}_h u_h := \; & h^{-2}(C_{1,1}(u_h)_{i-1,j-1} + C_{1,0}(u_h)_{i,j-1} + C_{1,1}(u_h)_{i+1,j-1} \\ & + C_{1,0}(u_h)_{i-1,j} + C_{0,0}(u_h)_{i,j} + C_{1,0}(u_h)_{i+1,j} \\ & + C_{1,1}(u_h)_{i-1,j+1} + C_{1,0}(u_h)_{i,j+1} + C_{1,1}(u_h)_{i+1,j+1}) = \sum_{(m,n)\in\Lambda_6} h^{-2} f^{(m,n)} C_{f,m,n},\end{aligned} \tag{3.18}$$

achieves the sixth order accuracy for $\Delta u + \kappa^2 u = f$ at the point (x_i, y_j) with reduced pollution effect, where $C_{f,m,n} := \sum_{k=-1}^{1}\sum_{\ell=-1}^{1} C_{k,\ell} Q_{7,m,n}^V(kh, \ell h)$ for all $(m,n) \in \Lambda_6$, $Q_{7,m,n}^V(x,y)$ is defined in (3.13), and

$$\begin{aligned}
&C_{-1,-1} = C_{-1,1} = C_{1,-1} = C_{1,1}, \quad C_{-1,0} = C_{0,-1} = C_{0,1} = C_{1,0},\\
&C_{1,1} = 1 - \tfrac{357462387}{25\times10^{10}}\kappa h + \tfrac{1001065991}{2\times10^{10}}(\kappa h)^2 - \tfrac{196477327}{2\times10^{12}}(\kappa h)^3 + \tfrac{1155977087}{10^{12}}(\kappa h)^4\\
&\qquad - \tfrac{116352513}{4\times10^{13}}(\kappa h)^5 + \tfrac{1255955641}{10^{14}}(\kappa h)^6,\\
&C_{1,0} = 4 - \tfrac{357462387}{625\times10^{8}}\kappa h + \tfrac{532995477}{25\times10^{11}}(\kappa h)^2 - \tfrac{267461861}{25\times10^{11}}(\kappa h)^3 - \tfrac{288674231}{10^{11}}(\kappa h)^4\\
&\qquad + \tfrac{2179972749}{5\times10^{14}}(\kappa h)^5 - \tfrac{3473210401}{5\times10^{13}}(\kappa h)^6,\\
&C_{0,0} = -20 + \tfrac{357462387}{125\times10^{8}}\kappa h + \tfrac{5798934009}{10^{9}}(\kappa h)^2 - \tfrac{969775457}{125\times10^{9}}(\kappa h)^3 - \tfrac{1963785709}{5\times10^{9}}(\kappa h)^4\\
&\qquad + \tfrac{4056581719}{10^{13}}(\kappa h)^5 + \tfrac{795951403}{10^{11}}(\kappa h)^6.
\end{aligned} \tag{3.19}$$

Moreover, the maximum accuracy order of a compact finite difference scheme for $\Delta u + \kappa^2 u = f$ at the point (x_i, y_j) is six.

3.1.2 Boundary and corner points

In this subsection, we discuss how to find a compact finite difference scheme centered at $(x_i, y_j) \in \partial\Omega$.

3.1.2.1 Boundary points

We first discuss in detail how the left boundary (i.e., $(x_i, y_j) \in \Gamma_1 = \{l_1\} \times (l_3, l_4)$) stencil is constructed. The stencils for the other three boundaries can afterwards be obtained by symmetry. If $\mathcal{B}_1 u = u = g_1$ on Γ_1, then the left boundary stencil can be directly obtained from (3.18)-(3.19) in Theorem 3.2 by replacing $(u_h)_{0,j-1}$, $(u_h)_{0,j}$, and $(u_h)_{0,j+1}$ with $g_1(y_{j-1})$, $g_1(y_j)$, and $g_1(y_{j+1})$ respectively, where $y_j \in (l_3, l_4)$, and moving terms involving these known boundary values to the right-hand side of (3.18). The other three boundary sides are dealt in a similar straightforward fashion if a Dirichlet boundary condition is present. On the other hand, the stencils for the other two boundary conditions are not trivial at all. The following theorem provides the explicit 6-point stencil of accuracy order at least six with reduced pollution effect for the left boundary operator $\mathcal{B}_1 \in \{\frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa \mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$. The proof of the following result is deferred to Section 3.3.

Theorem 3.3. *Assume $\Omega = (l_1, l_2) \times (l_3, l_4)$. Let $(u_h)_{i,j}$ be the numerical approximated solution of the exact solution u of the Helmholtz equation (3.1) at the point (x_i, y_j). Consider the following discretization stencil centered at $(x_0, y_j) \in \Gamma_1$ for $\mathcal{B}_1 u = g_1$ on Γ_1 with $\mathcal{B}_1 \in \{\frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa \mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$:*

$$\mathcal{L}_h^{\mathcal{B}_1} u_h := \begin{array}{ccc} h^{-1}(C_{0,1}^{\mathcal{B}_1}(u_h)_{0,j-1} & +C_{1,1}^{\mathcal{B}_1}(u_h)_{1,j-1} & \\ +C_{0,0}^{\mathcal{B}_1}(u_h)_{0,j} & +C_{1,0}^{\mathcal{B}_1}(u_h)_{1,j} & = \sum_{(m,n)\in\Lambda_6} h^{-1} f^{(m,n)} C_{f,m,n}^{\mathcal{B}_1} + \sum_{n=0}^{7} h^{-1} g_1^{(n)} C_{g_1,n}^{\mathcal{B}_1}, \\ +C_{0,1}^{\mathcal{B}_1}(u_h)_{0,j+1} & +C_{1,1}^{\mathcal{B}_1}(u_h)_{1,j+1}) & \end{array} \tag{3.20}$$

where $\{C_{k,\ell}^{\mathcal{B}_1}\}_{k\in\{0,1\},\ell\in\{-1,0,1\}}$ are polynomials of κh, $C_{f,m,n}^{\mathcal{B}_1} = \sum_{k=0}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}^{\mathcal{B}_1} Q_{7,m,n}^{V}(kh, \ell h)$ for all $(m,n) \in \Lambda_6$, $Q_{7,m,n}^{V}$ is defined in (3.13), $g_1^{(n)} := \frac{d^n g_1}{dy^n}(y_j)$, $C_{g_1,n}^{\mathcal{B}_1} = -\sum_{k=0}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}^{\mathcal{B}_1} G_{7,1,n}^{V}(kh, \ell h)$ for all $n = 0, \ldots, 7$, $G_{7,1,n}^{V}$ is defined in (3.14), $C_{0,-1}^{\mathcal{B}_1} = C_{0,1}^{\mathcal{B}_1}$, and $C_{1,-1}^{\mathcal{B}_1} = C_{1,1}^{\mathcal{B}_1}$.

(1) For $\mathcal{B}_1 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa \mathbf{I}_d$, the coefficients for defining $\mathcal{L}_h^{\mathcal{B}_1} u_h$ in (3.20) are given by

$$\begin{aligned}
C_{1,1}^{\mathcal{B}_1} &= 1 - \tfrac{218737123}{10^9}\kappa h + \tfrac{6698622893i}{10^{10}}\kappa h - \tfrac{1620223367}{10^{10}}(\kappa h)^2 - \tfrac{1202725989i}{10^{10}}(\kappa h)^2 \\
&\quad + \tfrac{3105005559}{10^{11}}(\kappa h)^3 - \tfrac{1252107029i}{10^{11}}(\kappa h)^3 - \tfrac{3412232989}{10^{12}}(\kappa h)^4 - \tfrac{1505046263i}{10^{12}}(\kappa h)^4, \\
C_{0,1}^{\mathcal{B}_1} &= 2 - \tfrac{218737123}{5\times 10^8}\kappa h + \tfrac{1139724579i}{10^9}\kappa h - \tfrac{3034055489}{10^{10}}(\kappa h)^2 - \tfrac{1967977733i}{10^{10}}(\kappa h)^2 \\
&\quad + \tfrac{1090897501}{25\times 10^9}(\kappa h)^3 - \tfrac{7785677273i}{10^{11}}(\kappa h)^3 + \tfrac{98544681}{4\times 10^9}(\kappa h)^4 + \tfrac{1218033221i}{5\times 10^{10}}(\kappa h)^4, \\
C_{1,0}^{\mathcal{B}_1} &= 4 - \tfrac{8749484921}{10^{10}}\kappa h + \tfrac{2279449157i}{10^9}\kappa h - \tfrac{946955529}{2\times 10^9}(\kappa h)^2 - \tfrac{1967977733i}{5\times 10^9}(\kappa h)^2 \\
&\quad + \tfrac{2905342517}{5\times 10^{10}}(\kappa h)^3 - \tfrac{1542150899i}{5\times 10^{10}}(\kappa h)^3 + \tfrac{2645544603}{10^{12}}(\kappa h)^4 + \tfrac{302693249i}{25\times 10^9}(\kappa h)^4, \\
C_{0,0}^{\mathcal{B}_1} &= -10 + \tfrac{218737123}{10^8}\kappa h + \tfrac{202754213i}{2\times 10^9}\kappa h + \tfrac{7851597997}{10^{10}}(\kappa h)^2 - \tfrac{2846864471i}{10^{10}}(\kappa h)^2 \\
&\quad - \tfrac{1147746931}{5\times 10^9}(\kappa h)^3 + \tfrac{2236631341i}{10^{10}}(\kappa h)^3 - \tfrac{1738692843}{5\times 10^{10}}(\kappa h)^4 - \tfrac{898631349i}{25\times 10^9}(\kappa h)^4.
\end{aligned} \tag{3.21}$$

Then the finite difference scheme in (3.20) achieves sixth order accuracy for $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa u = g_1$ at the point $(x_0, y_j) \in \Gamma_1$ with reduced pollution effect. The maximum accuracy order of a 6-point finite difference scheme for $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa u = g_1$ at the point $(x_0, y_j) \in \Gamma_1$ is six.

(2) For $\mathcal{B}_1 = \frac{\partial}{\partial \boldsymbol{\nu}}$, the coefficients for defining $\mathcal{L}_h^{\mathcal{B}_1} u_h$ in (3.20) are given by

$$\begin{aligned}
C_{1,1}^{\mathcal{B}_1} &= 1 + \tfrac{1915061419}{25\times 10^9}(\kappa h)^2 + \tfrac{3019639439}{10^{12}}(\kappa h)^4, \\
C_{0,1}^{\mathcal{B}_1} &= 2 + \tfrac{665061419}{125\times 10^8}(\kappa h)^2 - \tfrac{1071383831}{2\times 10^{12}}(\kappa h)^4, \\
C_{1,0}^{\mathcal{B}_1} &= 4 + \tfrac{106409827}{10^9}(\kappa h)^2 - \tfrac{1071383831}{10^{12}}(\kappa h)^4, \\
C_{0,0}^{\mathcal{B}_1} &= -10 + \tfrac{1316987716}{5\times 10^8}(\kappa h)^2 - \tfrac{1240891409}{10^{10}}(\kappa h)^4.
\end{aligned} \tag{3.22}$$

Then the finite difference scheme in (3.20) achieves seventh order accuracy for $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} = g_1$ at the point $(x_0, y_j) \in \Gamma_1$ with reduced pollution effect. Moreover, the maximum accuracy order of a 6-point finite difference scheme for $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} = g_1$ at the point $(x_0, y_j) \in \Gamma_1$ is seven.

By symmetry, we can immediately state the stencils for the other three boundary sides. Same accuracy order results as in Theorem 3.3 hold. First, consider the following discretization stencil for $\mathcal{B}_2 u = g_2$ on Γ_2 with $\mathcal{B}_2 \in \{\frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa \mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$ centered at $(x_{N_1}, y_j) \in \Gamma_2$:

$$\mathcal{L}_h^{\mathcal{B}_2} u_h := \sum_{k=-1}^{0} \sum_{\ell=-1}^{1} h^{-1} C_{k,\ell}^{\mathcal{B}_2} (u_h)_{N_1+k, j+\ell} = \sum_{(m,n)\in\Lambda_6} h^{-1} f^{(m,n)} C_{f,m,n}^{\mathcal{B}_2} + \sum_{n=0}^{7} h^{-1} g_2^{(n)} C_{g_2,n}^{\mathcal{B}_2},$$

where $C_{-k,\ell}^{\mathcal{B}_2} = C_{k,\ell}^{\mathcal{B}_1}$ for all $k \in \{0,1\}$, $\ell \in \{-1,0,1\}$, $C_{f,m,n}^{\mathcal{B}_2} = \sum_{k=-1}^{0} \sum_{\ell=-1}^{1} C_{7,m,n}^{V}(kh, \ell h)$ for all $(m,n) \in \Lambda_6$, $g_2^{(n)} := \frac{d^n g_2}{dy^n}(y_j)$, $C_{g_2,n}^{\mathcal{B}_2} = \sum_{k=-1}^{0} \sum_{\ell=-1}^{1} C_{k,\ell}^{\mathcal{B}_2} G_{7,1,n}^{V}(kh, \ell h)$ for all $n = 0, \ldots, 7$.

Second, the stencil for $\mathcal{B}_3 u = g_3$ on Γ_3 with $\mathcal{B}_3 \in \{\frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa \mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$ centered at $(x_i, y_0) \in \Gamma_3$ is

$$\mathcal{L}_h^{\mathcal{B}_3} u_h := \sum_{k=-1}^{1} \sum_{\ell=0}^{1} h^{-1} C_{k,\ell}^{\mathcal{B}_3} (u_h)_{i+k,\ell} = \sum_{(m,n)\in\Lambda_6} h^{-1} f^{(m,n)} C_{f,m,n}^{\mathcal{B}_3} + \sum_{n=0}^{7} h^{-1} g_3^{(n)} C_{g_3,n}^{\mathcal{B}_3},$$

where $C_{\ell,k}^{\mathcal{B}_3} = C_{k,\ell}^{\mathcal{B}_1}$ for all $k \in \{0,1\}$, $\ell \in \{-1,0,1\}$, $C_{f,m,n}^{\mathcal{B}_3} = \sum_{k=-1}^{1} \sum_{\ell=0}^{1} C_{k,\ell}^{\mathcal{B}_3} Q_{7,m,n}^{H}(kh, \ell h)$ for all $(m,n) \in \Lambda_6$, $Q_{7,m,n}^{H}$ is defined in (3.17), $g_3^{(n)} := \frac{d^n g_3}{dx^n}(x_i)$, $C_{g_3,n}^{\mathcal{B}_3} = -\sum_{k=-1}^{1} \sum_{\ell=0}^{1} C_{k,\ell}^{\mathcal{B}_3} G_{7,n,1}^{H}(kh, \ell h)$ for all $n = 0, \ldots, 7$, and $G_{7,n,1}^{H}$ is defined in (3.17).

Third, the stencil for $\mathcal{B}_4 u = g_4$ on Γ_4 with $\mathcal{B}_4 \in \{\frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa \mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$ centered at $(x_i, y_{N_2}) \in \Gamma_4$ is

$$\mathcal{L}_h^{\mathcal{B}_4} u_h := \sum_{k=-1}^{1} \sum_{\ell=-1}^{0} h^{-1} C_{k,\ell}^{\mathcal{B}_4} (u_h)_{i+k,N_2+\ell} = \sum_{(m,n)\in\Lambda_6} h^{-1} f^{(m,n)} C_{f,m,n}^{\mathcal{B}_4} + \sum_{n=0}^{7} h^{-1} g_4^{(n)} C_{g_4,n}^{\mathcal{B}_4},$$

where $C_{\ell,-k}^{\mathcal{B}_4} = C_{k,\ell}^{\mathcal{B}_1}$ for all $k \in \{0,1\}$, $\ell \in \{-1,0,1\}$, $C_{f,m,n}^{\mathcal{B}_4} = \sum_{k=-1}^{1} \sum_{\ell=-1}^{0} C_{k,\ell}^{\mathcal{B}_4} Q_{7,m,n}^{H}(kh, \ell h)$ for all $(m,n) \in \Lambda_6$, $g_4^{(n)} := \frac{d^n g_4}{dx^n}(x_i)$, and $C_{g_4,n}^{\mathcal{B}_4} = \sum_{k=-1}^{1} \sum_{\ell=-1}^{0} C_{k,\ell}^{\mathcal{B}_4} G_{7,n,1}^{H}(kh, \ell h)$ for all $n = 0, \ldots, 7$.

3.1.2.2 Corner points

For clarity of presentation, let us consider the following boundary configuration

$$\begin{aligned} &\mathcal{B}_1 u = \tfrac{\partial u}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa u = g_1 \ \text{ on } \ \Gamma_1, \quad \mathcal{B}_2 u = u = g_2 \ \text{ on } \ \Gamma_2, \\ &\mathcal{B}_3 u = \tfrac{\partial u}{\partial \boldsymbol{\nu}} = g_3 \ \text{ on } \ \Gamma_3, \qquad\quad \mathcal{B}_4 u = \tfrac{\partial u}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa u = g_4 \ \text{ on } \ \Gamma_4. \end{aligned} \tag{3.23}$$

See Fig. 3.1 for an illustration.

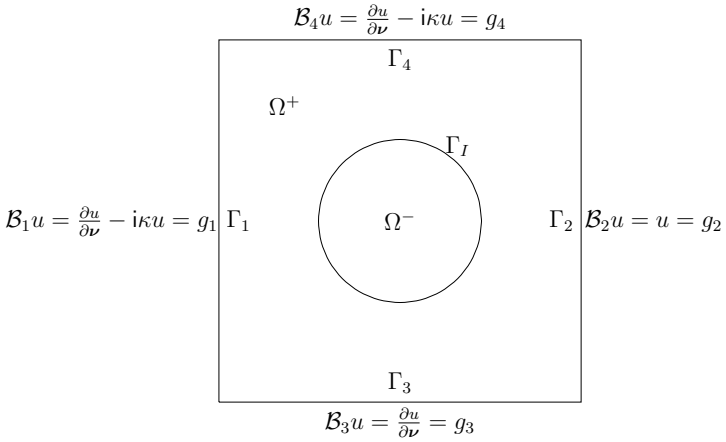


Figure 3.1: Boundary configuration in (3.23), where $\psi(x,y) = x^2 + y^2 - 2$.

The corners coming from other boundary configurations can be handled in a similar way. When a corner involves at least one Dirichlet boundary condition, we can use Theorem 3.3 and subsequent remarks to handle it. We denote the bottom left corner (the intersection of Γ_1 and Γ_3) by $\mathcal{R}_1$, and the top left corner (the intersection of Γ_1 and Γ_4) by $\mathcal{R}_2$. In what follows, we discuss in detail how the bottom and top left stencils are constructed. The following two theorems provide the 4-point stencils of accuracy order at least six with reduced pollution effect for the left corners. Their proofs are deferred to Section 3.3.

Theorem 3.4. *Assume $\Omega = (l_1, l_2) \times (l_3, l_4)$. Let $(u_h)_{i,j}$ be the numerical approximated solution of the exact solution u of the Helmholtz equation (3.1) at the point (x_i, y_j). Then the following discretization stencil centered at the corner point (x_0, y_0):*

$$\begin{aligned}\mathcal{L}_h^{\mathcal{R}_1} u_h :=& \begin{array}{ll} h^{-1}(C_{0,0}^{\mathcal{R}_1}(u_h)_{0,0} & +C_{1,0}^{\mathcal{R}_1}(u_h)_{1,0} \\ +C_{0,1}^{\mathcal{R}_1}(u_h)_{0,1} & +C_{1,1}^{\mathcal{R}_1}(u_h)_{1,1}) \end{array} \\ =& \sum_{(m,n)\in\Lambda_6} h^{-1} f^{(m,n)} C_{f,m,n}^{\mathcal{R}_1} + \sum_{n=0}^{7} h^{-1} g_1^{(n)} C_{g_1,n}^{\mathcal{R}_1} + \sum_{n=0}^{7} h^{-1} g_3^{(n)} C_{g_3,n}^{\mathcal{R}_1},\end{aligned} \tag{3.24}$$

where

$$\begin{aligned} C_{1,1}^{\mathcal{R}_1} &= 1 - \tfrac{2041589737}{10^{10}}\kappa h + \tfrac{6666011379i}{10^{10}}\kappa h - \tfrac{1213438849}{10^{10}}(\kappa h)^2 - \tfrac{254718888i}{25\times 10^8}(\kappa h)^2 \\ &\quad + \tfrac{2199377569}{10^{11}}(\kappa h)^3 + \tfrac{4307308979i}{5\times 10^{11}}(\kappa h)^3 - \tfrac{5536966589}{10^{12}}(\kappa h)^4 - \tfrac{1556373503i}{10^{12}}(\kappa h)^4, \\ C_{1,0}^{\mathcal{R}_1} &= 2 - \tfrac{2041589737}{5\times 10^9}\kappa h + \tfrac{566601138i}{5\times 10^8}\kappa h - \tfrac{156034209}{10^9}(\kappa h)^2 - \tfrac{1629433157i}{10^{10}}(\kappa h)^2 \\ &\quad + \tfrac{1855012159}{10^{11}}(\kappa h)^3 + \tfrac{453336943i}{2\times 10^{10}}(\kappa h)^3 - \tfrac{3170819689}{5\times 10^{11}}(\kappa h)^4 + \tfrac{25677723i}{8\times 10^9}(\kappa h)^4, \\ C_{0,1}^{\mathcal{R}_1} &= 2 - \tfrac{2041589737}{5\times 10^9}\kappa h + \tfrac{566601138i}{5\times 10^8}\kappa h - \tfrac{556752189}{25\times 10^8}(\kappa h)^2 - \tfrac{1629433157i}{10^{10}}(\kappa h)^2 \\ &\quad + \tfrac{3216071983}{10^{11}}(\kappa h)^3 - \tfrac{3955100649i}{10^{11}}(\kappa h)^3 + \tfrac{1546871341}{10^{11}}(\kappa h)^4 + \tfrac{231176972i}{125\times 10^8}(\kappa h)^4, \\ C_{0,0}^{\mathcal{R}_1} &= -5 + \tfrac{510397434}{5\times 10^8}\kappa h + \tfrac{6699431033i}{10^{11}}\kappa h + \tfrac{2002755557}{10^{10}}(\kappa h)^2 - \tfrac{369405469i}{2\times 10^9}(\kappa h)^2 \\ &\quad - \tfrac{285280517}{25\times 10^8}(\kappa h)^3 + \tfrac{326982886i}{25\times 10^8}(\kappa h)^3 + \tfrac{35165403}{25\times 10^9}(\kappa h)^4 - \tfrac{9939550949i}{10^{12}}(\kappa h)^4, \end{aligned} \tag{3.25}$$

$g_1^{(n)} := \frac{d^n g_1}{dy^n}(y_0)$, $g_3^{(n)} := \frac{d^n g_3}{dx^n}(x_0)$ for all $n = 0, \ldots, 7$, and $\{C_{f,m,n}^{\mathcal{R}_1}\}_{(m,n)\in\Lambda_6}$, $\{C_{g_1,n}^{\mathcal{R}_1}\}_{n=0}^7$, $\{C_{g_3,n}^{\mathcal{R}_1}\}_{n=0}^7$ are well-defined stencil coefficients that uniquely depend on $\{C_{k,\ell}^{\mathcal{R}_1}\}_{k,\ell\in\{0,1\}}$, achieves sixth order for $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - i\kappa u = g_1$ and $\mathcal{B}_3 u = \frac{\partial u}{\partial \boldsymbol{\nu}} = g_3$ at the point (x_0, y_0) with reduced pollution effect. Moreover, the maximum accuracy order of a 4-point finite difference scheme for $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - i\kappa u = g_1$ and $\mathcal{B}_3 u = \frac{\partial u}{\partial \boldsymbol{\nu}} = g_3$ at the point (x_0, y_0) is six.

Theorem 3.5. *Assume $\Omega = (l_1, l_2) \times (l_3, l_4)$. Let $(u_h)_{i,j}$ be the numerical approximated solution of the exact solution u of the Helmholtz equation (3.1) at the point (x_i, y_j). Then*

the following discretization stencil centered at the corner point (x_0, y_{N_2})*:*

$$\begin{aligned}\mathcal{L}_h^{\mathcal{R}_2} u_h :=& \; h^{-1}(C_{1,0}^{\mathcal{R}_2}(u_h)_{0,N_2-1} + C_{1,-1}^{\mathcal{R}_2}(u_h)_{1,N_2-1} \\ & + C_{0,0}^{\mathcal{R}_2}(u_h)_{0,N_2} + C_{1,0}^{\mathcal{R}_2}(u_h)_{1,N_2}) \\ =& \sum_{(m,n)\in\Lambda_6} h^{-1} f^{(m,n)} C_{f,m,n}^{\mathcal{R}_2} + \sum_{n=0}^{7} h^{-1} g_1^{(n)} C_{g_1,n}^{\mathcal{R}_2} + \sum_{n=0}^{7} h^{-1} g_4^{(n)} C_{g_4,n}^{\mathcal{R}_2},\end{aligned} \tag{3.26}$$

where

$$\begin{aligned}C_{1,-1}^{\mathcal{R}_2} &= 1 - \tfrac{535927359}{5\times10^9}\kappa h + \tfrac{131913924i}{10^8}\kappa h - \tfrac{4650641357}{10^{10}}(\kappa h)^2 - \tfrac{3255802571i}{10^{11}}(\kappa h)^2 \\ &\quad - \tfrac{1802358661}{10^{13}}(\kappa h)^3 - \tfrac{137039551i}{25\times10^8}(\kappa h)^3 - \tfrac{116115549}{625\times10^8}(\kappa h)^4 - \tfrac{390383949i}{2\times10^{11}}(\kappa h)^4, \\ C_{1,0}^{\mathcal{R}_2} &= 2 - \tfrac{428741887}{2\times10^9}\kappa h + \tfrac{2238278479i}{10^9}\kappa h - \tfrac{5558059089}{10^{10}}(\kappa h)^2 - \tfrac{278023284i}{125\times10^8}(\kappa h)^2 \\ &\quad + \tfrac{1525711827}{5\times10^{11}}(\kappa h)^3 - \tfrac{57317954i}{5\times10^8}(\kappa h)^3 + \tfrac{2099795921}{10^{11}}(\kappa h)^4 + \tfrac{1100929919i}{2\times10^{11}}(\kappa h)^4, \\ C_{0,0}^{\mathcal{R}_2} &= -5 + \tfrac{1339818397}{25\times10^8}\kappa h + \tfrac{2043038021i}{10^{10}}\kappa h - \tfrac{1519079742}{5\times10^8}(\kappa h)^2 - \tfrac{2830355397i}{5\times10^9}(\kappa h)^2 \\ &\quad - \tfrac{82143257}{5\times10^8}(\kappa h)^3 + \tfrac{3401956461i}{10^{10}}(\kappa h)^3, + \tfrac{1420360677}{5\times10^9}(\kappa h)^4 + \tfrac{4391249797i}{10^{11}}(\kappa h)^4,\end{aligned} \tag{3.27}$$

$g_1^{(n)} := \frac{d^n g_1}{dy^n}(y_{N_2})$, $g_4^{(n)} := \frac{d^n g_4}{dx^n}(x_0)$ *for all* $n = 0, \ldots, 7$, *and* $\{C_{f,m,n}^{\mathcal{R}_2}\}_{(m,n)\in\Lambda_6}$, $\{C_{g_1,n}^{\mathcal{R}_2}\}_{n=0}^7$, $\{C_{g_4,n}^{\mathcal{R}_2}\}_{n=0}^7$ *are well-defined stencil coefficients that uniquely depend on* $\{C_{k,\ell}^{\mathcal{R}_2}\}_{k\in\{0,1\},\ell\in\{-1,0\}}$ *with* $C_{0,-1}^{\mathcal{R}_2} = C_{1,0}^{\mathcal{R}_2}$*, achieves seventh order accuracy for* $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - \imath\kappa u = g_1$ *and* $\mathcal{B}_4 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - \imath\kappa u = g_4$ *at the point* (x_0, y_{N_2}) *with reduced pollution effect. Moreover, the maximum accuracy order of a 4-point finite difference scheme for* $\mathcal{B}_1 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - \imath\kappa u = g_1$ *and* $\mathcal{B}_4 u = \frac{\partial u}{\partial \boldsymbol{\nu}} - \imath\kappa u = g_4$ *at the point* (x_0, y_{N_2}) *is seven.*

Note that the right-hand sides of (3.24) and (3.26) can be explicitly recovered. See the proofs of Theorems 3.4 and 3.5 in Section 3.3 for details.

3.1.3 Irregular points

Let (x_i, y_j) be an irregular point (i.e., both $d_{i,j}^+$ and $d_{i,j}^-$ are nonempty) and let us take a base point $(x_i^*, y_j^*) \in \Gamma_I \cap (x_i - h, x_i + h) \times (y_j - h, y_j + h)$ on the interface Γ_I and inside $(x_i - h, x_i + h) \times (y_j - h, y_j + h)$. By (3.5), we have

$$x_i^* = x_i - v_0 h \quad \text{and} \quad y_j^* = y_j - w_0 h \quad \text{with} \quad -1 < v_0, w_0 < 1 \quad \text{and} \quad (x_i^*, y_j^*) \in \Gamma_I. \tag{3.28}$$

Let $u_\pm$ and $f_\pm$ represent the solution u and source term f in Ω^+ or Ω^-, respectively. Similar to (3.6), the following notations are used

$$u_\pm^{(m,n)} := \frac{\partial^{m+n} u_\pm}{\partial^m x \partial^n y}(x_i^*, y_j^*), \qquad f_\pm^{(m,n)} := \frac{\partial^{m+n} f_\pm}{\partial^m x \partial^n y}(x_i^*, y_j^*),$$
$$g_D^{(m,n)} := \frac{\partial^{m+n} g_D}{\partial^m x \partial^n y}(x_i^*, y_j^*), \qquad g_N^{(m,n)} := \frac{\partial^{m+n} g_N}{\partial^m x \partial^n y}(x_i^*, y_j^*).$$

Since the interface curve Γ_I is smooth and the solution u and the source term f are assumed to be piecewise smooth, we can extend u_+ and f_+ on Ω^+ into smooth functions in a neighborhood of (x_i^*, y_j^*). The same applies to u_- and f_- on Ω^-. Identity similar to (3.15) still holds:

$$u_\pm(x + x_i^*, y + y_j^*) = \sum_{(m,n)\in\Lambda_{M+1}^{V,1}} u_\pm^{(m,n)} G_{M,m,n}^V(x,y) + \sum_{(m,n)\in\Lambda_{M_f-1}} f_\pm^{(m,n)} Q_{M_f,m,n}^V(x,y) + \mathscr{O}(h^{M+2}),$$

for $x, y \in (-2h, 2h)$, where $\Lambda_{M+1}^{V,1}$ is defined in (3.10), Λ_{M_f-1} is defined in (3.7), $G_{M,m,n}^V(x,y)$ is defined in (3.14), $Q_{M_f,m,n}^V(x,y)$ is defined in (3.13). As in [49, 50], we assume that we have a parametric equation for Γ_I near the base point (x_i^*, y_j^*). I.e.,

$$x = r(t) + x_i^*, \quad y = s(t) + y_j^*, \quad (r'(t))^2 + (s'(t))^2 > 0 \quad \text{for } t \in (-\epsilon, \epsilon) \quad \text{with} \quad \epsilon > 0, \tag{3.29}$$

where r and s are smooth functions.

Theorem 3.6. *Let u be the solution to the Helmholtz interface problem in (3.1) and the base point $(x_i^*, y_j^*) \in \Gamma_I$ be parameterized near (x_i^*, y_j^*) by (3.29). Then*

$$u_-^{(m',n')} = u_+^{(m',n')} + \sum_{(m,n)\in\Lambda_{M-1}} \left(T^+_{m',n',m,n} f_+^{(m,n)} + T^-_{m',n',m,n} f_-^{(m,n)} \right) + \sum_{(m,n)\in\Lambda_{M+1}} T^{g_D}_{m',n',m,n} g_D^{(m,n)}$$
$$+ \sum_{(m,n)\in\Lambda_M} T^{g_N}_{m',n',m,n} g_N^{(m,n)}, \qquad \forall\, (m',n') \in \Lambda_{M+1}^{V,1},$$

where all the transmission coefficients $T^\pm, T^{g_D}, T^{g_N}$ are uniquely determined by $r^{(k)}(0)$, $s^{(k)}(0)$, and κ for $k = 0, \ldots, M+1$.

Proof. The proof closely follows from the proof of [49, Theorem 2.3]. □

Next, we state the compact finite difference stencil for interior irregular points.

Theorem 3.7. *Let $(u_h)_{i,j}$ be the numerical solution of (3.1) at an interior irregular point (x_i, y_j). Pick a base point (x_i^*, y_j^*) as in (3.28). Then the following compact scheme centered*

at the interior irregular point (x_i, y_j)

$$\begin{aligned}\mathcal{L}_h^{\Gamma_I} :=\ & h^{-1}(C_{1,1}(u_h)_{i-1,j-1} + C_{1,0}(u_h)_{i,j-1} + C_{1,1}(u_h)_{i+1,j-1} \\ & + C_{1,0}(u_h)_{i-1,j} + C_{0,0}(u_h)_{i,j} + C_{1,0}(u_h)_{i+1,j} \\ & + C_{1,1}(u_h)_{i-1,j+1} + C_{1,0}(u_h)_{i,j+1} + C_{1,1}(u_h)_{i+1,j+1}) \\ =\ & \sum_{(m,n)\in\Lambda_6} h^{-1} f_+^{(m,n)} J_{m,n}^+ + \sum_{(m,n)\in\Lambda_6} h^{-1} f_-^{(m,n)} J_{m,n}^- + \sum_{(m,n)\in\Lambda_8} h^{-1} g_D^{(m,n)} J_{m,n}^{g_D} \\ & + \sum_{(m,n)\in\Lambda_7} h^{-1} g_N^{(m,n)} J_{m,n}^{g_N},\end{aligned}$$

achieves seventh order accuracy, where $\{C_{k,\ell}\}_{k,\ell\in\{-1,0,1\}}$ *are defined in* (3.19), $J_{m,n}^{\pm} := J_{m,n}^{\pm,0} + J_{m,n}^{\pm,T}$ *for all* $(m,n) \in \Lambda_6$,

$$J_{m,n}^{\pm,0} := \sum_{(k,\ell)\in d_{i,j}^{\pm}} C_{k,\ell} Q_{7,m,n}^V((v_0+k)h, (w_0+\ell)h), \quad J_{m,n}^{\pm,T} := \sum_{(m',n')\in\Lambda_8^{V,1}} I_{m',n'}^- T_{m',n',m,n}^{\pm}, \quad \forall (m,n)\in\Lambda_6,$$

$$J_{m,n}^{g_D} := \sum_{(m',n')\in\Lambda_8^{V,1}} I_{m',n'}^- T_{m',n',m,n}^{g_D}, \quad \forall (m,n)\in\Lambda_8, \quad J_{m,n}^{g_N} := \sum_{(m',n')\in\Lambda_8^{V,1}} I_{m',n'}^- T_{m',n',m,n}^{g_N}, \quad \forall (m,n)\in\Lambda_7,$$

$$I_{m,n}^- := \sum_{(k,\ell)\in d_{i,j}^-} C_{k,\ell} G_{7,m,n}^V((v_0+k)h, (w_0+\ell)h), \quad \forall (m,n)\in\Lambda_8^{V,1}.$$

Moreover, the maximum accuracy order of a compact finite difference stencil for $\Delta u + \kappa^2 u = f$ *at an interior irregular point* (x_i, y_j) *is seven.*

Proof. The proof closely follows from the proof of [49, Theorem 2.4]. □

3.2 Numerical experiments

In this section, we let $\Omega = (l_1, l_2)^2$. For a given $J \in \mathbb{N}_0$, we define $h := (l_2 - l_1)/N_1$ with $N_1 := 2^J$. Recall the definition of (x_i, y_j) in (3.4). Let $u(x, y)$ be the exact solution of (3.1) and $(u_h)_{i,j}$ be the numerical solution at (x_i, y_j) using the mesh size h. We shall evaluate our proposed finite difference scheme in the 2-norm by the relative error $\frac{\|u_h - u\|_2}{\|u\|_2}$ if the exact solution u is available, and by the error $\|u_h - u_{h/2}\|_2$ if the exact solution is not known, where

$$\|u_h - u\|_2^2 := h^2 \sum_{i=0}^{N_1}\sum_{j=0}^{N_1} \left((u_h)_{i,j} - u(x_i, y_j)\right)^2, \qquad \|u_h - u_{h/2}\|_2^2 := h^2 \sum_{i=0}^{N_1}\sum_{j=0}^{N_1} \left((u_h)_{i,j} - (u_{h/2})_{2i,2j}\right)^2.$$

In the following numerical experiments, '[28]', '[120]' and '[125]' correspond to the sixth order compact FDMs proposed in [28], [120] and [125] respectively. 'Proposed' corresponds to the sixth order compact FDM with reduced pollution effect in Section 3.1 of this paper.

Recall that $\frac{2\pi}{\kappa h}$ corresponds to the number of points per wavelength.

3.2.1 Numerical examples with no interfaces

We provide four numerical experiments here.

Example 3.1. Consider the problem (3.1) in $\Omega = (0,1)^2$ with $f = 0$ and all Dirichlet boundary conditions such that the boundary data $g_1, \dots, g_4$ are picked such that the exact solution $u(x, y, \theta) = \exp(\mathrm{i}\kappa(\cos(\theta)x + \sin(\theta)y))$ is the plane wave with the angle θ. We define the following average error for plane wave solutions along all different angles θ by

$$\frac{\|u_h - u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}} := \frac{1}{N_3} \sum_{k=0}^{N_3-1} \sqrt{\frac{\sum_{i=0}^{N_1} \sum_{j=0}^{N_1} ((u_h)_{i,j,k} - u(x_i, y_j, \theta_k))^2}{\sum_{i=0}^{N_1} \sum_{j=0}^{N_1} (u(x_i, y_j, \theta_k))^2}},$$

where $\theta_k = kh_\theta, h_\theta = 2\pi/N_3$ for $N_3 \in \mathbb{N}_0$, and $(u_h)_{i,j,k}$ is the value of the numerical solution u_h at the grid point (x_i, y_j) with a plane wave angle θ_k. See Table 3.1 for numerical results.

	$\kappa = 50, N_3 = 50$					$\kappa = 150, N_3 = 30$					$\kappa = 450, N_3 = 30$				
	[28]	Proposed				[28]	Proposed				[28]	Proposed			
J	$\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$	$\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$	order	$\frac{2\pi}{\kappa h}$	r	$\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$	$\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$	order	$\frac{2\pi}{\kappa h}$	r	$\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$	$\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$	order	$\frac{2\pi}{\kappa h}$	r
4	9.83E+0	4.87E-01		2.0	**20.2**										
5	1.57E-02	1.01E-03	8.9	4.0	**15.5**										
6	5.01E-05	1.20E-05	6.4	8.0	**4.19**	3.67E+0	6.25E-02		2.7	**58.7**					
7	2.35E-07	1.77E-07	6.1	16.1	**1.33**	6.04E-03	6.82E-04	6.5	5.4	**8.85**					
8	2.78E-09	2.72E-09	6.0	32.2	**1.02**	2.56E-05	9.25E-06	6.2	10.7	**2.77**	1.26E+0	5.43E-02		3.6	**23.1**
9						1.78E-07	1.40E-07	6.0	21.4	**1.27**	4.72E-03	7.83E-04	6.1	7.1	**6.03**
10											2.25E-05	1.13E-05	6.1	14.3	**1.99**
11											1.85E-07	1.75E-07	6.0	28.6	**1.06**

Table 3.1: Numerical results for Example 3.1 with $h = 1/2^J$. The ratio r is equal to $\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$ of [28] divided by $\frac{\|u_h-u\|_{2,\mathsf{w}}}{\|u\|_{2,\mathsf{w}}}$ of our proposed method. In other words, for the same mesh size h with $h = 2^{-J}$, the error of [28] is r times larger than that of our proposed method.

Example 3.2. Consider the problem (3.1) in $\Omega = (0,1)^2$ with the boundary conditions

$$u(0, y) = g_1, \quad \text{and} \quad u(1, y) = g_2 \quad \text{for} \quad y \in (0,1),$$
$$u(x, 0) = g_3, \quad \text{and} \quad u_y(x, 1) - \mathrm{i}\kappa u(x, 1) = 0 \quad \text{for} \quad x \in (0,1),$$

where $g_1, \dots, g_4$ and f are chosen such that the exact solution $u = (y-1)\cos(\alpha x)\sin(\beta(y-1))$ with $\alpha, \beta \in \mathbb{R}$. See Table 3.2 for numerical results for various choices of α and β.

		$\alpha=50,\ \beta=290$					$\alpha=100,\ \beta=275$					$\alpha=150,\ \beta=255$				
		[120]	[125]	Proposed			[120]	[125]	Proposed			[120]	[125]	Proposed		
J	$\frac{2\pi}{\kappa h}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	r_1	r_2	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	r_1	r_2	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	r_1	r_2
7	2.7	1.1E+0	9.8E-02	3.8E-02	**29**	**2.6**	2.4E+0	2.1E-01	4.4E-02	**54**	**4.6**	4.4E+0	1.2E-01	5.8E-02	**77**	**2.1**
8	5.4	8.6E-03	6.1E-04	1.3E-04	**65**	**4.6**	1.2E-02	1.3E-03	3.1E-04	**40**	**4.4**	1.7E-02	8.3E-04	1.3E-04	**134**	**6.5**
9	10.7	1.2E-04	8.4E-06	2.8E-06	**43**	**3.0**	1.7E-04	1.8E-05	5.7E-06	**30**	**3.2**	2.4E-04	1.1E-05	2.0E-06	**121**	**5.7**
10	21.4	1.8E-06	1.2E-07	4.6E-08	**39**	**2.6**	2.6E-06	2.7E-07	9.2E-08	**28**	**2.9**	3.7E-06	1.7E-07	3.3E-08	**114**	**5.1**
		$\alpha=200,\ \beta=200$					$\alpha=250,\ \beta=160$					$\alpha=290,\ \beta=50$				
		[120]	[125]	Proposed			[120]	[125]	Proposed			[120]	[125]	Proposed		
J	$\frac{2\pi}{\kappa h}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	r_1	r_2	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	r_1	r_2	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	r_1	r_2
7	2.7	1.1E+0	1.3E-01	1.4E-01	**8**	**0.9**	6.0E+0	1.8E-01	4.8E-02	**125**	**3.7**	8.9E+0	1.3E-01	5.5E-02	**162**	**2.4**
8	5.4	7.5E-03	9.7E-04	3.8E-04	**20**	**2.6**	4.0E-02	1.1E-03	8.1E-05	**492**	**14.1**	9.8E-03	7.4E-04	1.5E-04	**66**	**4.9**
9	10.7	1.1E-04	1.3E-05	3.4E-06	**33**	**3.9**	5.6E-04	1.6E-05	2.1E-06	**264**	**7.6**	1.5E-04	1.0E-05	1.6E-06	**92**	**6.2**
10	21.4	1.7E-06	2.0E-07	4.5E-08	**38**	**4.4**	8.6E-06	2.3E-07	3.7E-08	**234**	**6.3**	2.3E-06	1.5E-07	2.3E-08	**101**	**6.4**

Table 3.2: Numerical results of Example 3.2 with $h = 1/2^J$ and $\kappa = 300$. The ratio r_1 is equal to $\frac{\|u_h-u\|_2}{\|u\|_2}$ of [120] divided by $\frac{\|u_h-u\|_2}{\|u\|_2}$ of our proposed method and the ratio r_2 is equal to $\frac{\|u_h-u\|_2}{\|u\|_2}$ of [125] divided by $\frac{\|u_h-u\|_2}{\|u\|_2}$ of our proposed method. In other words, for the same grid size h with $h = 2^{-J}$, the errors of [120] and [125] are r_1 and r_2 times larger than those of our proposed method, respectively.

Example 3.3. Consider the problem (3.1) in $\Omega = (0,1)^2$ with boundary conditions in (3.23). I.e., $\mathcal{B}_1 u = \frac{\partial u}{\partial \nu} - \mathrm{i}\kappa u = g_1$ on Γ_1, $\mathcal{B}_2 u = u = g_2$ on Γ_2, $\mathcal{B}_3 u = \frac{\partial u}{\partial \nu} = g_3$ on Γ_3 and $\mathcal{B}_4 u = \frac{\partial u}{\partial \nu} - \mathrm{i}\kappa u = g_4$ on Γ_4, where $g_1, \ldots, g_4$ and f are chosen such that the exact solution $u = \sin(\alpha x + \beta y)$ with $\alpha, \beta \in \mathbb{R}$. See Section 3.2.1 for numerical results for various choices of α and β.

	$\kappa=450,\ \alpha=400,\ \beta=200$					$\kappa=600,\ \alpha=300,\ \beta=500$				
J	$\frac{2\pi}{\kappa h}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	order	$\|u_h-u_{h/2}\|_2$	order	$\frac{2\pi}{\kappa h}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	order	$\|u_h-u_{h/2}\|_2$	order
7	1.79	1.3753E+01		9.8073E+00		1.34	9.0200E+01		6.4272E+01	
8	3.57	1.7358E-02	9.630	1.2212E-02	9.649	2.68	9.4259E-02	9.902	6.6801E-02	9.910
9	7.15	1.6528E-04	6.715	1.1540E-04	6.725	5.36	2.7428E-04	8.425	1.9430E-04	8.425
10	14.30	2.4370E-06	6.084	1.6971E-06	6.087	10.72	1.7971E-06	7.254	1.2453E-06	7.286
11	28.60	3.9410E-08	5.950			21.45	4.5869E-08	5.292		

Table 3.3: Numerical results of Example 3.3 with $h = 1/2^J$ using our proposed method.

Example 3.4. Consider the problem (3.1) in $\Omega = (0,1)^2$ with boundary conditions in (3.23), where $f(x,y) = \kappa^2 \sin(8x)\cos(6y)$, $g_1 = \sin(5y)$, $g_2 = 0$, $g_3 = (x-1)\sin(4x)$, and $g_4 = \cos(5x)$. Note that the exact solution u is unknown in this example. See Section 3.2.1 and Fig. 3.2 for numerical results.

	$\kappa = 200$				$\kappa = 400$				$\kappa = 800$			
J	$\frac{2\pi}{\kappa h}$	$\|u_h - u_{h/2}\|_2$	order	$\|u_h\|_2$	$\frac{2\pi}{\kappa h}$	$\|u_h - u_{h/2}\|_2$	order	$\|u_h\|_2$	$\frac{2\pi}{\kappa h}$	$\|u_h - u_{h/2}\|_2$	order	$\|u_h\|_2$
6	2.01	8.776E-01		5.81E-01								
7	4.02	3.716E-03	7.88	9.84E-01	2.01	7.936E-01		5.28E-01				
8	8.04	4.430E-05	6.39	9.81E-01	4.02	7.410E-03	6.74	9.76E-01	2.01	8.453E-01		5.08E-01
9	16.08			9.80E-01	8.04	8.579E-05	6.43	9.75E-01	4.02	1.486E-02	5.83	9.70E-01
10					16.08			9.74E-01	8.04	1.715E-04	6.44	9.70E-01
11									16.08			9.69E-01

Table 3.4: Numerical results of Example 3.4 with $h = 1/2^J$ using our proposed method.

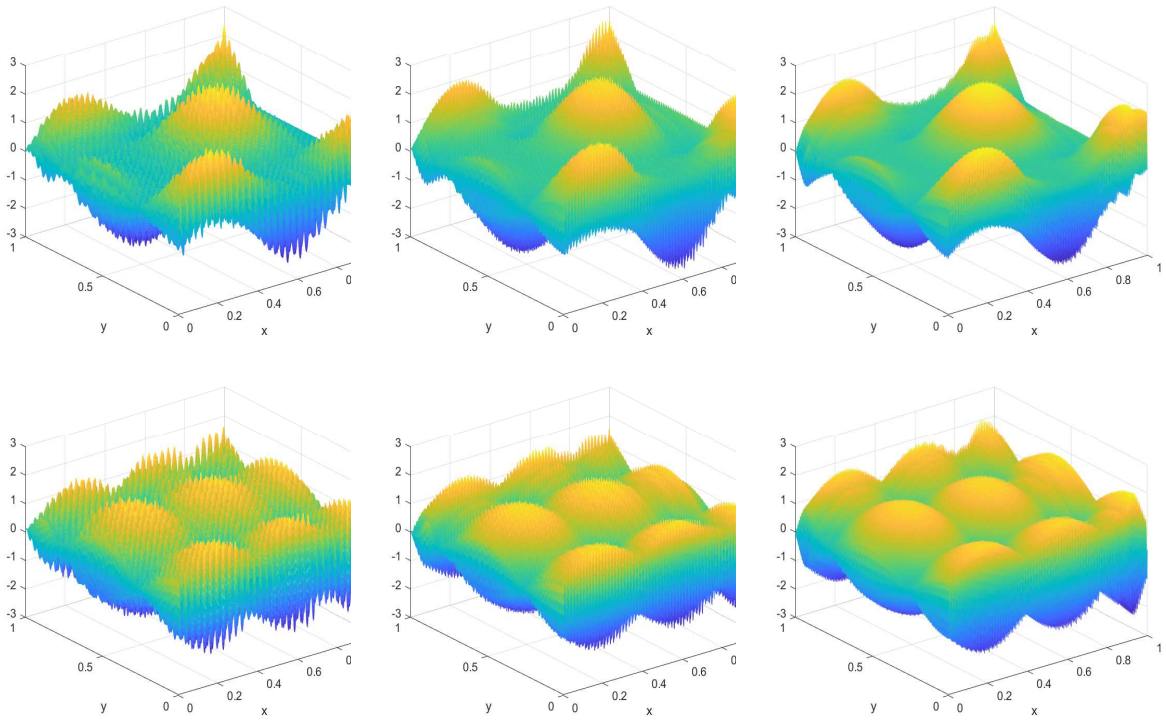


Figure 3.2: First row: the real part of u_h in Example 3.4, where $\kappa = 200$ and $h = 1/2^9$ (left), $\kappa = 400$ and $h = 1/2^{10}$ (middle), $\kappa = 800$ and $h = 1/2^{11}$ (right). Second row: the imaginary part of u_h in Example 3.4, where $\kappa = 200$ and $h = 1/2^9$ (left), $\kappa = 400$ and $h = 1/2^{10}$ (middle), $\kappa = 800$ and $h = 1/2^{11}$ (right).

3.2.2 Numerical examples with interfaces

We provide three numerical experiments here.

Example 3.5. Consider the problem (3.1) in $\Omega = (-3/2, 3/2)^2$ with boundary conditions in (3.23), where $\kappa = 100$, $\Gamma_I := \{(x, y) \in \Omega \ : \ \psi(x, y) = 0\}$ with $\psi(x, y) = y^2/2 + x^2/(1 + x^2) - 1/2$ (see Fig. 3.3 (left)), $g_D = -1$, and $g_N = 0$. The boundary data $g_1, \ldots, g_4$ and $f_\pm$ are chosen such that the exact solution u is given by $u_+ = u\chi_{\Omega^+} = \cos(50x)\cos(80y)$ and $u_- = u\chi_{\Omega^-} = \cos(50x)\cos(80y) + 1$. See Table 3.5 for numerical results.

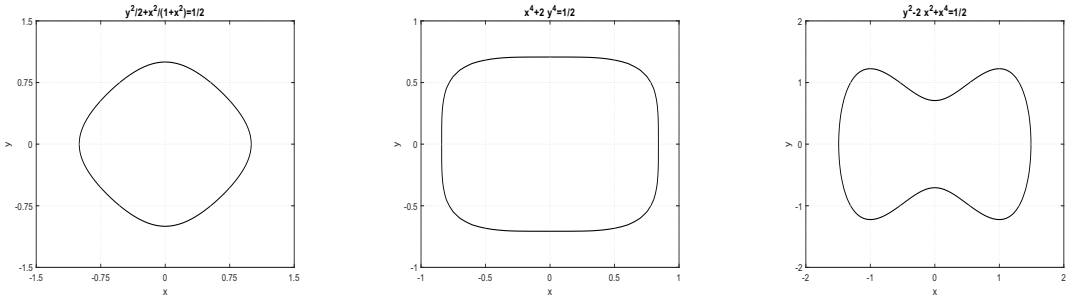


Figure 3.3: $y^2/2 + x^2/(1 + x^2) = 1/2$ (left), $x^4 + 2y^4 = 1/2$ (middle), and $y^2 - 2x^2 + x^4 = 1/2$ (right).

Example 3.6. Consider the problem (3.1) in $\Omega = (-1, 1)^2$ with boundary conditions in (3.23), where $\kappa = 300$, $\Gamma_I := \{(x, y) \in \Omega \ : \ \psi(x, y) = 0\}$ with $\psi(x, y) = x^4 + 2y^4 - 1/2$ (see Fig. 3.3 (middle)), $f_+ = 75^2 \sin(3(x+y))$, $f_- = 75^2 \cos(4x)\cos(3y)$, $g_D = \sin(2\pi x)\sin(2\pi y) + 3$, and $g_N = \cos(2\pi x)\cos(2\pi y)$. The following boundary data are given by $g_1 = e^y + e^{-y}$, $g_2 = 0$, $g_3 = (x - 1)e^x$, and $g_4 = \sin(2x)$. Note that the exact solution u is unknown in this example. See Table 3.5 for numerical results.

Example 3.7. Consider the problem (3.1) in $\Omega = (-2, 2)^2$ with boundary conditions in (3.23), where $\kappa = 150$, $\Gamma_I := \{(x, y) \in \Omega \ : \ \psi(x, y) = 0\}$ with $\psi(x, y) = y^2 - 2x^2 + x^4 - 1/2$ (see Fig. 3.3 (right)), $f_+ = \sin(5(x - y))$, $f_- = 10^4 \sin(5x)\sin(5y)$, $g_D = \sin(2\pi(x - y))$, and $g_N = \cos(2\pi(x + y))$. The following boundary data are given by $g_1 = \cos(y)\sin(y)$, $g_2 = 0$, $g_3 = \sin(2x - 4)$, and $g_4 = e^x \sin(x)$. Note that the exact solution u is unknown in this example. See Table 3.5 for numerical results.

3.3 Proofs of Theorems 3.2 to 3.5

In this section, we prove the main results stated in Section 3.1. The idea of proofs is to first construct all possible compact stencils with the maximum accuracy order and then to

	Example 3.5 with $h = \frac{3}{2^J}$					Example 3.6 with $h = \frac{2}{2^J}$				Example 3.7 with $h = \frac{4}{2^J}$			
J	$\frac{2\pi}{h\kappa}$	$\frac{\|u_h-u\|_2}{\|u\|_2}$	order	$\|u_h - u_{h/2}\|_2$	order	$\frac{2\pi}{h\kappa}$	$\|u_h - u_{h/2}\|_2$	order	$\|u_{h/2}\|_2$	$\frac{2\pi}{h\kappa}$	$\|u_h - u_{h/2}\|_2$	order	$\|u_{h/2}\|_2$
7	2.7	1.28E+00		2.90E+00									
8	5.4	2.44E-03	9.0	5.51E-03	9.0	2.7	1.06E+01		7.039	2.7	8.19E+00		3.467
9	10.7	5.82E-06	8.7	1.31E-05	8.7	5.4	1.49E-02	9.5	7.037	5.4	7.96E-03	10.0	3.469
10	21.4	3.98E-08	7.2	9.27E-08	7.1	10.7	1.69E-04	6.5	7.035	10.7	7.66E-05	6.7	3.468

Table 3.5: Numerical results of Examples 3.5 to 3.7 with $h = (l_2 - l_1)/2^J$ using our proposed method.

minimize the average truncation error of plane waves over the free parameters of stencils to reduce pollution effect.

Proof of Theorem 3.2. We first find all stencil coefficients $\{C_{k,\ell}\}_{k,\ell\in\{-1,0,1\}}$ and $\{C_{f,m,n}\}_{(m,n)\in\Lambda_{M_f-1}}$ such that

$$\sum_{k=-1}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}u(x_i+kh, y_j+\ell h) = \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)}C_{f,m,n} + \mathscr{O}(h^{M+2}), \qquad h\to 0,$$

for some $M, M_f \in \mathbb{N}_0$ and $M_f \geqslant M$. Afterwards, we set the remaining free parameters by minimizing the average truncation error of plane waves. Approximating $u(x_i+kh, y_j+\ell h)$ as in (3.15), we have

$$\sum_{(m,n)\in\Lambda^{V,1}_{M+1}} u^{(m,n)}I_{m,n} + \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)}\left(J_{m,n} - C_{f,m,n}\right) = \mathcal{O}(h^{M+2}), \quad h\to 0, \tag{3.30}$$

where we define

$$I_{m,n} := \sum_{k=-1}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}G^V_{M,m,n}(kh,\ell h), \quad \text{and} \quad J_{m,n} := \sum_{k=-1}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}Q^V_{M_f,m,n}(kh,\ell h). \tag{3.31}$$

Solving (3.30) is equivalent to solving

$$I_{m,n} = \mathscr{O}(h^{M+2}), \qquad h\to 0, \ \text{ for all } \ (m,n)\in\Lambda^{V,1}_{M+1}, \tag{3.32}$$

$$C_{f,m,n} = J_{m,n} + \mathscr{O}(h^{M+2}), \qquad h\to 0, \ \text{ for all } \ (m,n)\in\Lambda_{M_f-1}. \tag{3.33}$$

We set $C_{k,\ell} := \sum_{j=0}^{M+1} c_{k,\ell,j}(\kappa h)^j$, where $c_{k,\ell,j}\in\mathbb{R}$ for all $k,\ell\in\{-1,0,1\}$. Furthermore, we let $C_{-1,-1} = C_{-1,1} = C_{1,-1} = C_{1,1}$ and $C_{-1,0} = C_{0,-1} = C_{0,1} = C_{1,0}$ for symmetry. By calculation, we find that $M = 6$ is the maximum positive integer such that the linear system (3.32) has a non-trivial solution. All such non-trivial solutions for $M = 6$ can be uniquely

written (up to a constant multiple) as

$$\begin{aligned}
C_{1,1} &= c_9(\kappa h)^7 + c_3(\kappa h)^6 + c_2(\kappa h)^5 + c_1(\kappa h)^4 + (-12c_2 + c_4 - 6c_6 + 24c_{10} + 6c_{11} + 24c_9)(\kappa h)^3 + (1/15 \\
&\quad + 4c_1 + 2c_5 - 8c_7 - 2c_8 - 8c_3)(\kappa h)^2 + (-240c_2 + 15c_4 - 120c_6 + 480c_{10} + 120c_{11} + 480c_9)(\kappa h) + 1, \\
C_{1,0} &= c_{10}(\kappa h)^7 + c_7(\kappa h)^6 + c_6(\kappa h)^5 + c_5(\kappa h)^4 + c_4(\kappa h)^3 + (1/15 + 16c_1 + 8c_5 - 32c_7 - 8c_8 - 32c_3)(\kappa h)^2 \\
&\quad + (-960c_2 + 60c_4 - 480c_6 + 1920c_{10} + 480c_{11} + 1920c_9)(\kappa h) + 4, \\
C_{0,0} &= c_{11}(\kappa h)^7 + c_8(\kappa h)^6 + (92c_2 - (9/2)c_4 + 44c_6 - 192c_{10} - 48c_{11} - 192c_9)(\kappa h)^5 + (-3/10 + 20c_1 + 8c_5 \\
&\quad - 48c_7 - 12c_8 - 48c_3)(\kappa h)^4 + (-1392c_2 + 82c_4 - 696c_6 + 2784c_{10} + 696c_{11} + 2784c_9)(\kappa h)^3 \\
&\quad + (82/15 - 80c_1 - 40c_5 + 160c_7 + 40c_8 + 160c_3)(\kappa h)^2 + (4800c_2 - 300c_4 + 2400c_6 - 9600c_{10} \\
&\quad - 2400c_{11} - 9600c_9)(\kappa h) - 20,
\end{aligned} \tag{3.34}$$

where $c_i \in \mathbb{R}$ for $i = 1, \ldots, 11$ are free parameters. Note that any interior symmetric compact stencil has accuracy order 6 if and only if the 7th-degree Taylor polynomials of the stencil coefficients are given by (3.34). Choosing $M_f = 7$ in (3.31) and (3.33) yields the right-hand side of (3.18).

Next, consider a general compact stencil $\{C^{\mathsf{w}}_{k,\ell}\}_{k,\ell\in\{-1,0,1\}}$ parameterized by $C^{\mathsf{w}}_{1,1}, C^{\mathsf{w}}_{1,0} \in \mathbb{R}$ satisfying

$$C^{\mathsf{w}}_{-1,-1} = C^{\mathsf{w}}_{-1,1} = C^{\mathsf{w}}_{1,-1} = C^{\mathsf{w}}_{1,1}, \quad C^{\mathsf{w}}_{-1,0} = C^{\mathsf{w}}_{0,-1} = C^{\mathsf{w}}_{0,1} = C^{\mathsf{w}}_{1,0}, \quad \text{and} \quad C^{\mathsf{w}}_{0,0} = -20,$$

where we normalized the stencil by $C^{\mathsf{w}}_{0,0} = -20$. Take a plane wave solution $u(x, y, \theta) := \exp(\mathsf{i}\kappa(\cos(\theta)x + \sin(\theta)y))$ for any $\theta \in [0, 2\pi)$. Clearly, we have $\Delta u + \kappa^2 u = 0$. Hence, the truncation error associated with the general compact stencil coefficients $\{C^{\mathsf{w}}_{k,\ell}\}_{k,\ell\in\{-1,0,1\}}$ at the grid point $(x_i, y_j) \notin \partial\Omega$ is $h^{-2}(T(\theta|\kappa h))_{x_i,y_j}$, where

$$(T(\theta|\kappa h))_{x_i,y_j} := \sum_{k=-1}^{1}\sum_{\ell=-1}^{1} C^{\mathsf{w}}_{k,\ell} \exp(\mathsf{i}\kappa(\cos(\theta)(x_i + kh) + \sin(\theta)(y_j + \ell h))).$$

Recall that $\frac{2\pi}{\kappa h}$ is the number of points per wavelength. Hence, it is reasonable to choose $\kappa h \in [1/4, 1]$. Without loss of generality, we let $(x_i, y_j) = (0, 0)$. Define $S := \{\frac{1}{4} + \frac{3s}{4000} : s = 0, \ldots, 1000\}$ and let

$$(\tilde{C}^{\mathsf{w}}_{1,1}(\kappa h), \tilde{C}^{\mathsf{w}}_{1,0}(\kappa h)) := \operatorname*{arg\,min}_{C^{\mathsf{w}}_{1,1}, C^{\mathsf{w}}_{1,0}\in\mathbb{R}} \int_0^{2\pi} |(T(\theta|\kappa h))_{0,0}|^2 d\theta, \quad \kappa h \in S. \tag{3.35}$$

We use the Simpson's 3/8 rule with 900 uniform sampling points to calculate $\int_0^{2\pi} |(T(\theta|\kappa h))_{0,0}|^2 d\theta$. Now, we link $C_{0,0}, C_{1,0}, C_{1,1}$ in (3.34) with $C^{\mathsf{w}}_{0,0}, \tilde{C}^{\mathsf{w}}_{1,0}(\kappa h), \tilde{C}^{\mathsf{w}}_{1,1}(\kappa h)$ in (3.35) for $\kappa h \in S$. To further simplify the presentation of our stencil coefficients, we set

$c_9 = c_{10} = c_{11} = 0$ in (3.34) so that the coefficients of the polynomials in (3.34) for degree 7 are zero. Because $C^{\mathsf{w}}_{0,0} = -20$ is our normalization, we determine the free parameters c_i for $i = 1, \dots, 8$ in (3.34) by considering the following least-square problem:

$$(\tilde{c}_1, \tilde{c}_2, \dots, \tilde{c}_8) := \underset{c_1,c_2,\dots,c_8\in\mathbb{R}}{\arg\min} \sum_{\kappa h\in S} |C_{1,1}(\kappa h) - \tilde{C}^{\mathsf{w}}_{1,1}(\kappa h) C_{0,0}(\kappa h)/(-20)|^2 + |C_{1,0}(\kappa h) - \tilde{C}^{\mathsf{w}}_{1,0}(\kappa h) C_{0,0}(\kappa h)/(-20)|^2.$$

For simplicity of presentation, we replace each above calculated coefficient $\tilde{c}_i$ with its approximated fractional form $[10^8\tilde{c}_i]/10^8$, where $[\cdot]$ is a rounding operation to the nearest integer. Plugging these approximated fractional forms into coefficients c_i for $i = 1, \dots, 8$ in (3.34), we obtain (3.19). □

Proof of Theorem 3.3. We only prove item (1). The proof of item (2) is very similar. We start by finding all stencil coefficients $\{C^{\mathcal{B}_1}_{k,\ell}\}_{k\in\{0,1\},\ell\in\{-1,0,1\}}$, $\{C^{\mathcal{B}_1}_{f,m,n}\}_{(m,n)\in\Lambda_{M_f-1}}$ and $\{C^{\mathcal{B}_1}_{g_1,n}\}_{n=0}^{M_{g_1}}$ such that

$$\sum_{k=0}^{1}\sum_{\ell=-1}^{1} C^{\mathcal{B}_1}_{k,\ell} u(x_i+kh, y_j+\ell h) = \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)} C^{\mathcal{B}_1}_{f,m,n} + \sum_{n=0}^{M_{g_1}} g_1^{(n)} C^{\mathcal{B}_1}_{g_1,n} + \mathscr{O}(h^{M+2}), \qquad h\to 0 \tag{3.36}$$

for some $M_f, M_{g_1}, M \in \mathbb{N}_0$, $M_f \geqslant M$ and $M_{g_1} \geqslant M$. Afterwards, we set the remaining free parameters by minimizing the average truncation error of plane waves.

Since $-u_x - \mathsf{i}\kappa u = g_1$ on Γ_1, we have $u^{(1,n)} = -\mathsf{i}\kappa u^{(0,n)} - g_1^{(n)}$ for all $n = 0, \dots, M_{g_1}$. By (3.15),

$$\begin{aligned}
u(x+x_i^*, y+y_j^*) &= \sum_{n=0}^{M+1} u^{(0,n)} G^V_{M,0,n}(x,y) + \sum_{n=0}^{M} u^{(1,n)} G^V_{M,1,n}(x,y) \\
&\quad + \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)} Q^V_{M_f,m,n}(x,y) + \mathscr{O}(h^{M+2}) \\
&= \sum_{n=0}^{M+1} u^{(0,n)} G^V_{M,0,n}(x,y) + \sum_{n=0}^{M_{g_1}} u^{(1,n)} G^V_{M_{g_1},1,n}(x,y) + \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)} Q^V_{M_f,m,n}(x,y) + \mathscr{O}(h^{M+2}) \\
&= \sum_{n=0}^{M+1} u^{(0,n)} G^V_{M,0,n}(x,y) - \sum_{n=0}^{M_{g_1}} \big(\mathsf{i}\kappa u^{(0,n)} + g_1^{(n)}\big) G^V_{M_{g_1},1,n}(x,y) \\
&\quad + \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)} Q^V_{M_f,m,n}(x,y) + \mathscr{O}(h^{M+2}) \\
&= u^{(0,M+1)} G^V_{M,0,M+1}(x,y) + \sum_{n=0}^{M} u^{(0,n)} \Big(G^V_{M,0,n}(x,y) - \mathsf{i}\kappa G^V_{M,1,n}(x,y)\Big) - \sum_{n=0}^{M_{g_1}} g_1^{(n)} G^V_{M_{g_1},1,n}(x,y) \\
&\quad + \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)} Q^V_{M_f,m,n}(x,y) + \mathscr{O}(h^{M+2}), \quad \text{for } x,y\in(-2h,2h).
\end{aligned}$$

Approximating $u(x_i + kh, y_j + \ell h)$ in (3.36), we have

$$\sum_{n=0}^{M+1} u^{(0,n)} I_n^{\mathcal{B}_1} + \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)} \left(J_{m,n}^{\mathcal{B}_1} - C_{f,m,n}^{\mathcal{B}_1}\right) + \sum_{n=0}^{M_{g_1}} g_1^{(n)} \left(K_n^{\mathcal{B}_1} - C_{g_1,n}^{\mathcal{B}_1}\right) = \mathscr{O}(h^{M+2}), \quad (3.37)$$

as $h \to 0$, where

$$\begin{aligned} I_n^{\mathcal{B}_1} &:= \sum_{k=0}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}^{\mathcal{B}_1} \left(G_{M,0,n}^{V}(kh,\ell h) - \mathsf{i}\kappa G_{M,1,n}^{V}(kh,\ell h)(1-\delta_{n,M+1})\right), \\ J_{m,n}^{\mathcal{B}_1} &:= \sum_{k=0}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}^{\mathcal{B}_1} Q_{M_f,m,n}^{V}(kh,\ell h), \quad K_n^{\mathcal{B}_1} := -\sum_{k=0}^{1}\sum_{\ell=-1}^{1} C_{k,\ell}^{\mathcal{B}_1} G_{M_{g_1},1,n}^{V}(kh,\ell h), \end{aligned} \quad (3.38)$$

$\delta_{a,a} = 1$, and $\delta_{a,b} = 0$ for $a \neq b$. We set $C_{k,\ell}^{\mathcal{B}_1} := \sum_{j=0}^{M+1}(c_{k,\ell,j} + \mathsf{i}d_{k,\ell,j})(\kappa h)^j$, where $c_{k,\ell,j}, d_{k,\ell,j} \in \mathbb{R}$ for all $k \in \{0,1\}$ and $\ell \in \{-1,0,1\}$. Furthermore, we let $C_{0,-1}^{\mathcal{B}_1} = C_{0,1}^{\mathcal{B}_1}$ and $C_{1,-1}^{\mathcal{B}_1} = C_{1,1}^{\mathcal{B}_1}$ for symmetry. By calculation, we find that $M = 5$ is the maximum positive integer such that the linear system of (3.37) has a non-trivial solution. To further simplify such a solution, we set coefficients associated with κh of degrees higher than 4 to zero; i.e., we now have polynomials of κh, whose highest degree is now 4. All such non-trivial solutions for $M = 5$ can be uniquely written (up to a constant multiple) as

$$\begin{aligned} C_{1,1}^{\mathcal{B}_1} &= (c_3 + \mathsf{i}c_7)(\kappa h)^4 + (c_2 + \mathsf{i}c_6)(\kappa h)^3 + 12(\mathsf{i}c_8 - (7\mathsf{i}/3)c_2 + (7\mathsf{i}/3)c_5 + (13\mathsf{i}/3)c_7 + (7/3)c_1 + (13/3)c_3 + c_4 \\ &\quad + (7/3)c_6 - 4/135)(\kappa h)^2 - 60(\mathsf{i}c_1 + 2\mathsf{i}c_3 + (\mathsf{i}/2)c_4 + \mathsf{i}c_6 - 4\mathsf{i}/225 - (1/2)c_8 + c_2 - c_5 - 2c_7)\kappa h + 1 \\ C_{0,1}^{\mathcal{B}_1} &= (c_1 + \mathsf{i}c_5)(\kappa h)^4 + 13(\mathsf{i}c_1 + (20\mathsf{i}/13)c_3 + (7\mathsf{i}/26)c_4 + (12\mathsf{i}/13)c_6 - 17\mathsf{i}/1170 - (7/26)c_8 + (12/13)c_2 - c_5 \\ &\quad - (20/13)c_7)(\kappa h)^3 + 18(\mathsf{i}c_8 - (22\mathsf{i}/9)c_2 + (22\mathsf{i}/9)c_5 + (40\mathsf{i}/9)c_7 + (22/9)c_1 + (40/9)c_3 + c_4 + (22/9)c_6 \\ &\quad - 11/324)(\kappa h)^2 - 120(\mathsf{i}c_1 + (2\mathsf{i})c_3 + (\mathsf{i}/2)c_4 + \mathsf{i}c_6 - 29\mathsf{i}/1800 - (1/2)c_8 + c_2 - c_5 - 2c_7)\kappa h + 2 \\ C_{1,0}^{\mathcal{B}_1} &= (c_4 + \mathsf{i}c_8)(\kappa h)^4 + 18(\mathsf{i}c_1 + (4\mathsf{i}/3)c_3 + (\mathsf{i}/6)c_4 + (8\mathsf{i}/9)c_6 - \mathsf{i}/90 - (1/6)c_8 + (8/9)c_2 - c_5 \\ &\quad - (4/3)c_7)(\kappa h)^3 + 36(\mathsf{i}c_8 - (22\mathsf{i}/9)c_2 + (22\mathsf{i}/9)c_5 + (40\mathsf{i}/9)c_7 + (22/9)c_1 + (40/9)c_3 + c_4 + (22/9)c_6 \\ &\quad - 49/1620)(\kappa h)^2 - 240(\mathsf{i}c_1 + (2\mathsf{i})c_3 + (\mathsf{i}/2)c_4 + \mathsf{i}c_6 - 29\mathsf{i}/1800 - (1/2)c_8 + c_2 - c_5 - 2c_7)\kappa h + 4 \\ C_{0,0}^{\mathcal{B}_1} &= -4(\mathsf{i}c_8 - (3\mathsf{i}/2)c_2 + (2\mathsf{i})c_5 + (7\mathsf{i}/2)c_7 + 2c_1 + (7/2)c_3 + c_4 + (3/2)c_6 - 1/80)(\kappa h)^4 - 80(\mathsf{i}c_1 + (2\mathsf{i})c_3 \\ &\quad + (\mathsf{i}/2)c_4 + (39\mathsf{i}/40)c_6 - 7\mathsf{i}/720 - (1/2)c_8 + (39/40)c_2 - c_5 - 2c_7)(\kappa h)^3 + 84(\mathsf{i}c_8 - (32\mathsf{i}/21)c_2 \\ &\quad + (32\mathsf{i}/21)c_5 + (74\mathsf{i}/21)c_7 + (32/21)c_1 + (74/21)c_3 + c_4 + (32/21)c_6 + 1/3780)(\kappa h)^2 \\ &\quad + 600(\mathsf{i}c_1 + (2\mathsf{i})c_3 + (\mathsf{i}/2)c_4 + \mathsf{i}c_6 - 29\mathsf{i}/4500 - (1/2)c_8 + c_2 - c_5 - 2c_7)\kappa h - 10, \end{aligned} \quad (3.39)$$

where each $c_i \in \mathbb{R}$ for $i = 1, \ldots, 8$ are free parameters. Choosing $M_f = M_{g_1} = 7$ in (3.37) and (3.38) yields the right-hand side of (3.20).

Next, consider a compact stencil $\{C_{k,\ell}^{\mathsf{w}}\}_{k\in\{0,1\},\ell\in\{-1,0,1\}}$ parameterized by $C_{1,1}^{\mathsf{w}}, C_{0,1}^{\mathsf{w}}, C_{1,0}^{\mathsf{w}} \in$

$\mathbb{C}$ with

$$C^{\mathsf{w}}_{1,-1} = C^{\mathsf{w}}_{1,1}, \quad C^{\mathsf{w}}_{0,-1} = C^{\mathsf{w}}_{0,1}, \quad \text{and} \quad C^{\mathsf{w}}_{0,0} = -10,$$

where we normalized the general stencil by $C^{\mathsf{w}}_{0,0} = -10$. Take a plane wave solution $u(x,y,\theta) := \exp(\mathsf{i}\kappa(\cos(\theta)x + \sin(\theta)y))$ for any $\theta \in [0, 2\pi)$. Clearly, we have $\Delta u + \kappa^2 u = 0$ and $-u_x - \mathsf{i}\kappa u = g_1 \neq 0$ on Γ_1, where g_1 and its derivatives are explicitly known by plugging the plane wave solution $u(x,y,\theta)$ into the boundary condition. Hence, the truncation error associated with the compact general stencil coefficients $\{C^{\mathsf{w}}_{k,\ell}\}_{k\in\{0,1\},\ell\in\{-1,0,1\}}$ at the grid point $(x_0, y_j) \in \Gamma_1$ is $h^{-1}(T(\theta|\kappa h))_{x_i,y_j}$, where

$$(T(\theta|\kappa h))_{x_0,y_j} := \sum_{k=0}^{1}\sum_{\ell=-1}^{1} C^{\mathsf{w}}_{k,\ell} \exp(\mathsf{i}\kappa(\cos(\theta)(x_0+kh) + \sin(\theta)(y_j+\ell h))) \\ + \sum_{n=0}^{7} g_1^{(n)} \sum_{k=0}^{1}\sum_{\ell=-1}^{1} C^{\mathsf{w}}_{k,\ell} G^{V}_{7,1,n}(kh, \ell h).$$

Without loss of generality, we let $(x_0, y_j) = (0,0)$. Afterwards, we follow a similar minimization procedure as in the proof of Theorem 3.2 to obtain the concrete stencils in Theorem 3.3. □

Proof of Theorem 3.4. We start by finding all stencil coefficients $\{C^{\mathcal{R}_1}_{k,\ell}\}_{k,\ell\in\{0,1\}}$, $\{C^{\mathcal{R}_1}_{f,m,n}\}_{(m,n)\in\Lambda_{M_f-1}}$, $\{C^{\mathcal{R}_1}_{g_1,n}\}_{n=0}^{M_{g_1}}$, and $\{C^{\mathcal{R}_1}_{g_3,n}\}_{n=0}^{M_{g_3}}$ such that

$$\sum_{k=0}^{1}\sum_{\ell=0}^{1} C^{\mathcal{R}_1}_{k,\ell} u(x_0+kh, y_0+\ell h) = \sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)} C^{\mathcal{R}_1}_{f,m,n} + \sum_{n=0}^{M_{g_1}} g_1^{(n)} C^{\mathcal{R}_1}_{g_1,n} + \sum_{n=0}^{M_{g_3}} g_3^{(n)} C^{\mathcal{R}_1}_{g_3,n} + \mathscr{O}(h^{M+2}), \tag{3.40}$$

$h \to 0$, for some $M, M_f, M_{g_1}, M_{g_3} \in \mathbb{N}_0$, $M_f \geqslant M$, $M_{g_1} \geqslant M$ and $M_{g_3} \geqslant M$. Afterwards, we set the remaining free parameters by minimizing the average truncation error of plane waves.

Note that we have

$$u^{(1,n)} = -\mathsf{i}\kappa u^{(0,n)} - g_1^{(n)} \quad \text{and} \quad u^{(m,1)} = -g_3^{(m)}, \quad \text{for all } m, n \in \mathbb{N}_0. \tag{3.41}$$

Let $C^{\mathcal{R}_1}_{k,\ell} := C^{\mathcal{R}_1,V}_{k,\ell} + C^{\mathcal{R}_1,H}_{k,\ell}$ for $k,\ell \in \{0,1\}$, where $C^{\mathcal{R}_1,V}_{k,\ell}$ and $C^{\mathcal{R}_1,H}_{k,\ell}$ are to be determined polynomials of h. Approximating $u(x_0+kh, y_0+\ell h)$ with (3.15), (3.16), and using (3.41), we have

$$\sum_{k=0}^{1}\sum_{\ell=0}^{1} (C^{\mathcal{R}_1,V}_{k,\ell} + C^{\mathcal{R}_1,H}_{k,\ell}) u(x_0+kh, y_0+\ell h) = \sum_{n=0}^{M+1} u^{(0,n)} I^{\mathcal{R}_1,V}_n + \sum_{m=0}^{M+1} u^{(m,0)} I^{\mathcal{R}_1,H}_m$$

$$+\sum_{(m,n)\in\Lambda_{M_f-1}} f^{(m,n)}J^{\mathcal{R}_1}_{m,n}+\sum_{n=0}^{M_{g_1}} g_1^{(n)}K^{\mathcal{R}_1,V}_n+\sum_{m=0}^{M_{g_3}} g_3^{(m)}K^{\mathcal{R}_1,H}_m+\mathscr{O}(h^{M+2}), \tag{3.42}$$

where

$$I^{\mathcal{R}_1,V}_n:=\sum_{k=0}^{1}\sum_{\ell=0}^{1}C^{\mathcal{R}_1,V}_{k,\ell}\left(G^V_{M,0,n}(kh,\ell h)-\mathsf{i}\kappa G^V_{M,1,n}(kh,\ell h)(1-\delta_{n,M+1})\right),$$

$$I^{\mathcal{R}_1,H}_m:=\sum_{k=0}^{1}\sum_{\ell=0}^{1}C^{\mathcal{R}_1,H}_{k,\ell}G^H_{M,m,0}(kh,\ell h),\quad J^{\mathcal{R}_1}_{m,n}:=\sum_{k=0}^{1}\sum_{\ell=0}^{1}(C^{\mathcal{R}_1,V}_{k,\ell}Q^V_{M_f,m,n}(kh,\ell h)+C^{\mathcal{R}_1,H}_{k,\ell}Q^H_{M_f,m,n}(kh,\ell h)),$$

$$K^{\mathcal{R}_1,V}_n:=-\sum_{k=0}^{1}\sum_{\ell=0}^{1}C^{\mathcal{R}_1,V}_{k,\ell}G^V_{M_{g_1},1,n}(kh,\ell h),\quad\text{and}\quad K^{\mathcal{R}_1,H}_m:=-\sum_{k=0}^{1}\sum_{\ell=0}^{1}C^{\mathcal{R}_1,H}_{k,\ell}G^H_{M_{g_3},m,1}(kh,\ell h).$$

By replacing the left-hand side of (3.40) with (3.42), replacing $u^{(m,0)}$ for $m=2,\dots,M+1$ with (3.9), using (3.41), and rearranging some terms, we obtain

$$\begin{aligned}
&u^{(0,0)}\left(I^{\mathcal{R}_1,V}_0+I^{\mathcal{R}_1,H}_0-\mathsf{i}\kappa I^{\mathcal{R}_1,H}_1+\sum_{p=1}^{\lfloor\frac{M+1}{2}\rfloor}(-1)^p\kappa^{2p}I^{\mathcal{R}_1,H}_{2p}+\mathsf{i}\sum_{p=1}^{\lfloor\frac{M}{2}\rfloor}(-1)^{p+1}\kappa^{2p+1}I^{\mathcal{R}_1,H}_{2p+1}\right)\\
&\quad+\sum_{\ell=0}^{\lfloor\frac{M}{2}\rfloor}u^{(0,2\ell+1)}I^{\mathcal{R}_1,V}_{2\ell+1}\\
&\quad+\sum_{\ell=1}^{\lfloor\frac{M}{2}\rfloor}u^{(0,2\ell)}\left(\sum_{p=\max\{\ell,1\}}^{\lfloor\frac{M+1}{2}\rfloor}(-1)^p\binom{p}{\ell}\kappa^{2(p-\ell)}I^{\mathcal{R}_1,H}_{2p}+\mathsf{i}\sum_{p=\max\{\ell,1\}}^{\lfloor\frac{M}{2}\rfloor}(-1)^{p+1}\binom{p}{\ell}\kappa^{2(p-\ell)+1}I^{\mathcal{R}_1,H}_{2p+1}+I^{\mathcal{R}_1,V}_{2\ell}\right)\\
&\quad+u^{(0,2\lfloor\frac{M+1}{2}\rfloor)}\left((-1)^{\lfloor\frac{M+1}{2}\rfloor}I^{\mathcal{R}_1,H}_{2\lfloor\frac{M+1}{2}\rfloor}+I^{\mathcal{R}_1,V}_{2\lfloor\frac{M+1}{2}\rfloor}\right)\left(1-\delta_{\lfloor\frac{M+1}{2}\rfloor,\lfloor\frac{M}{2}\rfloor}\right)\\
&\quad+\sum_{\ell=0}^{\lfloor\frac{M_{g_1}-1}{2}\rfloor}g_1^{(2\ell+1)}(K^{\mathcal{R}_1,V}_{2\ell+1}-C^{\mathcal{R}_1}_{g_1,2\ell+1})\\
&\quad+\sum_{\ell=0}^{\lfloor\frac{M_{g_1}}{2}\rfloor}g_1^{(2\ell)}\left(K^{\mathcal{R}_1,V}_{2\ell}+\sum_{p=\max\{\ell,1\}}^{\lfloor\frac{M_{g_1}}{2}\rfloor}(-1)^{p+1}\binom{p}{\ell}\kappa^{2(p-\ell)}I^{\mathcal{R}_1,H}_{2p+1}-I^{\mathcal{R}_1,H}_1\delta_{\ell,0}-C^{\mathcal{R}_1}_{g_1,2\ell}\right)\\
&\quad+\sum_{\ell=0}^{M_{g_3}}g_3^{(\ell)}(K^{\mathcal{R}_1,H}_\ell-C^{\mathcal{R}_1}_{g_3,\ell})+\sum_{j=0}^{\lfloor\frac{M_f}{2}-1\rfloor}\sum_{\ell=0}^{M_f-2j-2}f^{(\ell,2j+1)}(J^{\mathcal{R}_1}_{\ell,2j+1}-C^{\mathcal{R}_1}_{f,\ell,2j+1})\\
&\quad+\sum_{\gamma\in\{0,1\}}\sum_{\ell=0}^{\lfloor\frac{M_f+1-\gamma}{2}\rfloor-1}\sum_{j=0}^{\lfloor\frac{M_f+1-\gamma}{2}\rfloor-\ell-1}f^{(2\ell+\gamma,2j)}\left(\sum_{p=\max\{j+\ell+1,1\}}^{\lfloor\frac{M_f+1-\gamma}{2}\rfloor}\right.\\
&\quad\left.(-1)^{p-\ell-1}\binom{p-\ell-1}{j}\kappa^{2(p-\ell-j-1)}I^{\mathcal{R}_1,H}_{2p+\gamma}+J^{\mathcal{R}_1}_{2\ell+\gamma,2j}-C^{\mathcal{R}_1}_{f,2\ell+\gamma,2j}\right)=\mathscr{O}(h^{M+2}),\quad h\to 0.
\end{aligned}$$

We set $C^{\mathcal{R}_1,V}_{k,\ell}=\sum_{j=0}^{M+1}(a_{k,\ell,j}+\mathsf{i}b_{k,\ell,j})(\kappa h)^j$ and $C^{\mathcal{R}_1,H}_{k,\ell}=\sum_{j=0}^{M+1}(c_{k,\ell,j}+\mathsf{i}d_{k,\ell,j})(\kappa h)^j$, where

$a_{k,\ell,j}$, $b_{k,\ell,j}$, $c_{k,\ell,j}$, $d_{k,\ell,j} \in \mathbb{R}$ for all $k \in \{0,1\}$ and $\ell \in \{-1,0,1\}$. By calculation, $M = 5$ is the maximum positive integer such that the linear system, obtained by setting each coefficient of $u^{(0,n)}$ for $n = 0,\ldots,6$ to be $\mathscr{O}(h^7)$ as $h \to 0$, has a non-trivial solution. Afterwards, to further simplify such a solution, we can set remaining coefficients associated with $(\kappa h)^5$ or $(\kappa h)^6$ to zero.

By using the minimization procedure described in the proofs of Theorems 3.2 and 3.3, we can verify that $C_{0,1}^{\mathcal{R}_1,V} = C_{1,1}^{\mathcal{R}_1,V} = C_{0,0}^{\mathcal{R}_1,H} = C_{1,0}^{\mathcal{R}_1,H} = 0$, $C_{0,0}^{\mathcal{R}_1,V} = C_{0,0}^{\mathcal{R}_1}$, $C_{1,0}^{\mathcal{R}_1,V} = C_{1,0}^{\mathcal{R}_1}$, $C_{0,1}^{\mathcal{R}_1,H} = C_{0,1}^{\mathcal{R}_1}$, and $C_{1,1}^{\mathcal{R}_1,H} = C_{1,1}^{\mathcal{R}_1}$, where $\{C_{k,\ell}^{\mathcal{R}_1}\}_{k,\ell\in\{0,1\}}$ are defined in (3.25). Given these $\{C_{k,\ell}^{\mathcal{R}_1,V}\}_{k,\ell\in\{0,1\}}$ and $\{C_{k,\ell}^{\mathcal{R}_1,H}\}_{k,\ell\in\{0,1\}}$, we set $M_f = M_{g_1} = M_{g_3} = 7$ and plug them into the following relations

$$\begin{aligned}
&C_{g_1,2\ell}^{\mathcal{R}_1} = K_{2\ell}^{\mathcal{R}_1,V} + \sum_{p=\max\{\ell,1\}}^{\left\lfloor \frac{M_{g_1}}{2} \right\rfloor} (-1)^{p+1} \binom{p}{\ell} \kappa^{2(p-\ell)} I_{2p+1}^{\mathcal{R}_1,H} - I_1^{\mathcal{R}_1,H} \delta_{\ell,0}, \quad \ell = 0,\ldots,\left\lfloor \frac{M_{g_1}}{2} \right\rfloor,\\
&C_{g_1,2\ell+1}^{\mathcal{R}_1} = K_{2\ell+1}^{\mathcal{R}_1,V}, \quad \ell = 0,\ldots,\left\lfloor \frac{M_{g_1}-1}{2} \right\rfloor, \qquad C_{g_3,\ell}^{\mathcal{R}_1} = K_{\ell}^{\mathcal{R}_1,H}, \quad \ell = 0,\ldots,M_{g_3},\\
&C_{f,\ell,2j+1}^{\mathcal{R}_1} = J_{\ell,2j+1}^{\mathcal{R}_1}, \quad \ell = 0,\ldots,M_f - 2j - 2, j = 0,\ldots,\left\lfloor \frac{M_f}{2} - 1 \right\rfloor, \quad \text{and}\\
&C_{f,2\ell+\gamma,2j}^{\mathcal{R}_1} = \sum_{p=\max\{j+\ell+1,1\}}^{\left\lfloor \frac{M_f+1-\gamma}{2} \right\rfloor} (-1)^{p-\ell-1} \binom{p-\ell-1}{j} \kappa^{2(p-\ell-j-1)} I_{2p+\gamma}^{\mathcal{R}_1,H} + J_{2\ell+\gamma,2j}^{\mathcal{R}_1},
\end{aligned} \tag{3.43}$$

where $\gamma \in \{0,1\}$, $j = 0,\ldots,\left\lfloor \frac{M_f+1-\gamma}{2} \right\rfloor - \ell - 1$, and $\ell = 0,\ldots,\left\lfloor \frac{M_f+1-\gamma}{2} \right\rfloor - 1$. This completes the proof of Theorem 3.4. □

Proof of Theorem 3.5. The proof is almost identical to the proof of Theorem 3.4. Note that we need to replace $u^{(m,1)} = -g_3^{(m)}$ with $u^{(m,1)} = \mathsf{i}\kappa u^{(m,0)} + g_4^{(m)}$ for all $m \in \mathbb{N}_0$ in (3.41). □

Chapter 4

Sharp Wavenumber-explicit Stability Bounds for 2D Helmholtz Equations

We have considered the discretization aspects of the 2D Helmholtz equation on a rectangular domain in Chapter 3. In this chapter, we study the stability of its solution. In particular, we want to consider the following 2D Helmholtz equation:

$$\mathcal{L}u := \Delta u + \kappa^2 u = -f \quad \text{in} \quad \Omega := (0,1)^2 \tag{4.1}$$

with the following boundary conditions

$$\begin{aligned} &\mathcal{B}_1 u = g_1 \quad \text{on} \quad \Gamma_1 := (0,1)\times\{0\}, \qquad &&\mathcal{B}_3 u = g_3 \quad \text{on} \quad \Gamma_3 := (0,1)\times\{1\}, \\ &\mathcal{B}_2 u = g_2 \quad \text{on} \quad \Gamma_2 := \{1\}\times(0,1), \qquad &&\mathcal{B}_4 u = g_4 \quad \text{on} \quad \Gamma_4 := \{0\}\times(0,1), \end{aligned} \tag{4.2}$$

where $\kappa > 0$ is a constant wavenumber, $f \in L_2(\Omega)$ is the source term, and $g_j \in L_2(\Gamma_j)$ for $j = 1,\ldots,4$ are boundary data. The different boundary conditions are prescribed by boundary operators $\mathcal{B}_1,\ldots,\mathcal{B}_4$, which belong to one of the three boundary operators: $\mathbf{I}_d$ (i.e., $\mathbf{I}_d u = u$) for a Dirichlet boundary condition; $\frac{\partial}{\partial \boldsymbol{\nu}}$ for a Neumann boundary condition; or $\frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d$ for the impedance boundary condition, where $\boldsymbol{\nu}$ is the outward normal vector. We shall assume that at least one impedance boundary condition is present. Without loss of generality, we assume that the impedance boundary condition is always imposed on Γ_4, i.e., $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d$. More specifically, we are interested in the following boundary configurations

$$\mathcal{B}_1, \mathcal{B}_3 \in \{\mathbf{I}_d, \tfrac{\partial}{\partial \boldsymbol{\nu}}\}, \quad \mathcal{B}_2 \in \{\mathbf{I}_d, \tfrac{\partial}{\partial \boldsymbol{\nu}}, \tfrac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d\}, \quad \text{and} \quad \mathcal{B}_4 = \tfrac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d. \tag{4.3}$$

See Fig. 4.1 for the domain and boundary configurations of the 2D Helmholtz equation (4.1)–(4.3).

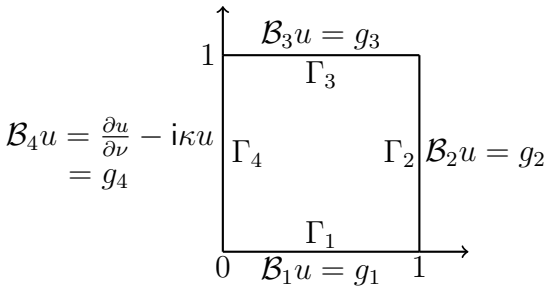


Figure 4.1: Boundary configuration in (4.2) and (4.3) for the 2D Helmholtz equation (4.1).

We prove the existence and uniqueness of the solution to the 2D Helmholtz equation (4.1)–(4.3), state several sharp wavenumber-explicit stability bounds, propose a lifting strategy, and provide several examples to illustrate the optimality of our stability bounds in Section 4.1. The technical proofs of several theorems are deferred to Section 4.2.

Results in this chapter are based on [76].

4.1 Main results on sharp wavenumber-explicit stability bounds

First, we study the existence and uniqueness of the solution to the Helmholtz equation (4.1)–(4.3) and then derive several relevant sharp wavenumber-explicit stability bounds.

Let Γ_D be the union of all boundaries on which the Dirichlet condition is imposed (i.e., $u = g_D$ on Γ_D), Γ_N be the union of all boundaries on which the Neumann condition is imposed (i.e., $\frac{\partial u}{\partial \boldsymbol{\nu}} = g_N$ on Γ_N), and Γ_R be the union of all boundaries on which the impedance boundary condition is imposed (i.e., $\frac{\partial u}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa u = g_R$ on Γ_R).

Define

$$\mathcal{H} := \{u \in H^1(\Omega) \ : \ u = 0 \quad \text{on} \quad \Gamma_D\},$$

where Γ_D is allowed to be an empty set. If $\Gamma_D = \emptyset$, then $\mathcal{H} = H^1(\Omega)$. For the homogeneous Dirichlet boundary condition $u = g_D = 0$ on Γ_D, the weak formulation of the 2D Helmholtz equation (4.1)–(4.3) is to find $u \in \mathcal{H}$ such that

$$a(u,v) := \int_\Omega (\nabla u \cdot \nabla \overline{v} - \kappa^2 u\overline{v}) - \mathsf{i}\kappa \int_{\Gamma_R} u\overline{v} = \int_\Omega f\overline{v} + \int_{\Gamma_R} g_R\overline{v} + \int_{\Gamma_N} g_N\overline{v} \qquad \forall\, v \in \mathcal{H}. \tag{4.4}$$

The existence and uniqueness of the solution to problem (4.4) can be proved by using the Fredholm alternative and the unique continuation principle [61, Theorem 2.1]. For the convenience of the reader, we shall include an explicit proof for our problem to make the

presentation self-contained.

Proposition 4.1. *There is a unique solution $u \in \mathcal{H}$ satisfying the problem (4.4).*

Proof. The sesquilinear form $a(\cdot,\cdot)$ is bounded, since $|a(u,u)| \leqslant \max(1,\kappa^2)\|u\|^2_{1,\Omega}$. Also, the Gårding's inequality [101, (2.7)] is satisfied, since $\Re(a(u,u)) = \|u\|^2_{1,\Omega} - (\kappa^2+1)\|u\|^2_{0,\Omega}$. We also know that $\mathcal{H}$ is compactly embedded in $L_2(\Omega)$. Hence, by the Fredholm alternative [101, Theorems 2.34 and 2.27], the solution to the variational problem (4.4) exists as long as we can show its uniqueness.

We now prove the uniqueness. Suppose that $f = g_R = g_N = 0$ in (4.4). We have to prove that the solution u must be 0. Then, recalling that $\Gamma_R \neq \emptyset$, we have $\Im(a(u,u)) = \kappa\|u\|_{0,\Gamma_R}$. Since κ is positive, we have $u = 0$ on Γ_R almost everywhere. Let $\tilde{\Omega}$ be an extended domain Ω such that $\tilde{\Omega} = (-\varepsilon, 1) \times (0,1)$ if $\Gamma_R = \Gamma_4$, $\tilde{\Omega} = (0, 1+\varepsilon) \times (0,1)$ if $\Gamma_R = \Gamma_2$, or $\tilde{\Omega} = (-\varepsilon, 1+\varepsilon) \times (0,1)$ if $\Gamma_R = \Gamma_2 \cup \Gamma_4$ for some $\varepsilon > 0$. Let $\tilde{u}$ be the function u with zero extension in $\tilde{\Omega}$. Note that $\tilde{u} \in H^1(\tilde{\Omega})$. Since $\langle \nabla u, \nabla v\rangle_\Omega - \kappa^2\langle u, v\rangle_\Omega = 0$ for all $v \in \mathcal{H}$, we have $\langle \nabla\tilde{u}, \nabla v\rangle_{\tilde{\Omega}} - \kappa^2\langle \tilde{u}, v\rangle_{\tilde{\Omega}} = 0$ for all $v \in H^1(\tilde{\Omega})$. Also noting that $\tilde{u} = 0$ in $\tilde{\Omega} \setminus \Omega$, by [61, Theorem 2.1] or [1, Theorem 1.1], we conclude that $u = 0$. The existence and uniqueness of the solution to the problem (4.4) for the Helmholtz equations have been proved. □

Furthermore, the existence of a unique solution still holds true even in the presence of inhomogeneous Dirichlet boundary conditions on Γ_D due to lifting.

4.1.1 Stability bounds for inhomogeneous vertical boundary conditions

To establish stability bounds for inhomogeneous boundary conditions only on the vertical sides, we assume the horizontal sides take homogenous boundary conditions such that

$$\mathcal{B}_1, \mathcal{B}_3 \in \{\mathbf{I}_d, \tfrac{\partial}{\partial \boldsymbol{\nu}}\} \quad \text{and} \quad \mathcal{B}_1 u = g_1 = 0 \ \text{ on } \ \Gamma_1, \quad \mathcal{B}_3 u = g_3 = 0 \ \text{ on } \ \Gamma_3. \tag{4.5}$$

Define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We shall use one of the following four orthonormal bases $\{Z_{j,n}\}_{n\in\mathbb{N}_0}$, $j = 1,\ldots,4$ in $L_2(\mathcal{I})$ with $\mathcal{I} := [0,1]$:

$$\begin{aligned} &Z_{1,n} := \sqrt{2}\sin(n\pi\cdot) \qquad \text{and} \qquad Z_{2,0} := 1, \quad Z_{2,n} := \sqrt{2}\cos(n\pi\cdot), \qquad n \in \mathbb{N}, \\ &Z_{3,n} := \sqrt{2}\sin((n+\tfrac{1}{2})\pi\cdot) \quad \text{and} \qquad Z_{4,n} := \sqrt{2}\cos((n+\tfrac{1}{2})\pi\cdot), \qquad n \in \mathbb{N}_0. \end{aligned} \tag{4.6}$$

To maintain a unified presentation, we often use $Z_{1,0} := 0$ instead of dropping $Z_{1,0}$. For $g \in L_2(\mathcal{I})$, we let $\tilde{g}$ be the function g with the zero extension outside the interval $\mathcal{I}$, and

define 2-periodic functions $G_1, \ldots, G_4$ whose values on $(-1, 1]$ are defined by

$$\begin{aligned} G_1(x) &:= \tilde{g}(x) - \tilde{g}(-x), & G_2(x) &:= \tilde{g}(x) + \tilde{g}(-x), \\ G_3(x) &:= (\tilde{g}(x) - \tilde{g}(-x))e^{\mathrm{i}x\pi/2}, & G_4(x) &:= (\tilde{g}(x) + \tilde{g}(-x))e^{\mathrm{i}x\pi/2}. \end{aligned} \tag{4.7}$$

For $g \in L_2(\mathcal{I})$ and $j = 1, \ldots, 4$, we have

$$g = \sum_{n\in\mathbb{N}_0} \widehat{g}(n) Z_{j,n} \quad \text{with} \quad \widehat{g}(n) := \int_0^1 g(x) Z_{j,n}(x) dx, \quad \forall n \in \mathbb{N}_0, \tag{4.8}$$

where we used the convention $Z_{1,0} := 0$. Let $Z'_{j,n}$ stand for the derivative of $Z_{j,n}$. It is also easy to observe that $\{Z'_{j,n}\}_{n\in\mathbb{N}_0}$ is an orthogonal system in $L_2(\mathcal{I})$ satisfying $\int_0^1 Z'_{j,m}(x) Z'_{j,n}(x) dx = 0$ as long as $m \neq n$. For $\{Z_{j,n}\}_{n\in\mathbb{N}_0}$ with $j = 1, 2$, we let $\{\sigma_n = n\pi\}_{n\in\mathbb{N}_0}$. For $\{Z_{j,n}\}_{n\in\mathbb{N}_0}$ with $j = 3, 4$, we let $\{\sigma_n = (n + \frac{1}{2})\pi\}_{n\in\mathbb{N}_0}$. We refer to $\{\sigma_n\}_{n\in\mathbb{N}_0}$ as eigenvalues, $\{Z_{j,n}\}_{n\in\mathbb{N}_0}$ as eigenfunctions, and $\{(\sigma_n^2, Z_{j,n})\}_{n\in\mathbb{N}_0}$ as eigenpairs. Due to the identities in (4.8), given eigenpairs $\{(\sigma_n^2, Z_{j,n})\}_{n\in\mathbb{N}_0}$ for some $j \in \{1, 2, 3, 4\}$ and $s \geqslant 0$, we say that $g \in \mathcal{Z}^s(\Gamma_\ell)$ for $\ell \in \{1, 2, 3, 4\}$, if $g \in L^2(\Gamma_\ell)$ and

$$\|g\|^2_{\mathcal{Z}^s(\Gamma_\ell)} := \sum_{n=0}^{\infty} |\widehat{g}(n)|^2 \sigma_n^{2s} < \infty \quad \text{with} \quad \widehat{g}(n) := \int_{\Gamma_\ell} g(x) Z_{j,n}(x) dx \quad \forall n \in \mathbb{N}_0. \tag{4.9}$$

Note that $\int_{\Gamma_\ell} g(x) Z_{j,n}(x) dx = \int_0^1 g(x) Z_{j,n}(x) dx$ for $\ell = 1, \ldots, 4$. Such a Hilbert space has been used in stability estimates of the Helmholtz equation; e.g, see [14, Section 2.2] and [106, Section 3].

In this subsection, let our eigenvalues be $\{\mu_n = n\pi\}_{n\in\mathbb{N}_0}$ or $\{\mu_n = (n + \frac{1}{2})\pi\}_{n\in\mathbb{N}_0}$, and our eigenfunctions be

$$Y_n(y) = \begin{cases} Z_{1,n}(y), & \text{if} \quad \mu_n = n\pi \neq 0 \text{ and } \mathcal{B}_1 = \mathcal{B}_3 = \mathbf{I}_d, \\ Z_{2,n}(y), & \text{if} \quad \mu_n = n\pi \neq 0 \text{ and } \mathcal{B}_1 = \mathcal{B}_3 = \frac{\partial}{\partial \boldsymbol{\nu}}, \\ Z_{3,n}(y), & \text{if} \quad \mu_n = (n + \frac{1}{2})\pi \text{ and } \mathcal{B}_1 = \mathbf{I}_d, \mathcal{B}_3 = \frac{\partial}{\partial \boldsymbol{\nu}}, \\ Z_{4,n}(y), & \text{if} \quad \mu_n = (n + \frac{1}{2})\pi \text{ and } \mathcal{B}_1 = \frac{\partial}{\partial \boldsymbol{\nu}}, \mathcal{B}_3 = \mathbf{I}_d, \end{cases} \quad \forall n \in \mathbb{N}_0, \tag{4.10}$$

which together give us $\{(\mu_n^2, Y_n)\}_{n\in\mathbb{N}_0}$ as our eigenpairs. We are now ready to state our first set of stability bounds. The oscillating part of the solution predominantly contributes to the stability bound. Hence, the main idea of the proof is to find a delicate upper bound for its norm. We achieve this by establishing several technical norm estimates. The proof of the following theorem is deferred to Section 4.2.

Theorem 4.2. *Consider the Helmholtz equation in (4.1)–(4.2). Assume that (4.5) holds, $\mathcal{B}_2 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}, \frac{\partial}{\partial \boldsymbol{\nu}} - i\kappa\mathbf{I}_d\}$ with $\mathcal{B}_2 u = g_2 = 0$ on Γ_2, and $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - i\kappa\mathbf{I}_d$ with $\mathcal{B}_4 u = g_4 \in L_2(\Gamma_4)$ on Γ_4. Then the unique solution u to the Helmholtz equation in (4.1)–(4.2) with the source term f vanishing satisfies*

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} \leqslant \sqrt{12}\max\{\kappa, 1\}\|g_4\|_{0,\Gamma_4}, \qquad \forall\, \kappa > 0. \tag{4.11}$$

The following example demonstrates how the stability bound in Theorem 4.2 is sharp in the sense that the right-hand side of (4.11) holds up to a constant multiple (independent of κ and g_4).

Example 4.1. In what follows, suppose that the conditions of Theorem 4.2 hold, that is, $f = 0$, $\mathcal{B}_1, \mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$, and $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d$ by (4.3) with the boundary data $g_1 = g_2 = g_3 = 0$ in (4.2). Note that the boundary data $g_4 = \mathcal{B}_4 u$ on Γ_4. We have three choices for the boundary operator $\mathcal{B}_2$. Let us consider the first case $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d$ with $\kappa^2 := \mu_n^2 + \pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n + \frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = \frac{1}{2\kappa\pi}\left(-\sin(\pi x)\kappa + \cos(\pi x)\pi\mathsf{i}\right) Y_n(y)$ be the exact solution, where $Y_n(y)$ takes one of the forms in (4.10). Because $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d$, we have $g_4 = \mathcal{B}_4 u = Y_n$ on Γ_4 and hence $\|g_4\|_{0,\Gamma_4} = 1$. The exact solution $u = \frac{1}{2\kappa\pi}\left(-\sin(\pi x)\kappa + \cos(\pi x)\pi\mathsf{i}\right) Y_n(y)$ satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{\sqrt{2}}{2}\kappa\sqrt{\tfrac{1}{\pi^2} + \tfrac{1}{k^2}} \geqslant \tfrac{\sqrt{2}}{2\pi}\kappa\|g_4\|_{0,\Gamma_4}$$

with $\kappa := \sqrt{\mu_n^2 + \pi^2}$ and $\mu_n \in \{n\pi, (n + \frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

Next, we consider the second case $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$ with $\kappa^2 := \mu_n^2 + \frac{1}{4}\pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n + \frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = -\frac{2}{\pi}\sin(\frac{\pi}{2}x)Y_n(y)$ be the exact solution, where $Y_n(y)$ takes one of the forms in (4.10). Because $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d$, $g_4 = \mathcal{B}_4 u = Y_n$ on Γ_4 and hence $\|g_4\|_{0,\Gamma_4} = 1$. The exact solution $u = -\frac{2}{\pi}\sin(\frac{\pi}{2}x)Y_n(y)$ satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{2\sqrt{2}}{\pi}\kappa\|g_4\|_{0,\Gamma_4}$$

with $\kappa := \sqrt{\mu_n^2 + \frac{1}{4}\pi^2}$ and $\mu_n \in \{n\pi, (n + \frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

Finally, we consider the third case $\mathcal{B}_2 = \mathbf{I}_d$ with $\kappa^2 := \mu_n^2 + \pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n + \frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = -\frac{1}{\pi}\sin(\pi x)Y_n(y)$ be the exact solution, where $Y_n(y)$ takes one of the forms in (4.10). Because $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathsf{i}\kappa\mathbf{I}_d$, $g_4 = \mathcal{B}_4 u = Y_n$ on Γ_4 and hence $\|g_4\|_{0,\Gamma_4} = 1$. The exact solution $u = -\frac{1}{\pi}\sin(\pi x)Y_n(y)$ satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{\sqrt{2}}{\pi}\kappa\|g_4\|_{0,\Gamma_4}$$

with $\kappa := \sqrt{\mu_n^2 + \pi^2}$ and $\mu_n \in \{n\pi, (n+\frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

A similar example demonstrating the sharpness of the stability bound (4.11) for the third case was also presented in [41].

Next, we present our second set of stability bounds, whose proof is deferred to Section 4.2, since it again involves several technical norm estimates. We recall that if $\mathcal{B}_j u = u = 0$ on Γ_j for $j = 1, 3$, then $\mathcal{Z}^{1/2}(\Gamma_2)$ can be identified with a subspace of $H^{1/2}(\Gamma_2)$, which after the zero extension belongs to the space $H^{1/2}(\partial\Omega)$; e.g, see [106, Section 3].

Theorem 4.3. *Consider the Helmholtz equation in (4.1)–(4.2) with the source term f vanishing. Assume that (4.5) holds, $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - i\kappa \mathbf{I}_d$ with $\mathcal{B}_4 u = g_4 = 0$ on Γ_4, and $\mathcal{B}_2 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}, \frac{\partial}{\partial \boldsymbol{\nu}} - i\kappa \mathbf{I}_d\}$ with $\mathcal{B}_2 u = g_2 \in L_2(\Gamma_2)$ on Γ_2. Then,*

(1) For $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}} - i\kappa \mathbf{I}_d$, the unique solution u to the Helmholtz equation in (4.1)–(4.2) satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa \|u\|_{0,\Omega} \leqslant \sqrt{12}\, \max\{\kappa, 1\} \|g_2\|_{0,\Gamma_2}, \qquad \forall\, \kappa > 0.$$

(2) For $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$, the unique solution u to the Helmholtz equation in (4.1)–(4.2) satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa \|u\|_{0,\Omega} \leqslant \sqrt{20}\, \max\{\kappa^2, 1\} \|g_2\|_{0,\Gamma_2}, \qquad \forall\, \kappa > 0. \tag{4.12}$$

(3) For $\mathcal{B}_2 = \mathbf{I}_d$ and $g_2 \in \mathcal{Z}^{1/2}(\Gamma_2)$, the unique solution u to the Helmholtz equation in (4.1)–(4.2) satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa \|u\|_{0,\Omega} \leqslant \sqrt{14}\, \big(\max\{\kappa^2, 1\} \|g_2\|_{0,\Gamma_2} + \max\{\kappa^{\frac{1}{2}}, 1\} \|g_2\|_{\mathcal{Z}^{1/2}(\Gamma_2)}\big), \qquad \forall\, \kappa > 0. \tag{4.13}$$

An example demonstrating the sharpness of the stability bound in item (1) of Theorem 4.3 can be recovered from the first case discussed in Example 4.1, where both vertical sides have the impedance boundary conditions with only the left hand side being inhomogeneous, by replacing x in the solution u with $1 - x$. This way the nonzero vertical boundary condition is on the right-hand side (i.e., Γ_2). The following example demonstrates how the stability bounds (4.12) and (4.13) are sharp in the sense that the right-hand sides of (4.12) and (4.13) hold up to a constant multiple (independent of κ and g_2).

Example 4.2. In what follows, suppose that the conditions of Theorem 4.3 hold, that is, $f = 0$, $\mathcal{B}_1, \mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$, and $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa \mathbf{I}_d$ by (4.3) with the boundary data $g_1 = 0, g_3 = 0, g_4 = 0$ in (4.2). Note that the boundary data $g_2 = \mathcal{B}_2 u$ on Γ_2. Let us consider the first

case $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$ with $\kappa^2 := \mu_n^2 + \frac{1}{4}\pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n+\frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = \frac{1}{\pi^2}\left(-2\pi\cos(\frac{\pi}{2}x) + 4\kappa\sin(\frac{\pi}{2}x)\mathsf{i}\right) Y_n(y)$ be the exact solution, where $Y_n(y)$ takes one of the forms in (4.10). Then $g_2 = \mathcal{B}_2 u = Y_n$ on Γ_2 and hence $\|g_2\|_{0,\Gamma_2} = 1$. The exact solution $u = \frac{1}{\pi^2}\left(-2\pi\cos(\frac{\pi}{2}x) + 4\kappa\sin(\frac{\pi}{2}x)\mathsf{i}\right) Y_n(y)$ satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{4\sqrt{2}}{\pi}\kappa^2\sqrt{\tfrac{1}{\pi^2} + \tfrac{1}{4\kappa^2}} \geqslant \tfrac{4\sqrt{2}}{\pi^2}\kappa^2\|g_2\|_{0,\Gamma_2},$$

where $\kappa := \sqrt{\mu_n^2 + \frac{1}{4}\pi^2}$ and $\mu_n \in \{n\pi, (n+\frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

Finally, we consider the case $\mathcal{B}_2 = \mathbf{I}_d$ with $\kappa^2 := \mu_n^2 + \pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n+\frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = \frac{1}{\pi}\left(-\pi\cos(\pi x) + \sin(\pi x)\kappa\mathsf{i}\right) Y_n(y)$ be the exact solution, where $Y_n(y)$ takes the form of (4.10). Then $g_2 = \mathcal{B}_2 u = Y_n$ on Γ_2 and hence $\|g_2\|_{0,\Gamma_2} = 1$. Also, $\|g_2\|_{\mathcal{Z}^{1/2}(\Gamma_2)} = \mu_n^{1/2}$. The exact solution $u = \frac{1}{\pi}\left(-\pi\cos(\pi x) + \sin(\pi x)\kappa\mathsf{i}\right) Y_n(y)$ satisfies

$$\begin{aligned}\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{\sqrt{2}}{\pi}\sqrt{\kappa^4 + \pi^2\kappa^2} &\geqslant \tfrac{1}{\pi}\left(\kappa^2 + \pi\kappa\right) \geqslant \tfrac{1}{\pi}\left(\kappa^2 + \pi\kappa^{1/2}\mu_n^{1/2}\right)\\ &\geqslant \tfrac{1}{\pi}(\kappa^2\|g_2\|_{0,\Gamma_2} + \kappa^{1/2}\|g_2\|_{\mathcal{Z}^{1/2}(\Gamma_2)}),\end{aligned}$$

where $\kappa := \sqrt{\mu_n^2 + \pi^2}$ and $\mu_n \in \{n\pi, (n+\frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

4.1.2 Stability bounds for non-vanishing source terms

We can derive a stability estimate for $f \in L_2(\Omega)$ by using the variational formulation (4.4) and the Rellich's identity [32, Proposition 2.1]. A part of this problem (i.e., $\Gamma_D = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ and $\Gamma_R := \Gamma_4$) was addressed in [41, Appendix]. The proof of the following result is deferred to Section 3.

Theorem 4.4. *Consider the Helmholtz equation (4.1)-(4.3). Assume that $g_j = 0$ on Γ_j for all $j = 1, \ldots, 4$ and $f \in L_2(\Omega)$.*

(1) For $\mathcal{B}_2 = \mathbf{I}_d$, if the unique solution u to (4.4) is in $H^2(\Omega)$, then

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} \leqslant \sqrt{30}\max\{\kappa^2, 1\}\|f\|_{0,\Omega}, \quad \forall \kappa > 0. \tag{4.14}$$

(2) For $\mathcal{B}_2 \in \{\frac{\partial}{\partial \boldsymbol{\nu}}, \frac{\partial}{\partial \boldsymbol{\nu}} - \imath\kappa\mathbf{I}_d\}$, if the unique solution u to (4.4) is in $H^2(\Omega)$, then

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} \leqslant \sqrt{542}\max\{\kappa^2, \kappa^{-1/2}\}\|f\|_{0,\Omega}, \quad \forall \kappa > 0. \tag{4.15}$$

The following example demonstrates how the stability bounds (4.14) and (4.15) are sharp in the sense that the right-hand sides of (4.14) and (4.15) hold up to a constant multiple

(independent of κ and f).

Example 4.3. In what follows, suppose that the conditions of Theorem 4.4 hold, that is, $f \in L_2(\Omega)$, $\mathcal{B}_1, \mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$, $\mathcal{B}_4 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa\mathbf{I}_d$ by (4.3) with the boundary data $g_j = 0$ on Γ_j for all $j = 1, \ldots, 4$. Let us consider the first case $\mathcal{B}_2 = \mathbf{I}_d$ with $\kappa^2 := \mu_n^2 + \pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n + \frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = \frac{1}{\pi^3}\left(\pi\cos(\pi x) + \pi - 2\sin(\pi x)\kappa\mathrm{i}\right) Y_n(y)$, where $Y_n(y)$ takes one of the forms in (4.10). Then $f = Y_n(y)$ and hence $\|f\|_{0,\Omega} = 1$. The exact solution $u = \frac{1}{\pi^3}\left(\pi\cos(\pi x) + \pi - 2\sin(\pi x)\kappa\mathrm{i}\right) Y_n(y)$ satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{\sqrt{2}}{\pi^3}\kappa^2\left(\sqrt{1 + \tfrac{3\pi^2}{4\kappa^2} - \tfrac{\pi^4}{2\kappa^4}} + \sqrt{1 + \tfrac{3\pi^2}{4\kappa^2}}\right) \geqslant \tfrac{2\sqrt{2}}{\pi^3}\kappa^2\|f\|_{0,\Omega},$$

where $\kappa := \sqrt{\mu_n^2 + \pi^2}$ and $\mu_n \in \{n\pi, (n + \frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

Next, we consider the second case $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$ with $\kappa^2 := \mu_n^2 + \frac{1}{4}\pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n + \frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = \frac{1}{\pi^3}\left(4\pi - 8\kappa\sin(\frac{\pi}{2}x)\mathrm{i}\right) Y_n(y)$, where $Y_n(y)$ takes one of the forms in (4.10). Then $f = Y_n(y)$ and hence $\|f\|_{0,\Omega} = 1$. The exact solution $u = \frac{1}{\pi^3}\left(4\pi - 8\kappa\sin(\frac{\pi}{2}x)\mathrm{i}\right) Y_n(y)$ satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{4\sqrt{2}}{\pi^3}\kappa^2\left(\sqrt{1 + \tfrac{\pi^2}{2\kappa^2} - \tfrac{\pi^4}{8\kappa^4}} + \sqrt{1 + \tfrac{\pi^2}{2\kappa^2}}\right) \geqslant \tfrac{8\sqrt{2}}{\pi^3}\kappa^2\|f\|_{0,\Omega},$$

where $\kappa := \sqrt{\mu_n^2 + \frac{1}{4}\pi^2}$ and $\mu_n \in \{n\pi, (n + \frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

Finally, we consider the third case $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa\mathbf{I}_d$ with $\kappa^2 := \mu_n^2 + \pi^2$, where $\mu_n = n\pi$ or $\mu_n = (n + \frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Let $u = \frac{1}{\pi^3}\left(\pi - \kappa\sin(\pi x)i\right) Y_n(y)$, where $Y_n(y)$ takes one of the forms in (4.10). Then $f = Y_n(y)$ and hence $\|f\|_{0,\Omega} = 1$. The exact solution $u = \frac{1}{\pi^3}\left(\pi - \kappa\sin(\pi x)\mathrm{i}\right) Y_n(y)$ satisfies

$$\|\nabla u\|_{0,\Omega} + \kappa\|u\|_{0,\Omega} = \tfrac{1}{\sqrt{2}\pi^3}\kappa^2\left(\sqrt{1 + \tfrac{2\pi^2}{\kappa^2} - \tfrac{2\pi^4}{\kappa^4}} + \sqrt{1 + \tfrac{2\pi^2}{\kappa^2}}\right) \geqslant \tfrac{\sqrt{2}}{\pi^3}\kappa^2\|f\|_{0,\Omega},$$

where $\kappa := \sqrt{\mu_n^2 + \pi^2}$ and $\mu_n \in \{n\pi, (n + \frac{1}{2})\pi\}$ for all $n \in \mathbb{N}$.

4.1.3 Stability bounds for inhomogeneous horizontal boundary conditions using a lifting technique

In this section, we discuss how under certain assumptions, we can transfer the inhomogeneous horizontal boundary data to the vertical boundary conditions. This procedure is well known as lifting in the literature. As we shall soon see, we are actually considering a particular instance of lifting, where our auxiliary functions do not affect the source term at all. Consider

the Helmholtz equation (4.1)–(4.3). Without loss of generality, let us assume that only one of the horizontal boundary conditions is inhomogeneous and it is on Γ_1. We can use the same method to handle the case where both horizontal boundary conditions are inhomogeneous. Our goal is thus to explicitly construct an auxiliary function $\tilde{u}$ satisfying

$$\begin{aligned} &\mathcal{L}\tilde{u} := \Delta\tilde{u} + \kappa^2\tilde{u} = 0 \quad \text{in} \quad \Omega := (0,1)^2 \quad \text{with} \\ &\mathcal{B}_1\tilde{u} = g_1 \ \text{ on } \ \Gamma_1, \quad \mathcal{B}_3\tilde{u} = 0 \ \text{ on } \ \Gamma_3. \end{aligned} \tag{4.16}$$

We shall impose some conditions on g_1 to ensure that the traces of $\tilde{u}$ belong to the appropriate function spaces so that we can go back to the situations discussed in Section 4.1.

Expanding g_1 in terms of certain eigenfunctions is a vital step for the construction of the above auxiliary solution. For eigenvalues $\{\tilde{\mu}_n = n\pi\}_{n\in\mathbb{N}_0}$, we can use either eigenfunctions $\{\tilde{X}_n = Z_{1,n}\}_{n\in\mathbb{N}_0}$ or $\{\tilde{X}_n = Z_{2,n}\}_{n\in\mathbb{N}_0}$. For eigenvalues $\{\tilde{\mu}_n = (n+\frac{1}{2})\pi\}_{n\in\mathbb{N}_0}$, we can use either eigenfunctions $\{\tilde{X}_n = Z_{3,n}\}_{n\in\mathbb{N}_0}$ or $\{\tilde{X}_n = Z_{4,n}\}_{n\in\mathbb{N}_0}$. We first discuss how to properly choose $\tilde{\mu}_n, n \in \mathbb{N}_0$. Define

$$\begin{aligned} d_0 &:= \operatorname{dist}(\kappa^2, \pi^2\mathbb{Z}) = \inf_{n\in\mathbb{Z}} |\kappa^2 - n\pi^2| \quad \text{and} \\ d_1 &:= \operatorname{dist}(\kappa^2, \pi^2(\tfrac{1}{2}+\mathbb{Z})) = \inf_{n\in\mathbb{Z}} |\kappa^2 - (n+\tfrac{1}{2})\pi^2|. \end{aligned} \tag{4.17}$$

Note that $d_0, d_1 \in [0, \frac{1}{2}\pi^2]$ and $d_0 + d_1 = \frac{1}{2}\pi^2$. For $n \in \mathbb{N}_0$, we choose $\tilde{\mu}_n$ according to the following four cases:

$$\tilde{\mu}_n = \begin{cases} (n+\frac{1}{2})\pi, & \text{if } \mathcal{B}_1 = \mathcal{B}_3 \text{ and } d_0 \in [0, \frac{1}{8}\pi^2], \\ n\pi, & \text{if } \mathcal{B}_1 = \mathcal{B}_3 \text{ and } d_0 \notin [0, \frac{1}{8}\pi^2], \\ n\pi, & \text{if } \mathcal{B}_1 \neq \mathcal{B}_3 \text{ and } d_0 \in [0, \frac{1}{8}\pi^2] \cup [\frac{3}{8}\pi^2, \frac{1}{2}\pi^2], \\ (n+\frac{1}{2})\pi, & \text{if } \mathcal{B}_1 \neq \mathcal{B}_3 \text{ and } d_0 \notin [0, \frac{1}{8}\pi^2] \cup [\frac{3}{8}\pi^2, \frac{1}{2}\pi^2]. \end{cases} \tag{4.18}$$

The next result states that the choices in (4.18) are critical in ensuring that the following auxiliary solution $\tilde{u}$ satisfying (4.16) is well defined. Furthermore, sufficient conditions under which the Dirichlet trace of an auxiliary function $\tilde{u}$ belongs to $H^{1/2}(\partial\Omega)$ and the Neumann trace in the x-direction of an auxiliary function $\tilde{u}$ belongs to $L_2(\partial\Omega)$ are presented. This allows us to fall back to the cases discussed in Section 4.1; more specifically, with g_j replaced by $g_j - \mathcal{B}_j(\tilde{u})$ on Γ_j for each $j \in \{2,4\}$. The proof of the following result is deferred to Section 4.2.

Proposition 4.5. *Assume $g_1 \in \mathcal{Z}^{1/2}(\Gamma_1)$ if $\mathcal{B}_1 = \frac{\partial}{\partial\boldsymbol{\nu}}$. Otherwise, assume $g_1 \in \mathcal{Z}^{3/2}(\Gamma_1)$ if $\mathcal{B}_1 = \mathbf{I}_d$. Suppose that $\{\tilde{\mu}_n\}_{n\in\mathbb{N}_0}$ are chosen according to (4.18). Let the auxiliary function $\tilde{u}$*

take the following form

$$\tilde{u} = \sum_{n=0}^{\infty} \widehat{g}_1(n)\tilde{X}_n(x)\tilde{Y}_n(y) \quad \textit{with} \quad \widehat{g}_1(n) := \int_{\Gamma_1} g_1(x)\tilde{X}_n(x)dx, \tag{4.19}$$

where $\{Y_n\}_{n\in\mathbb{N}_0}$ *solve*

$$\tilde{Y}_n''(y) + (\kappa^2 - \tilde{\mu}_n^2)\tilde{Y}_n(y) = 0 \quad \textit{in} \quad \mathcal{I} := (0,1), \quad n \in \mathbb{N}_0, \tag{4.20}$$

$$\mathcal{B}_1\tilde{Y}_n(0) = 1, \quad \mathcal{B}_3\tilde{Y}_n(1) = 0, \tag{4.21}$$

with $\mathcal{B}_1, \mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$. *Then, the auxiliary function* $\tilde{u}$ *in* (4.19) *satisfies* (4.16) *and each term of* $\tilde{u}$ *is well defined. Furthermore, we have that* $\tilde{u} \in H^1(\Omega)$, *the (Dirichlet) trace of* $\tilde{u}$ *is in* $H^{1/2}(\partial\Omega)$, *and the trace of* $\tilde{u}_x$ *(i.e., the Neumann trace in the* x*-direction of* $\tilde{u}$*) is in* $L_2(\partial\Omega)$.

Next, we study upper bounds of an auxiliary function satisfying (4.16), which is defined in (4.19).

Theorem 4.6. *Consider an auxiliary function* $\tilde{u}$ *satisfying* (4.16), *which is defined in* (4.19) *and takes into account of* (4.18). *Then,*

(1) For $\mathcal{B}_1 = \frac{\partial}{\partial \boldsymbol{\nu}}$, $\mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$, *and* $g_1 \in L_2(\Gamma_1)$, *the auxiliary function* $\tilde{u}$ *satisfies*

$$\|\nabla\tilde{u}\|_{0,\Omega} + \kappa\|\tilde{u}\|_{0,\Omega} \leqslant 2\sqrt{717}\max\{\kappa, 1\}\|g_1\|_{0,\Gamma_1}, \quad \forall \kappa > 0. \tag{4.22}$$

(2) For $\mathcal{B}_1 = \mathbf{I}_d$, $\mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$, *and* $g_1 \in \mathcal{Z}^{1/2}(\Gamma_1)$, *the auxiliary function* $\tilde{u}$ *satisfies*

$$\|\nabla\tilde{u}\|_{0,\Omega} + \kappa\|\tilde{u}\|_{0,\Omega} \leqslant 2\sqrt{43}(\max\{\kappa^2, 1\}\|g_1\|_{0,\Gamma_1} + \max\{\kappa^{\frac{1}{2}}, 1\}\|g_1\|_{\mathcal{Z}^{1/2}(\Gamma_1)}), \quad \forall \kappa > 0. \tag{4.23}$$

Note that by symmetry, the same results as above hold when $\mathcal{B}_1\tilde{u} = 0$ on Γ_1 and $\mathcal{B}_3\tilde{u} = g_3$ on Γ_3 in (4.16). Also, the conditions imposed on g_1 in Theorem 4.6 are weaker compared to those in Proposition 4.5, because in the former, we are only interested in finding an upper bound of the norm of an auxiliary solution and do not consider whether its traces belong to particular spaces or not. The following example demonstrates how the stability bounds (4.22) and (4.23) are sharp in the sense that the right-hand sides of (4.22) and (4.23) hold up to a constant multiple (independent of κ and g_1).

Example 4.4. In what follows, suppose that the conditions of Theorem 4.6 hold. Note that the source term in (4.16) vanishes. Suppose that $\mathcal{B}_1 = \mathcal{B}_3 = \frac{\partial}{\partial \boldsymbol{\nu}}$ and $g_1 \in L_2(\Gamma_1)$.

Consider $\kappa^2 := (n\pi)^2 + \frac{1}{4}\pi^2$ for a temporarily fixed $n \in \mathbb{N}$. Since $d_0 \notin [0, \frac{1}{8}\pi^2]$ in (4.18), let $\tilde{u} = -\frac{2}{\pi}\tilde{X}_n(x)\cos(\frac{\pi}{2}(y-1))$ be an auxiliary solution, where $\tilde{X}_n(x) = \sqrt{2}\cos(\tilde{\mu}_n x)$ or $\tilde{X}_n(x) = \sqrt{2}\sin(\tilde{\mu}_n x)$ with $\tilde{\mu}_n = n\pi$. Then $g_1 = \mathcal{B}_1\tilde{u} = \tilde{X}_n$ on Γ_1 and hence $\|g_1\|_{0,\Gamma_1} = 1$. The auxiliary solution $\tilde{u} = -\frac{2}{\pi}\tilde{X}_n(x)\cos(\frac{\pi}{2}(y-1))$ satisfies

$$\|\nabla\tilde{u}\|_{0,\Omega} + \kappa\|\tilde{u}\|_{0,\Omega} = \tfrac{2\sqrt{2}}{\pi}\kappa\|g_1\|_{0,\Gamma_1} \quad \text{with} \quad \kappa := \sqrt{(n\pi)^2 + \tfrac{1}{4}\pi^2}, \quad \forall n \in \mathbb{N}.$$

Suppose that $\mathcal{B}_1 = \frac{\partial}{\partial\boldsymbol{\nu}}$, $\mathcal{B}_3 = \mathbf{I}_d$, and $g_1 \in L_2(\Gamma_1)$. Consider $\kappa^2 := (n\pi)^2 + \pi^2$ for a temporarily fixed $n \in \mathbb{N}$. Since $d_0 \in [0, \frac{1}{8}\pi^2]$, let $\tilde{u} = \frac{1}{\pi}\tilde{X}_n(x)\sin(\pi(y-1))$ be an auxiliary solution, where $\tilde{X}_n(x) = \sqrt{2}\cos(\tilde{\mu}_n x)$ or $\tilde{X}_n(x) = \sqrt{2}\sin(\tilde{\mu}_n x)$ with $\tilde{\mu}_n = n\pi$. Then $g_1 = \mathcal{B}_1\tilde{u} = \tilde{X}_n$ on Γ_1 and hence $\|g_1\|_{0,\Gamma_1} = 1$. The auxiliary solution $\tilde{u} = \frac{1}{\pi}\tilde{X}_n(x)\sin(\pi(y-1))$ satisfies

$$\|\nabla\tilde{u}\|_{0,\Omega} + \kappa\|\tilde{u}\|_{0,\Omega} = \tfrac{\sqrt{2}}{\pi}\kappa\|g_1\|_{0,\Gamma_1} \quad \text{with} \quad \kappa := \sqrt{(n\pi)^2 + \pi^2}, \quad \forall n \in \mathbb{N}.$$

Suppose that $\mathcal{B}_1 = \mathbf{I}_d$, $\mathcal{B}_3 = \frac{\partial}{\partial\boldsymbol{\nu}}$, and $g_1 \in \mathcal{Z}^{1/2}(\Gamma_1)$. Consider $\kappa^2 := (\theta_n+\theta_n^{-1})^2+(\theta_n-\frac{1}{2}\pi)^2$, where $\theta_n := (n+\frac{1}{2})\pi$ for a temporarily fixed $n \in \mathbb{N}$. Since $d_0 \in [\frac{3}{8}\pi^2, \frac{1}{2}\pi^2]$ in (4.18), let $\tilde{u} = \tilde{X}_n(x)\frac{\cos((\theta_n+\theta_n^{-1})(y-1))}{\cos((\theta_n+\theta_n^{-1}))}$ be an auxiliary solution, where $\tilde{X}_n(x) = \sqrt{2}\sin(\tilde{\mu}_n x)$ or $\tilde{X}_n(x) = \sqrt{2}\cos(\tilde{\mu}_n x)$ with $\tilde{\mu}_n = n\pi$. Then $g_1 = \mathcal{B}_1\tilde{u} = \tilde{X}_n$ on Γ_1 and hence $\|g_1\|_{0,\Gamma_1} = 1$. Also, $\|g_1\|_{\mathcal{Z}^{1/2}(\Gamma_1)} = \tilde{\mu}_n^{1/2}$. The auxiliary solution $\tilde{u} = \tilde{X}_n(x)\frac{\cos((\theta_n+\theta_n^{-1})(y-1))}{\cos((\theta_n+\theta_n^{-1}))}$ satisfies

$$\begin{aligned}
\|\nabla\tilde{u}\|_{0,\Omega}^2 + \kappa^2\|\tilde{u}\|_{0,\Omega}^2 &= \frac{\kappa^2+\tilde{\mu}_n^2\frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})}}{\cos^2(\theta_n+\theta_n^{-1})} = \frac{\kappa^2+\tilde{\mu}_n^2\frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})}}{\sin^2(\theta_n^{-1})} \\
&\geqslant \theta_n^2\kappa^2 + \tilde{\mu}_n^2\theta_n^2\frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})} = \theta_n^2(\theta_n+\theta_n^{-1})^2 + \tilde{\mu}_n^2\theta_n^2\left(1+\frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})}\right) \\
&\geqslant \min\left\{\frac{\theta_n^2(\theta_n+\theta_n^{-1})^2}{((\theta_n+\theta_n^{-1})^2+\tilde{\mu}_n^2)^2}, \frac{\tilde{\mu}_n^2\theta_n^2}{2((\theta_n+\theta_n^{-1})^2+\tilde{\mu}_n^2)}\right\}(\kappa^4+\kappa^2) \\
&\geqslant \min\left\{\frac{\theta_n^2}{4(\theta_n+\theta_n^{-1})^2}, \frac{\pi^2\theta_n^2}{4(\theta_n+\theta_n^{-1})^2}\right\}(\kappa^4+\kappa\tilde{\mu}_n) \\
&= \frac{\theta_1^2}{4(\theta_1+\theta_1^{-1})^2}(\kappa^4+\kappa\tilde{\mu}_n) = \frac{81\pi^4}{4(9\pi^2+4)^2}(\kappa^4+\kappa\tilde{\mu}_n),
\end{aligned}$$

where we used the fact that $|\sin(x)| \leqslant |x|$ for all $x \geqslant 0$ to arrive at the first inequality. Using the basic inequality $a^2+b^2 \geqslant \frac{1}{\sqrt{2}}(a+b)$ for nonnegative numbers a and b, we have

$$\|\nabla\tilde{u}\|_{0,\Omega} + \kappa\|\tilde{u}\|_{0,\Omega} \geqslant \tfrac{9\pi^2}{2\sqrt{2}(9\pi^2+4)}(\kappa^2\|g_1\|_{0,\Gamma_1} + \kappa^{\frac{1}{2}}\|g_1\|_{\mathcal{Z}^{1/2}(\Gamma_1)}),$$

where $\kappa := \sqrt{(\theta_n+\theta_n^{-1})^2+\tilde{\mu}_n^2}$ with $\theta_n = (n+\frac{1}{2})\pi$ and $\tilde{\mu}_n = \theta_n - \frac{1}{2}\pi$ for all $n \in \mathbb{N}$.

Suppose that $\mathcal{B}_1 = \mathbf{I}_d$, $\mathcal{B}_3 = \mathbf{I}_d$, $g_1 \in \mathcal{Z}^{1/2}(\Gamma_1)$, and $g_3 = 0$. Consider $\kappa^2 := (\theta_n + \theta_n^{-1})^2 + \theta_n^2$, where $\theta_n := n\pi$ for a temporarily fixed $n \in \mathbb{N}$. Since $d_0 \notin [0, \frac{1}{8}\pi^2]$ in (4.18),

let $\tilde{u} = -\tilde{X}_n(x)\frac{\sin((\theta_n+\theta_n^{-1})(y-1))}{\sin(\theta_n+\theta_n^{-1})}$ be an auxiliary solution, where $\tilde{X}_n = \sqrt{2}\sin(\tilde{\mu}_n x)$ or $\tilde{X}_n = \sqrt{2}\cos(\tilde{\mu}_n x)$ with $\tilde{\mu}_n = \theta_n = n\pi$. Then $g_1 = \mathcal{B}_1\tilde{u} = \tilde{X}_n$ on Γ_1 and hence $\|g_1\|_{0,\Gamma_1} = 1$. Also, $\|g_1\|_{\mathcal{Z}^{1/2}(\Gamma_1)} = \tilde{\mu}_n^{1/2}$. The auxiliary solution $\tilde{u} = -\tilde{X}_n(x)\frac{\sin((\theta_n+\theta_n^{-1})(y-1))}{\sin(\theta_n+\theta_n^{-1})}$ satisfies

$$\begin{aligned}\|\nabla\tilde{u}\|_{0,\Omega}^2 + \kappa^2\|\tilde{u}\|_{0,\Omega}^2 &= \frac{\kappa^2-\theta_n^2\frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})}}{\sin^2(\theta_n+\theta_n^{-1})} = \frac{(\theta_n+\theta_n^{-1})^2+\theta_n^2\left(1-\frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})}\right)}{\sin^2(\theta_n^{-1})}\\ &\geqslant \theta_n^2(\theta_n+\theta_n^{-1})^2 + \theta_n^4\left(1-\frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})}\right)\\ &\geqslant \frac{\theta_1^2}{4(\theta_1+\theta_1^{-1})^2}(\kappa^4+\kappa\tilde{\mu}_n) = \frac{\pi^4}{4(\pi^2+1)^2}(\kappa^4+\kappa\tilde{\mu}_n),\end{aligned}$$

where we used the same steps as in the previous case, and the fact that $\theta_1 + \theta_1^{-1} > \pi$ and hence $1 - \frac{\sin(2(\theta_n+\theta_n^{-1}))}{2(\theta_n+\theta_n^{-1})} > \frac{1}{2}$ for all $n \geqslant 1$ to move from the first inequality to the second inequality. Using the basic inequality $a^2 + b^2 \geqslant \frac{1}{\sqrt{2}}(a+b)$ for nonnegative numbers a and b, we have

$$\|\nabla\tilde{u}\|_{0,\Omega} + \kappa\|\tilde{u}\|_{0,\Omega} \geqslant \frac{\pi^2}{2\sqrt{2}(\pi^2+1)}(\kappa^2\|g_1\|_{0,\Gamma_1} + \kappa^{\frac{1}{2}}\|g_1\|_{\mathcal{Z}^{1/2}(\Gamma_1)}),$$

where $\kappa := \sqrt{(\theta_n+\theta_n^{-1})^2+\theta_n^2}$ with $\theta_n = \tilde{\mu}_n = n\pi$ for all $n \in \mathbb{N}$.

We close this section with two important final remarks. By the superposition principle, the stability bounds for the case where all boundary conditions are inhomogeneous and the source term vanishes can be recovered by using Theorem 4.6, subtracting the traces of the auxiliary solutions from g_2 on Γ_2 and g_4 on Γ_4, using Theorems 4.2 and 4.3, and finally adding all these bounds. Additionally, if the source term is nonzero, then we may also add the stability bound in Theorem 4.4.

Note that the geometric assumptions in [83] simplify into three cases for a unit square domain: (1) all sides have impedance boundary conditions, (2) three sides have impedance boundary conditions and one side has a Dirichlet/Neumann boundary condition, or (3) two adjacent sides have Dirichlet/Neumann boundary conditions and the other two have impedance boundary conditions. By two adjacent sides, we mean two sides that are connected to each other; e.g., Γ_1 and Γ_2, or Γ_4 and Γ_1 in (4.2). Whenever a Dirichlet/Neumann boundary condition is imposed, [83] assumes that it is homogeneous. We emphasize that the boundary configurations in this paper are completely different from the assumptions used in [83]. Hence, it is not surprising that our wavenumber-explicit stability bounds are also different.

4.2 Proofs of Theorems 4.2 to 4.4 and 4.6, Lemma 4.8, and Proposition 4.5

To prove Theorem 4.2, we need the following result.

Lemma 4.7. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be the solutions of the following problem*

$$X_n''(x)+(\kappa^2-\mu_n^2)X_n(x)=0 \quad \text{in} \quad \mathcal{I}:=(0,1), \quad n\in\mathbb{N}_0, \tag{4.24}$$

$$\mathcal{B}_4X_n(0)=\delta_{j,4}, \quad \mathcal{B}_2X_n(1)=\delta_{j,2}, \tag{4.25}$$

where $\{\mu_n=n\pi\}_{n\in\mathbb{N}_0}$ or $\{\mu_n=(n+\frac{1}{2})\pi\}_{n\in\mathbb{N}_0}$ with $j\in\{2,4\}$, $\delta_{j,j}=1$, and $\delta_{j,m}=0$ for $j\neq m$. Define

$$\lambda_n:=\sqrt{\left|1-\tfrac{\mu_n^2}{\kappa^2}\right|} \quad \text{and} \quad \mathring{\lambda}_n:=\begin{cases}\lambda_n & \text{if } \mu_n^2\leqslant\kappa^2,\\ i\lambda_n & \text{if } \mu_n^2>\kappa^2,\end{cases} \quad \forall n\in\mathbb{N}_0.$$

Recall that $\mathcal{B}_4=\frac{\partial}{\partial\boldsymbol{\nu}}-i\kappa\mathbf{I}_d$ in (4.3).

(1) If $\mathcal{B}_4X_n(0)=-X_n'(0)-i\kappa X_n(0)=1$ and $\mathcal{B}_2X_n(1)=X_n'(1)-i\kappa X_n(1)=0$ with $\mathcal{B}_2=\frac{\partial}{\partial\boldsymbol{\nu}}-i\kappa\mathbf{I}_d$, then the solutions $\{X_n\}_{n\in\mathbb{N}_0}$ to the problem (4.24)–(4.25) are given by

$$X_n(x)=\tfrac{\mathring{\lambda}_n((1-\mathring{\lambda}_n^2)\cos(\kappa\mathring{\lambda}_n)\sin(\kappa\mathring{\lambda}_n(1-x))-(1+\mathring{\lambda}_n^2)\sin(\kappa\mathring{\lambda}_nx))}{\kappa(4\mathring{\lambda}_n^2+(1-\mathring{\lambda}_n^2)^2\sin^2(\kappa\mathring{\lambda}_n))}+\tfrac{(1+\mathring{\lambda}_n^2)\cos(\kappa\mathring{\lambda}_nx)-(1-\mathring{\lambda}_n^2)\cos(\kappa\mathring{\lambda}_n)\cos(\kappa\mathring{\lambda}_n(1-x))}{\kappa(4\mathring{\lambda}_n^2+(1-\mathring{\lambda}_n^2)^2\sin^2(\kappa\mathring{\lambda}_n))}i.$$

If $\mathcal{B}_4X_n(0)=-X_n'(0)-i\kappa X_n(0)=0$ and $\mathcal{B}_2X_n(1)=X_n'(1)-i\kappa X_n(1)=1$ with $\mathcal{B}_2=\frac{\partial}{\partial\boldsymbol{\nu}}-i\kappa\mathbf{I}_d$, then the solutions $\{X_n\}_{n\in\mathbb{N}_0}$ to (4.24)–(4.25) are given above with x replaced by $1-x$. Moreover, for both cases, the norms $\|X_n\|_{0,\mathcal{I}}$ and $\|X_n'\|_{0,\mathcal{I}}$ are given by

$$\|X_n\|_{0,\mathcal{I}}^2=\tfrac{\kappa\mathring{\lambda}_n(1+\mathring{\lambda}_n^2)-(1-\mathring{\lambda}_n^2)\cos(\kappa\mathring{\lambda}_n)\sin(\kappa\mathring{\lambda}_n)}{2\kappa^3\mathring{\lambda}_n(4\mathring{\lambda}_n^2+(1-\mathring{\lambda}_n^2)^2\sin^2(\kappa\mathring{\lambda}_n))}, \quad \|X_n'\|_{0,\mathcal{I}}^2=\tfrac{\kappa\mathring{\lambda}_n^2(1+\lambda_n^2)-\mathring{\lambda}_n(\mathring{\lambda}_n^2-1)\cos(\kappa\mathring{\lambda}_n)\sin(\kappa\mathring{\lambda}_n)}{2\kappa(4\mathring{\lambda}_n^2+(1-\mathring{\lambda}_n^2)^2\sin^2(\kappa\mathring{\lambda}_n))}.$$

(2) If $\mathcal{B}_4X_n(0)=1$ and $\mathcal{B}_2X_n(1)=\alpha X_n'(1)+(1-\alpha)X_n(1)=0$ with $\alpha\in\{0,1\}$, then the solutions $\{X_n\}_{n\in\mathbb{N}_0}$ to the problem (4.24)–(4.25) are given by

$$X_n(x)=\tfrac{-\mathring{\lambda}_n(\cos(\kappa\mathring{\lambda}_n)\sin(\kappa\mathring{\lambda}_n(1-x))+\alpha\sin(\kappa\mathring{\lambda}_nx))}{\kappa((1-\mathring{\lambda}_n^2)\cos^2(\kappa\mathring{\lambda}_n)+\mathring{\lambda}_n^2-(1-\alpha)(1+\mathring{\lambda}_n^2))}+\tfrac{\cos(\kappa\mathring{\lambda}_n)\cos(\kappa\mathring{\lambda}_n(1-x))-(1-\alpha)\cos(\kappa\mathring{\lambda}_nx)}{\kappa((1-\mathring{\lambda}_n^2)\cos^2(\kappa\mathring{\lambda}_n)+\mathring{\lambda}_n^2-(1-\alpha)(1+\mathring{\lambda}_n^2))}i.$$

Moreover, the norms $\|X_n\|_{0,\mathcal{I}}$ *and* $\|X_n'\|_{0,\mathcal{I}}$ *are given by*

$$\|X_n\|_{0,\mathcal{I}}^2 = \frac{\sin(\kappa\mathring{\lambda}_n)\cos(\kappa\mathring{\lambda}_n)+(-1)^{1-\alpha}\kappa\mathring{\lambda}_n}{2\kappa^3\mathring{\lambda}_n\big(\big(1-\mathring{\lambda}_n^2\big)\cos^2(\kappa\mathring{\lambda}_n)+\lambda_n^2-(1-\alpha)(1+\lambda_n^2)\big)},$$
$$\|X_n'\|_{0,\mathcal{I}}^2 = \frac{\mathring{\lambda}_n\big(-\sin\big(\kappa\mathring{\lambda}_n\big)\cos\big(\kappa\mathring{\lambda}_n\big)+(-1)^{1-\alpha}\kappa\mathring{\lambda}_n\big)}{2\kappa\big(\cos^2\big(\kappa\mathring{\lambda}_n\big)(1-\lambda_n^2)+\lambda_n^2-(1-\alpha)(1+\lambda_n^2)\big)}.$$

(3) If $\mathcal{B}_4X_n(0) = 0$ *and* $\mathcal{B}_2X_n(1) = \alpha X_n'(1) + (1-\alpha)X_n(1) = 1$ *with* $\alpha \in \{0,1\}$*, then the solutions* $\{X_n\}_{n\in\mathbb{N}_0}$ *to the problem* (4.24)–(4.25) *are given by*

$$X_n(x) = \frac{\sin(\kappa\mathring{\lambda}_n x)(\alpha\cos(\kappa\mathring{\lambda}_n)-(1-\alpha)\sin(\kappa\mathring{\lambda}_n))-\cos(\kappa\mathring{\lambda}_n x)\mathring{\lambda}_n^2(\alpha\sin(\kappa\mathring{\lambda}_n)+(1-\alpha)\cos(\kappa\mathring{\lambda}_n))}{(\kappa\mathring{\lambda}_n)^\alpha((1-\mathring{\lambda}_n^2)\cos^2(\kappa\mathring{\lambda}_n)+\mathring{\lambda}_n^2-(1-\alpha)(1+\mathring{\lambda}_n^2))}$$
$$+\frac{\alpha\kappa^{-\alpha}\cos(\kappa\mathring{\lambda}_n(1-x))-(1-\alpha)\kappa^{-2\alpha}\mathring{\lambda}_n^{1-\alpha}\sin(\kappa\mathring{\lambda}_n(1-x))}{(1-\mathring{\lambda}_n^2)\cos^2(\kappa\mathring{\lambda}_n)+\mathring{\lambda}_n^2-(1-\alpha)(1+\mathring{\lambda}_n^2)}i.$$

Moreover, the norms $\|X_n\|_{0,\mathcal{I}}$ *and* $\|X_n'\|_{0,\mathcal{I}}$ *are given by*

$$\|X_n\|_{0,\mathcal{I}}^2 = \frac{\kappa\mathring{\lambda}_n(1+\mathring{\lambda}_n^2)-(1-\mathring{\lambda}_n^2)\cos(\kappa\mathring{\lambda}_n)\sin(\kappa\mathring{\lambda}_n)}{2(\kappa\mathring{\lambda}_n)^{2\alpha+1}(\alpha\mathring{\lambda}_n^2+(1-\alpha)+(-1)^{1-\alpha}(1-\mathring{\lambda}_n^2)\cos^2(\kappa\mathring{\lambda}_n))},$$
$$\|X_n'\|_{0,\mathcal{I}}^2 = \frac{\kappa\mathring{\lambda}_n(1+\mathring{\lambda}_n^2)+(1-\mathring{\lambda}_n^2)\sin(\kappa\mathring{\lambda}_n)\cos(\kappa\mathring{\lambda}_n)}{2(\kappa\mathring{\lambda}_n)^{2\alpha-1}(\alpha\mathring{\lambda}_n^2+(1-\alpha)+(-1)^{1-\alpha}(1-\mathring{\lambda}_n^2)\cos^2(\kappa\mathring{\lambda}_n))}.$$

Proof. Recall from the standard ordinary differential equation theory that the solution to (4.24)-(4.25) for each $n \in \mathbb{N}_0$ with $\mu_n^2 < \kappa^2$ takes the form

$$X_n(x) = A_n\exp(\mathrm{i}\kappa\lambda_n x) + B_n\exp(-ia\kappa\lambda_n x), \tag{4.26}$$

where A_n, B_n are uniquely determined by imposing the boundary conditions. Then,

$$\|X_n\|_{0,\mathcal{I}}^2 = \int_0^1 |\Re(X_n)|^2 + |\Im(X_n)|^2 dx \quad \text{and} \quad \|X_n'\|_{0,\mathcal{I}}^2 = \int_0^1 |\Re(X_n')|^2 + |\Im(X_n')|^2 dx.$$

For $n \in \mathbb{N}_0$ with $\mu_n^2 > \kappa^2$, each solution X_n and its norms can be directly obtained by replacing λ_n with $\mathrm{i}\lambda_n$ in (4.26). For $n \in \mathbb{N}_0$ such that $\mu_n^2 = \kappa^2$, the solution X_n and its norms can be obtained by letting λ_n tend to zero in (4.26). □

The following quantities will be used numerous times in the proofs of Theorems 4.2 and 4.3. Similar quantities will also be used multiple times in the proof of (4.6). Let $\{\mu_n = n\pi\}_{n\in\mathbb{N}_0}$ or $\{\mu_n = (n+\frac{1}{2})\pi\}_{n\in\mathbb{N}_0}$ and $\{X_n\}_{n\in\mathbb{N}_0}$ be solutions to (4.24)-(4.25) with boundary conditions explicitly given in the proofs. Define $N_p := \max\{n \in \mathbb{N}_0 : \mu_n^2 < \kappa^2\}$, $N_c \in \mathbb{N}$ such that $\mu_{N_c}^2 = \kappa^2$, $N_e := \min\{n \in \mathbb{N}_0 : \mu_n^2 > \kappa^2\}$, and

$$\phi_n := \|X_n'\|_{0,\mathcal{I}}^2 + (\mu_n^2+\kappa^2)\|X_n\|_{0,\mathcal{I}}^2, \quad 0 \leqslant n \leqslant N_p,$$

$$\theta_{N_c} := \|X'_{N_c}\|^2_{0,\mathcal{I}} + (\mu^2_{N_c} + \kappa^2)\|X_{N_c}\|^2_{0,\mathcal{I}}, \tag{4.27}$$
$$\psi_n := \|X'_n\|^2_{0,\mathcal{I}} + (\mu^2_n + \kappa^2)\|X_n\|^2_{0,\mathcal{I}}, \quad n \geqslant N_e.$$

Also recall that if $a, b \geqslant 0$, then the following inequality always holds

$$\sqrt{a^2+b^2} \leqslant a + b \leqslant \sqrt{2}\sqrt{a^2+b^2}. \tag{4.28}$$

Let $\mathbb{1}_A$ denote the indicator/characteristic function of the set A.

Proof of Theorem 4.2. Given the boundary assumptions, the solution u can be expressed as $u = \sum_{n=0}^{\infty} \widehat{g}_4(n) X_n(x) Y_n(y)$ with $\widehat{g}_4(n) := \int_{\Gamma_4} g_4(y) Y_n(y) dy$, where $\{X_n\}_{n\in\mathbb{N}_0}$ are stated in Lemma 4.7 and $\{Y_n\}_{n\in\mathbb{N}_0}$ are stated in (4.10).

Recall that $\lambda_n = \sqrt{|1-\mu_n^2\kappa^{-2}|}$ and observe that $\|Y'_n\|_{0,\mathcal{I}} = \mu_n$. By (4.6), since both $\{Y_n\}_{n\in\mathbb{N}_0}$ and $\{Y'_n\}_{n\in\mathbb{N}}$ are orthogonal systems in $L_2(\mathcal{I})$, we deduce that

$$\begin{aligned}
&\|\nabla u\|^2_{0,\Omega} + \kappa^2\|u\|^2_{0,\Omega} \\
&\quad = \|\sum_{n=0}^{\infty} \widehat{g}_4(n) X'_n Y_n\|^2_{0,\Omega} + \|\sum_{n=0}^{\infty} \widehat{g}_4(n) X_n Y'_n\|^2_{0,\Omega} + \kappa^2\|\sum_{n=0}^{\infty} \widehat{g}_4(n) X_n Y_n\|^2_{0,\Omega} \\
&\quad = \sum_{n=0}^{\infty} |\widehat{g}_4(n)|^2 \|X'_n\|^2_{0,\mathcal{I}} + \sum_{n=0}^{\infty} |\widehat{g}_4(n)\mu_n|^2 \|X_n\|^2_{0,\mathcal{I}} + \kappa^2 \sum_{n=0}^{\infty} |\widehat{g}_4(n)|^2 \|X_n\|^2_{0,\mathcal{I}} \\
&\quad = \sum_{n=0}^{\infty} |\widehat{g}_4(n)|^2 \left(\|X'_n\|^2_{0,\mathcal{I}} + (\mu_n^2+\kappa^2)\|X_n\|^2_{0,\mathcal{I}}\right) \\
&\quad \leqslant \max\left\{ \max_{0\leqslant n\leqslant N_p} \phi_n, \theta_{N_c}, \max_{n\geqslant N_e} \psi_n \right\} \sum_{n=0}^{\infty} |\widehat{g}_4(n)|^2,
\end{aligned} \tag{4.29}$$

where ϕ_n, θ_{N_c}, and ψ_n are defined as in (4.27).

Case I: suppose $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa \mathbf{I}_d$. Using item (1) of Lemma 4.7, we obtain

$$\phi_n = \frac{(1+\lambda_n^2) - \frac{\sin(2\kappa\lambda_n)}{2\kappa\lambda_n}(1-\lambda_n^2)^2}{(1+\lambda_n^2)^2 - \cos^2(\kappa\lambda_n)(1-\lambda_n^2)^2}, \quad \psi_n = \frac{\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n} 2(\lambda_n^2+1)^2 + 2(\lambda_n^2-1)}{(\lambda_n^2+1)^2(\cosh(2\kappa\lambda_n)-1) + 8\lambda_n^2}.$$

To obtain an upper bound for ϕ_n, we note that for all $n \leqslant N_p$ and $\kappa\lambda_n \in (0, \frac{\pi}{4}]$

$$\begin{aligned}
&1 - \kappa^2\lambda_n^2 \leqslant \cos^2(\kappa\lambda_n) \leqslant 1 - \kappa^2\lambda_n^2 + \tfrac{1}{3}\kappa^4\lambda_n^4, \quad 1 - \tfrac{2}{3}\kappa^2\lambda_n^2 \leqslant \tfrac{\sin(2\kappa\lambda_n)}{2\kappa\lambda_n}, \quad \text{and} \\
&1 - \kappa^2\lambda_n^2 + \tfrac{1}{3}\kappa^4\lambda_n^4 \leqslant 1 - \tfrac{2}{3}\kappa^2\lambda_n^2 \leqslant 1.
\end{aligned} \tag{4.30}$$

Moreover, for all $0 < \kappa < 1$,

$$\begin{aligned} &N_p = 0 \quad \text{(i.e., } \lambda_n = 1\text{)} \quad \text{if} \quad \{\mu_n = n\pi\}_{n\in\mathbb{N}_0}\text{; otherwise,} \\ &N_p \text{ does not exist if } \{\mu_n = (n+\tfrac{1}{2})\pi\}_{n\in\mathbb{N}_0}. \end{aligned} \tag{4.31}$$

Now, for $\kappa > 0$ and $n \leqslant N_p$,

$$\begin{aligned} \phi_n &\leqslant \tfrac{1}{2}\mathbb{1}_{\{\kappa<1\}} + \tfrac{(1+\lambda_n^2)^2-\left(1-\frac{2}{3}\kappa^2\lambda_n^2\right)(1-\lambda_n^2)^2}{(1+\lambda_n^2)^2-\left(1-\kappa^2\lambda_n^2+\frac{1}{3}\kappa^4\lambda_n^4\right)(1-\lambda_n^2)^2}\mathbb{1}_{\{\kappa\geqslant 1, 0<\lambda_n\leqslant\frac{\pi}{4\kappa}\}} + \tfrac{(1+\lambda_n^2)+(1-\lambda_n^2)^2}{(1+\lambda_n^2)^2-(1-\lambda_n^2)^2}\mathbb{1}_{\{\kappa\geqslant 1, \frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\ &\leqslant \tfrac{1}{2}\mathbb{1}_{\{\kappa<1\}} + \mathbb{1}_{\{\kappa\geqslant 1, 0<\lambda_n\leqslant\frac{\pi}{4\kappa}\}} + \left(\tfrac{1}{2\lambda_n^2} - \tfrac{1}{4} + \tfrac{1}{4}\lambda_n^2\right)\mathbb{1}_{\{\kappa\geqslant 1, \frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\ &\leqslant \tfrac{1}{2}\mathbb{1}_{\{\kappa<1\}} + \mathbb{1}_{\{\kappa\geqslant 1, 0<\lambda_n\leqslant\frac{\pi}{4\kappa}\}} + \tfrac{8}{\pi^2}\kappa^2\mathbb{1}_{\{\kappa\geqslant 1, \frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\ &\leqslant \max\{\tfrac{1}{2}, 1, \tfrac{8}{\pi^2}\}\max\{\kappa^2, 1\} \leqslant \max\{\kappa^2, 1\}, \end{aligned}$$

where we respectively used (4.31) and (4.30) to obtain the first and second terms of the first inequality. Next, to obtain an upper bound for ψ_n, we note that $\frac{\sinh(x)}{x} \leqslant \cosh(x)$ for all $x \in \mathbb{R}$ and so

$$\psi_n \leqslant F(\lambda_n, z) := \tfrac{z2(\lambda_n^2+1)^2+2(\lambda_n^2-1)}{(\lambda_n^2+1)^2(z-1)+8\lambda_n^2} \quad \text{with} \quad z := \cosh(2\kappa\lambda_n).$$

Then we have $\frac{dF}{dz} = \frac{2(\lambda_n^2+1)^2\lambda_n^2(5-\lambda_n^2)}{((z-1)\lambda_n^4+(2z+6)\lambda_n^2+z-1)^2}$. For $\kappa > 0$ and $0 < \lambda_n \leqslant \sqrt{5}$, F is increasing to $\lim_{z\to\infty} F(\lambda_n, z) = 2$. Since $\{\mu_n = n\pi\}_{n\in\mathbb{N}_0}$ or $\{\mu_n = (n+\frac{1}{2})\pi\}_{n\in\mathbb{N}_0}$, we also note that

$$\eta := \sqrt{\tfrac{\pi^2}{4} - 1} \leqslant \sqrt{\mu_n^2 - \kappa^2} = \kappa\lambda_n \quad \text{for } 0 < \kappa < 1, n \in \{n \in \mathbb{N}_0 : \mu_n^2 > \kappa^2\}. \tag{4.32}$$

Now, for $\kappa > 0$ and $n \geqslant N_e$,

$$\begin{aligned} \psi_n &\leqslant 2\mathbb{1}_{\{\kappa>0, \lambda_n\leqslant\sqrt{5}\}} + F(\lambda_n, z)\mathbb{1}_{\{\kappa>0, \lambda_n>\sqrt{5}\}} \leqslant 2\mathbb{1}_{\{\kappa>0, \lambda_n\leqslant\sqrt{5}\}} + \tfrac{2(z+1)}{z-1}\mathbb{1}_{\{\kappa>0, \lambda_n>\sqrt{5}\}} \\ &\leqslant 2\mathbb{1}_{\{\kappa>0, \lambda_n\leqslant\sqrt{5}\}} + \tfrac{2(\cosh(2\sqrt{5})+1)}{\cosh(2\sqrt{5})-1}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n>\sqrt{5}\}} + \tfrac{2(\cosh(2\eta)+1)}{\cosh(2\eta)-1}\mathbb{1}_{\{\kappa<1, \lambda_n>\sqrt{5}\}} \\ &\leqslant \max\left\{2, \tfrac{2(\cosh(2\sqrt{5})+1)}{\cosh(2\sqrt{5})-1}, \tfrac{2(\cosh(2\eta)+1)}{\cosh(2\eta)-1}\right\} \leqslant 3. \end{aligned}$$

Consequently,

$$\max\left\{\max_{0\leqslant n\leqslant N_p}\phi_n, \theta_{N_c}, \max_{n\geqslant N_e}\psi_n\right\} \leqslant \max\left\{\max\{\kappa^2, 1\}, \tfrac{2\kappa^2+9}{3\kappa^2+12}, 3\right\} \leqslant 3\max\{\kappa^2, 1\}.$$

Plugging in the above estimate back into (4.29), applying the Parseval's identity, and finally using (4.28), we have (4.11).

Case II: suppose $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$. Using item (2) of Lemma 4.7 with $\alpha = 1$, we obtain

$$\phi_n = \tfrac{1+(1-\lambda_n^2)\frac{\sin(2\kappa\lambda_n)}{2\kappa\lambda_n}}{(1-\lambda_n^2)\cos^2(\kappa\lambda_n)+\lambda_n^2}, \quad \psi_n = \tfrac{2\left(\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}(1+\lambda_n^2)+1\right)}{(\lambda_n^2+1)\cosh(2\kappa\lambda_n)+1-\lambda_n^2}.$$

Now, for $\kappa > 0$ and $n \leqslant N_p$,

$$\begin{aligned}
\phi_n &\leqslant \mathbb{1}_{\{\kappa<1\}} + \tfrac{2-\lambda_n^2}{(1-\lambda_n^2)(1-\kappa^2\lambda_n^2)+\lambda_n^2}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \min\{\frac{1}{\sqrt{2}}, \frac{\pi}{4k}\}\}} + \tfrac{2}{\lambda_n^2}\mathbb{1}_{\{\kappa\geqslant 1, \min\{\frac{1}{\sqrt{2}}, \frac{\pi}{4\kappa}\}<\lambda_n\leqslant 1\}} \\
&\leqslant \mathbb{1}_{\{\kappa<1\}} + \tfrac{2}{1+\lambda_n^2(\lambda_n^2-1)\kappa^2}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \min\{\frac{1}{\sqrt{2}}, \frac{\pi}{4\kappa}\}\}} + 4\kappa^2\mathbb{1}_{\{\kappa\geqslant 1, \min\{\frac{1}{\sqrt{2}}, \frac{\pi}{4\kappa}\}<\lambda_n\leqslant 1\}} \\
&\leqslant \mathbb{1}_{\{\kappa<1\}} + \tfrac{2}{1-\frac{\kappa^2}{4}}\mathbb{1}_{\{1\leqslant\kappa<\frac{\pi}{2\sqrt{2}}, \lambda_n \leqslant \frac{1}{\sqrt{2}}\}} + \tfrac{512\kappa^2}{256\kappa^2-16\pi^2\kappa^2+\pi^4}\mathbb{1}_{\{\kappa\geqslant\frac{\pi}{2\sqrt{2}}, \lambda_n\leqslant\frac{\pi}{4\kappa}\}} \\
&\quad + 4\kappa^2\mathbb{1}_{\{\kappa\geqslant 1, \min\{\frac{1}{\sqrt{2}}, \frac{\pi}{4\kappa}\}<\lambda_n\leqslant 1\}} \\
&\leqslant \mathbb{1}_{\{\kappa<1\}} + \tfrac{64}{32-\pi^2}\mathbb{1}_{\{1\leqslant\kappa<\frac{\pi}{2\sqrt{2}}, \lambda_n\leqslant\frac{1}{\sqrt{2}}\}} + \tfrac{512}{256-16\pi^2}\mathbb{1}_{\{\kappa\geqslant\frac{\pi}{2\sqrt{2}}, \lambda_n\leqslant\frac{\pi}{4\kappa}\}} + 4\kappa^2\mathbb{1}_{\{\kappa\geqslant 1, \min\{\frac{1}{\sqrt{2}}, \frac{\pi}{4\kappa}\}<\lambda_n\leqslant 1\}} \\
&\leqslant \max\{1, \tfrac{64}{32-\pi^2}, \tfrac{512}{256-16\pi^2}, 4\}\max\{\kappa^2, 1\} \leqslant 6\max\{\kappa^2, 1\},
\end{aligned}$$

where we respectively used (4.31) and (4.30) with $\kappa\lambda_n \in (0, \min\{\frac{\kappa}{\sqrt{2}}, \frac{\pi}{4}\}]$ to obtain the first and second terms of the first inequality. Next, to obtain an upper bound for ψ_n, we note that $\frac{x^2}{\cosh(x)} < \frac{5}{4}$ and $\frac{\sinh(x)}{x} \leqslant \cosh(x)$ for all $x \in \mathbb{R}$. For $\kappa > 0$ and $n \geqslant N_e$,

$$\begin{aligned}
\psi_n &= \tfrac{2\left(\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}(1+\lambda_n^2)+1-\lambda_n^2\right)}{\cosh(2\kappa\lambda_n)(1+\lambda_n^2)+1-\lambda_n^2} + \tfrac{2\lambda_n^2}{\lambda_n^2(\cosh(2\kappa\lambda_n)-1)+1+\cosh(2\kappa\lambda_n)} \\
&\leqslant 2 + \tfrac{2\lambda_n^2}{\lambda_n^2(\cosh(2\kappa\lambda_n)-1)+1+\cosh(2\kappa\lambda_n)} \\
&\leqslant 2 + \tfrac{2}{\cosh(2\kappa\lambda_n)-1}\mathbb{1}_{\{\kappa<1\}} + \tfrac{(2\kappa\lambda_n)^2}{2\kappa^2\cosh(2\kappa\lambda_n)}\mathbb{1}_{\{\kappa\geqslant 1\}} \\
&\leqslant 2 + \tfrac{2}{\cosh(2\eta)-1}\mathbb{1}_{\{\kappa<1\}} + \tfrac{5}{8}\mathbb{1}_{\{\kappa\geqslant 1\}} \leqslant \max\{2, \tfrac{2}{\cosh(2\eta)-1}, \tfrac{5}{8}\} \leqslant 3,
\end{aligned}$$

where we used (4.32) to arrive at the second term of the third inequality. Consequently,

$$\max\left\{\max_{0\leqslant n\leqslant N_p}\phi_n, \theta_{N_c}, \max_{n\geqslant N_e}\psi_n\right\} \leqslant \max\left\{6\max\{\kappa^2, 1\}, 2, 3\right\} = 6\max\{\kappa^2, 1\}.$$

Plugging in the above estimate back into (4.29), applying the Parseval's identity, and finally using (4.28), we have (4.11).

Case III: suppose $\mathcal{B}_2 = \mathbf{I}_d$. This configuration has been studied in [41], but we include the proof for the sake of completeness. Using item (2) of Lemma 4.7 with $\alpha = 0$, we obtain

$$\phi_n = \tfrac{1-(1-\lambda_n^2)\frac{\sin(2\kappa\lambda_n)}{2\kappa\lambda_n}}{1-(1-\lambda_n^2)\cos^2(\kappa\lambda_n)}, \quad \psi_n = \tfrac{2\left(\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}(1+\lambda_n^2)-1\right)}{(1+\lambda_n^2)\cosh(2\kappa\lambda_n)-(1-\lambda_n^2)}.$$

Now, for $\kappa > 0$ and $n \leqslant N_p$,

$$\begin{aligned}\phi_n &\leqslant \mathbb{1}_{\{\kappa<1\}} + \frac{1-(1-\lambda_n^2)\left(1-\frac{2}{3}\kappa^2\lambda_n^2\right)}{1-(1-\lambda_n^2)\left(1-\kappa^2\lambda_n^2+\frac{1}{3}\kappa^4\lambda_n^4\right)}\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{\pi}{4\kappa}\}} + \frac{2}{\lambda_n^2}\mathbb{1}_{\{\kappa\geqslant 1,\frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\ &\leqslant \mathbb{1}_{\{\kappa<1\}} + \mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{\pi}{4\kappa}\}} + \frac{32}{\pi^2}\kappa^2\mathbb{1}_{\{\kappa\geqslant 1,\frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\ &\leqslant \max\{1,\tfrac{32}{\pi^2}\}\max\{\kappa^2,1\} \leqslant 4\max\{\kappa^2,1\},\end{aligned}$$

where we respectively used (4.31) and (4.30) to obtain the first and second terms of the first inequality. Next, for $\kappa > 0$ and $n \geqslant N_e$, we have

$$\psi_n \leqslant \frac{2\left(\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}(1+\lambda_n^2)-1\right)}{(1+\lambda_n^2)\cosh(2\kappa\lambda_n)-1} \leqslant 2.$$

Consequently,

$$\max\left\{\max_{0\leqslant n\leqslant N_p}\phi_n, \theta_{N_c}, \max_{n\geqslant N_e}\psi_n\right\} \leqslant \max\left\{4\max\{\kappa^2,1\}, \tfrac{2\kappa^2+3}{3\kappa^2+3}, 2\right\} = 4\max\{\kappa^2,1\}.$$

Plugging in the above estimate back into (4.29), applying the Parseval's identity, and finally using (4.28), we have (4.11). □

Proof of Theorem 4.3. We shall only focus on items (2) and (3), since the proof of item (1) is identical to the proof of Theorem 4.2 (Case I). Given the boundary assumptions, the solution u can be expressed as $u = \sum_{n=0}^{\infty}\widehat{g}_2(n)X_n(x)Y_n(y)$ with $\widehat{g}_2(n) := \int_{\Gamma_2} g_2(y)Y_n(y)dy$, where $\{X_n\}_{n\in\mathbb{N}_0}$ are stated in Lemma 4.7 and $\{Y_n\}_{n\in\mathbb{N}_0}$ are stated in (4.10).

Recall that $\lambda_n := \sqrt{|1-\mu_n^2\kappa^{-2}|}$ and observe that $\|Y_n'\|_{0,\mathcal{I}} = \mu_n$. By (4.6), since $\{Y_n\}_{n\in\mathbb{N}_0}$ and $\{Y_n'\}_{n\in\mathbb{N}_0}$ are orthogonal systems in $L_2(\mathcal{I})$, we deduce that

$$\begin{aligned}&\|\nabla u\|_{0,\Omega}^2 + \kappa^2\|u\|_{0,\Omega}^2 \\ &\quad= \|\sum_{n=0}^{\infty}\widehat{g}_2(n)X_n'Y_n\|_{0,\Omega}^2 + \|\sum_{n=0}^{\infty}\widehat{g}_2(n)X_nY_n'\|_{0,\Omega}^2 + \kappa^2\|\sum_{n=0}^{\infty}\widehat{g}_2(n)X_nY_n\|_{0,\Omega}^2 \\ &\quad= \sum_{n=0}^{\infty}|\widehat{g}_2(n)|^2\|X_n'\|_{0,\mathcal{I}}^2 + \sum_{n=0}^{\infty}|\widehat{g}_2(n)\mu_n|^2\|X_n\|_{0,\mathcal{I}}^2 + \kappa^2\sum_{n=0}^{\infty}|\widehat{g}_2(n)|^2\|X_n\|_{0,\mathcal{I}}^2,\end{aligned} \tag{4.33}$$

Regrouping the terms, we have

$$\begin{aligned}&\sum_{n=0}^{\infty}|\widehat{g}_2(n)|^2(\|X_n'\|_{0,\mathcal{I}}^2 + (\mu_n^2+\kappa^2)\|X_n\|_{0,\mathcal{I}}^2) \\ &\qquad\qquad\leqslant \max\left\{\max_{0\leqslant n\leqslant N_p}\phi_n, \theta_{N_c}, \max_{n\geqslant N_e}\psi_n\right\}\sum_{n=0}^{\infty}|\widehat{g}_2(n)|^2,\end{aligned} \tag{4.34}$$

where ϕ_n, θ_{N_c}, and ψ_n are defined as in (4.27).

Item (2): suppose $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$. Using item (3) of Lemma 4.7 with $\alpha = 1$, we obtain

$$\phi_n = \tfrac{(1+\lambda_n^2)-\frac{\sin(2\kappa\lambda_n)}{2\kappa\lambda_n}(1-\lambda_n^2)^2}{\lambda_n^2((1-\lambda_n^2)\cos^2(\kappa\lambda_n)+\lambda_n^2)}, \quad \psi_n = \tfrac{2\left((1+\lambda_n^2)^2\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}+(\lambda_n^2-1)\right)}{\lambda_n^2((\lambda_n^2+1)(\cosh(2\kappa\lambda_n)-1)+2)}.$$

First, we note that for $\kappa \geqslant 1$,

$$\tfrac{d}{d\lambda_n}\left(2\kappa^2\lambda_n^4 - (4\kappa^2+3)\lambda_n^2 + 2\kappa^2 + 9\right) = 2\lambda_n(4(\lambda_n^2-1)\kappa^2 - 3) \leqslant 0, \quad \forall \lambda_n \in (0, \sqrt{\tfrac{3}{4\kappa^2}+1}].$$

Now, for $\kappa > 0$ and $n \leqslant N_p$,

$$\begin{aligned}
\phi_n &\leqslant 2\mathbb{1}_{\{\kappa<1\}} + \tfrac{(1+\lambda_n^2)-\left(1-\frac{2}{3}\kappa^2\lambda_n^2\right)(1-\lambda_n^2)^2}{\lambda_n^2((1-\lambda_n^2)(1-\kappa^2\lambda_n^2)+\lambda_n^2)}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \frac{\pi}{4\kappa}\}} + \tfrac{(1+\lambda_n^2)+(1-\lambda_n^2)^2}{\lambda_n^4}\mathbb{1}_{\{\kappa\geqslant 1, \frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}}, \\
&\leqslant 2\mathbb{1}_{\{\kappa<1\}} + \tfrac{2\kappa^2\lambda_n^4-(4\kappa^2+3)\lambda_n^2+2\kappa^2+9}{3+3\lambda_n^2(\lambda_n^2-1)\kappa^2}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \frac{\pi}{4\kappa}\}} + \left(\tfrac{512}{\pi^4}\kappa^4 - \tfrac{16}{\pi^2}\kappa^2 + 1\right)\mathbb{1}_{\{\kappa\geqslant 1, \frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\
&\leqslant 2\mathbb{1}_{\{\kappa<1\}} + \tfrac{2\kappa^2+9}{3(1-\lambda_n^2\kappa^2)}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \frac{\pi}{4\kappa}\}} + \left(\tfrac{512}{\pi^4}\kappa^4 - \tfrac{16}{\pi^2}\kappa^2 + 1\right)\mathbb{1}_{\{\kappa\geqslant 1, \frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\
&\leqslant 2\mathbb{1}_{\{\kappa<1\}} + \tfrac{2\kappa^2+9}{3(1-\frac{\pi^2}{16})}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \frac{\pi}{4\kappa}\}} + \left(\tfrac{512}{\pi^4}\kappa^4 - \tfrac{16}{\pi^2}\kappa^2 + 1\right)\mathbb{1}_{\{\kappa\geqslant 1, \frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}} \\
&\leqslant \max\left\{2, \tfrac{11}{3}(1-\tfrac{\pi^2}{16})^{-1}, \tfrac{512}{\pi^4} - \tfrac{16}{\pi^2} + 1\right\}\max\{\kappa^4, 1\} \leqslant 10\max\{\kappa^4, 1\},
\end{aligned}$$

where we respectively used (4.31) and (4.30) to obtain the first and second terms of the first inequality. Next, we note that for all $n \geqslant N_e$ and $\kappa\lambda_n \in (0, 1]$,

$$\tfrac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n} \leqslant 1 + \tfrac{4}{3}\kappa^2\lambda_n^2 \quad \text{and} \quad 1 + 2\kappa^2\lambda_n^2 \leqslant \cosh(2\kappa\lambda_n). \tag{4.35}$$

Now, for $\kappa > 0$ and $n \geqslant N_e$,

$$\begin{aligned}
\psi_n &\leqslant 2\left(\tfrac{(1+\lambda_n^2)\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}+1}{\lambda_n^2(\cosh(2\kappa\lambda_n)-1)}\right)\mathbb{1}_{\{\kappa<1\}\cup\{\kappa\geqslant 1, \lambda_n > \frac{1}{\kappa}\}} + \tfrac{(1+\lambda_n^2)^2\left(1+\frac{4}{3}\kappa^2\lambda_n^2\right)+(\lambda_n^2-1)}{\lambda_n^2}\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \frac{1}{\kappa}\}} \\
&\leqslant 2\left(\tfrac{\kappa^2}{\kappa^2\lambda_n^2}+1\right)\left(\tfrac{\cosh(2\kappa\lambda_n)+1}{\cosh(2\kappa\lambda_n)-1}\right)\mathbb{1}_{\{\kappa<1\}\cup\{\kappa\geqslant 1, \lambda_n > \frac{1}{\kappa}\}} \\
&\quad + \left(\lambda_n^2 + 3 + \tfrac{4}{3}\kappa^2\lambda_n^4 + \tfrac{8}{3}\kappa^2\lambda_n^2 + \tfrac{4}{3}\kappa^2\right)\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \frac{1}{\kappa}\}} \\
&\leqslant 2\left(\tfrac{1}{\eta^2}+1\right)\left(\tfrac{\cosh(2\eta)+1}{\cosh(2\eta)-1}\right)\mathbb{1}_{\{\kappa<1\}} + \left(\tfrac{7}{3\kappa^2} + \tfrac{17}{3} + \tfrac{4}{3}\kappa^2\right)\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n \leqslant \frac{1}{\kappa}\}} \\
&\quad + \left(2\kappa^2+2\right)\left(\tfrac{\cosh(2)+1}{\cosh(2)-1}\right)\mathbb{1}_{\{\kappa\geqslant 1, \lambda_n > \frac{1}{\kappa}\}} \\
&\leqslant \max\left\{2\left(\tfrac{1}{\eta^2}+1\right)\left(\tfrac{\cosh(2\eta)+1}{\cosh(2\eta)-1}\right), \tfrac{28}{3}, 4\left(\tfrac{\cosh(2)+1}{\cosh(2)-1}\right)\right\}\max\{\kappa^2, 1\} \leqslant 10\max\{\kappa^2, 1\},
\end{aligned}$$

where we used (4.35) to arrive at the second term of the first inequality and (4.32) to arrive

at the first term of the third inequality. Consequently,

$$\max\left\{\max_{0\leqslant n\leqslant N_p}\phi_n,\theta_{N_c},\max_{n\geqslant N_e}\psi_n\right\}\leqslant\max\left\{10\max\{\kappa^4,1\},\tfrac{2}{3}\kappa^2+3,10\max\{\kappa^2,1\}\right\}$$
$$=10\max\{\kappa^4,1\}.$$

Plugging in the above estimate back into (4.34), applying the Parseval's identity, and finally using (4.28), we have (4.12).

Item (3): suppose $\mathcal{B}_2=\mathbf{I}_d$. We note that for $\kappa\geqslant 1$ and $\lambda_n\in(0,\sqrt{\frac{3}{2\kappa^2}+\frac{1}{2}}]$

$$\tfrac{d}{d\lambda_n}\Big(3+\lambda_n^2(\lambda_n^2-1)\kappa^4+3(1-\lambda_n^2)\kappa^2\Big)=2\lambda_n((2\lambda_n^2-1)\kappa^4-3\kappa^2)\leqslant 0.$$

By item (3) of Lemma 4.7 with $\alpha=0$, we obtain for $\kappa>0$ and $n\leqslant N_p$,

$$\begin{aligned}\phi_n&=\tfrac{\kappa^2\left((1+\lambda_n^2)-(1-\lambda_n^2)^2\frac{\sin(2\kappa\lambda_n)}{2\kappa\lambda_n}\right)}{1-(1-\lambda_n^2)\cos^2(\kappa\lambda_n)}\leqslant 2\mathbb{1}_{\{\kappa<1\}}+\tfrac{\kappa^2((1+\lambda_n^2)+(1-\lambda_n^2)^2)}{\lambda_n^2}\mathbb{1}_{\{\kappa\geqslant 1,\frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}}\\&\quad+\kappa^2\left(\tfrac{(1-\lambda_n^2)\left(1-(1-\lambda_n^2)\left(1-\frac{2}{3}\kappa^2\lambda_n^2\right)\right)}{1-(1-\lambda_n^2)\left(1-\kappa^2\lambda_n^2+\frac{1}{3}\kappa^4\lambda_n^4\right)}+\tfrac{2\lambda_n^2}{1-(1-\lambda_n^2)\left(1-\kappa^2\lambda_n^2+\frac{1}{3}\kappa^4\lambda_n^4\right)}\right)\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{\pi}{4\kappa}\}}\\&\leqslant 2\mathbb{1}_{\{\kappa<1\}}+\kappa^2\left((1-\lambda_n^2)+\tfrac{6}{3+\lambda_n^2(\lambda_n^2-1)\kappa^4+3(1-\lambda_n^2)\kappa^2}\right)\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{\pi}{4\kappa}\}}\\&\quad+\kappa^2\left(\lambda_n^2-1+\tfrac{2}{\lambda_n^2}\right)\mathbb{1}_{\{\kappa\geqslant 1,\frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}}\\&\leqslant 2\mathbb{1}_{\{\kappa<1\}}+\kappa^2\left(1+\tfrac{1536}{(768-16\pi^2)\kappa^2+(768+\pi^4-48\pi^2)}\right)\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{\pi}{4\kappa}\}}+\tfrac{32}{\pi^2}\kappa^4\mathbb{1}_{\{\kappa\geqslant 1,\frac{\pi}{4\kappa}<\lambda_n\leqslant 1\}}\\&\leqslant\max\left\{2,1+\tfrac{1536}{(768-16\pi^2)+(768+\pi^4-48\pi^2)},\tfrac{32}{\pi^2}\right\}\max\{\kappa^4,1\}\leqslant 4\max\{\kappa^4,1\},\end{aligned}$$

where we respectively used (4.31) and (4.30) to obtain the first and second terms of the first inequality, and substitute $\lambda_n=\frac{\pi}{4\kappa}$ into the second term of the third inequality. Next, for $n\geqslant N_e$ and $\lambda_n\in(0,\infty)$, we have

$$\kappa^2\|X_n\|_{0,\mathcal{I}}^2=\kappa^2\tfrac{\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}(1+\lambda_n^2)-(1-\lambda_n^2)}{\cosh(2\kappa\lambda_n)(1+\lambda_n^2)-(1-\lambda_n^2)}\leqslant\kappa^2.$$

Next, for all $\lambda_n\in\mathbb{R}$, we note that $\lambda_n^2(1+\lambda_n^2)^{-\frac{3}{2}}\leqslant\frac{2\sqrt{3}}{9}$,

$$\tfrac{\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}-1}{\cosh(2\kappa\lambda_n)-1}\leqslant\lim_{\kappa\lambda_n\to 0}\tfrac{\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}-1}{\cosh(2\kappa\lambda_n)-1}\leqslant\tfrac{1}{3},\quad\text{and}\quad\tfrac{\lambda_n(\frac{3}{2}+\lambda_n^2)}{(1+\lambda_n^2)^{\frac{3}{2}}}\leqslant\lim_{\lambda_n\to\infty}\tfrac{\lambda_n(\frac{3}{2}+\lambda_n^2)}{(1+\lambda_n^2)^{\frac{3}{2}}}\leqslant 1.$$

By item (3) of Lemma 4.7 with $\alpha=0$, we obtain for $\kappa>0$ and $n\geqslant N_e$

$$\tfrac{\|X_n'\|_{0,\mathcal{I}}^2+\mu_n^2\|X_n\|_{0,\mathcal{I}}^2}{\mu_n}=\tfrac{\kappa((1+3\lambda_n^2+2\lambda_n^4)\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}+\lambda_n^2-1)}{(1+\lambda_n^2)^{\frac{3}{2}}\cosh(2\kappa\lambda_n)-(1-\lambda_n^2)(1+\lambda_n^2)^{\frac{1}{2}}}$$

$$
\begin{aligned}
&\leqslant \frac{\kappa((1+3\lambda_n^2+2\lambda_n^4)\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}+\lambda_n^2+1)}{(1+\lambda_n^2)^{\frac{3}{2}}(\cosh(2\kappa\lambda_n)-1)}\mathbb{1}_{\{\kappa<1\}} \\
&\quad+\frac{\kappa((1+3\lambda_n^2+2\lambda_n^4)\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}+\lambda_n^2-1)}{(1+\lambda_n^2)^{\frac{1}{2}}((1+\lambda_n^2)\cosh(2\kappa\lambda_n)-(1-\lambda_n^2))}\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{1}{\kappa}\}\cup\{\kappa\geqslant 1,\lambda_n>\frac{1}{\kappa}\}} \\
&\leqslant \left(\frac{(1+3\lambda_n^2+2\lambda_n^4)\sinh(2\kappa\lambda_n)}{2\lambda_n^4(\cosh(2\kappa\lambda_n)-1)}+\frac{1}{(1+\lambda_n^2)^{\frac{1}{2}}(\cosh(2\kappa\lambda_n)-1)}\right)\mathbb{1}_{\{\kappa<1\}} \\
&\quad+\frac{\kappa((1+3\lambda_n^2+2\lambda_n^4)\left(1+\frac{4}{3}\kappa^2\lambda_n^2\right)+\lambda_n^2-1)}{(1+\lambda_n^2)^{\frac{1}{2}}((1+\lambda_n^2)(1+2\kappa^2\lambda_n^2)-(1-\lambda_n^2))}\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{1}{\kappa}\}} \\
&\quad+\left(\frac{\left(\frac{\sinh(2\kappa\lambda_n)}{2\kappa\lambda_n}-1\right)}{(1+\lambda_n^2)^{\frac{3}{2}}(\cosh(2\kappa\lambda_n)-1)}\kappa+\left(\frac{\lambda_n\left(\frac{3}{2}+\lambda_n^2\right)}{(1+\lambda_n^2)^{\frac{3}{2}}}\right)\left(\frac{\sinh(2\kappa\lambda_n)}{\cosh(2\kappa\lambda_n)-1}\right)\right. \\
&\qquad\left.+\frac{\lambda_n^2}{(1+\lambda_n^2)^{\frac{3}{2}}(\cosh(2\kappa\lambda_n)-1)}\kappa\right)\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n>\frac{1}{\kappa}\}} \\
&\leqslant \left(\left(\frac{1}{2\eta^4}+\frac{3}{2\eta^2}+1\right)\left(\frac{\sinh(2\eta)}{\cosh(2\eta)-1}\right)+\frac{1}{\cosh(2\eta)-1}\right)\mathbb{1}_{\{\kappa<1\}} \\
&\quad+\frac{(4\lambda_n^4+6\lambda_n^2+2)\kappa^3+3\kappa(\lambda_n^2+2)}{3+(3\lambda_n^2+3)\kappa^2}\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{1}{\kappa}\}}+\left(\frac{1}{3}\kappa+\frac{\sinh(2)}{\cosh(2)-1}+\frac{2\sqrt{3}}{9(\cosh(2)-1)}\kappa\right)\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n>\frac{1}{\kappa}\}} \\
&\leqslant \frac{(2\eta^4+3\eta+1)\sinh(2\eta)+2\eta^4}{2\eta^4(\cosh(2\eta)-1)}\mathbb{1}_{\{\kappa<1\}}+\frac{(4\lambda_n^4+6\lambda_n^2+2)\kappa^3+3\kappa(\lambda_n^2+2)}{3\kappa^2}\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{1}{\kappa}\}} \\
&\quad+\frac{3(\cosh(2)-1)+9\sinh(2)+2\sqrt{3}}{9(\cosh(2)-1)}\kappa\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n>\frac{1}{\kappa}\}} \\
&\leqslant \frac{(2\eta^4+3\eta+1)\sinh(2\eta)+2\eta^4}{2\eta^4(\cosh(2\eta)-1)}\mathbb{1}_{\{\kappa<1\}}+\left(\frac{7}{3\kappa^3}+\frac{4}{\kappa}+\frac{2}{3}\kappa\right)\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n\leqslant\frac{1}{\kappa}\}} \\
&\quad+\frac{3(\cosh(2)-1)+9\sinh(2)+2\sqrt{3}}{9(\cosh(2)-1)}\kappa\mathbb{1}_{\{\kappa\geqslant 1,\lambda_n>\frac{1}{\kappa}\}} \\
&\leqslant \max\left\{\frac{(2\eta^4+3\eta+1)\sinh(2\eta)+2\eta^4}{2\eta^4(\cosh(2\eta)-1)},7,\frac{3(\cosh(2)-1)+9\sinh(2)+2\sqrt{3}}{9(\cosh(2)-1)}\right\}\max\{\kappa,1\} \\
&\leqslant 7\max\{\kappa,1\},
\end{aligned}
$$

where we used (4.35) to arrive at the last term of the second inequality and (4.32) to arrive at the first term of the third inequality. Continuing from (4.33), we have

$$
\begin{aligned}
&\sum_{n=0}^{\infty}|\widehat{g_2}(n)|^2\|X_n'\|_{0,\mathcal{I}}^2+\sum_{n=0}^{\infty}|\widehat{g_2}(n)\mu_n|^2\|X_n\|_{0,\mathcal{I}}^2+\kappa^2\sum_{n=0}^{\infty}|\widehat{g_2}(n)|^2\|X_n\|_{0,\mathcal{I}}^2 \\
&\quad=\sum_{n=0}^{N_e-1}|\widehat{g_2}(n)|^2\left(\|X_n'\|_{0,\mathcal{I}}^2+(\mu_n^2+\kappa^2)\|X_n\|_{0,\mathcal{I}}^2\right) \\
&\qquad+\sum_{n=N_e}^{\infty}|\widehat{g_2}(n)|^2\mu_n\left(\frac{\|X_n'\|_{0,\mathcal{I}}^2+\mu_n^2\|X_n\|_{0,\mathcal{I}}^2}{\mu_n}\right)+\sum_{n=N_e}^{\infty}\kappa^2|\widehat{g_2}(n)|^2\|X_n\|_{0,\mathcal{I}}^2 \\
&\quad\leqslant \max\left\{\max_{0\leqslant n\leqslant N_p}\phi_n,\frac{\kappa^2(2\kappa^2+9)}{3(\kappa^2+1)}\right\}\sum_{n=0}^{N_e-1}|\widehat{g_2}(n)|^2+7\max\{\kappa,1\}\sum_{n=N_e}^{\infty}|\widehat{g_2}(n)|^2\mu_n+\kappa^2\sum_{n=N_e}^{\infty}|\widehat{g_2}(n)|^2 \\
&\quad\leqslant 4\max\{\kappa^4,1\}\sum_{n=0}^{\infty}|\widehat{g_2}(n)|^2+7\max\{\kappa,1\}\sum_{n=0}^{\infty}|\widehat{g_2}(n)|^2\mu_n
\end{aligned}
$$

$$\leqslant 7(\max\{\kappa^4,1\}\|g_2\|^2_{0,\Gamma_2}+\max\{\kappa,1\}\|g_2\|^2_{\mathcal{Z}^{1/2}(\Gamma_2)}),$$

where we used our assumptions that $\mathcal{Z}^{1/2}(\Gamma_2)$ and applied the Parseval's identity to arrive at the last line. Finally by (4.28), the stability estimate in (4.13) is proved. □

Proof of Theorem 4.4. All three cases start the same way. Letting $v=u$ in (4.4), we have

$$\|\nabla u\|^2_{0,\Omega}-\kappa^2\|u\|^2_{0,\Omega}-\mathrm{i}\kappa\|u\|^2_{0,\Gamma_R}=\langle f,u\rangle_\Omega, \tag{4.36}$$

where $\Gamma_R=\Gamma_2\cup\Gamma_4$ if $\mathcal{B}_2=\frac{\partial}{\partial\boldsymbol{\nu}}-\mathrm{i}\kappa\mathbf{I}_d$; otherwise, $\Gamma_R=\Gamma_4$ if $\mathcal{B}_2\in\{\mathbf{I}_d,\frac{\partial}{\partial\nu}\}$. Separately considering the real and imaginary parts of (4.36), and applying the Cauchy-Schwarz inequality, we have

$$\|\nabla u\|^2_{0,\Omega}\leqslant\kappa^2\|u\|^2_{0,\Omega}+\|f\|_{0,\Omega}\|u\|_{0,\Omega},\quad \kappa\|u\|^2_{0,\Gamma_R}\leqslant\|f\|_{0,\Omega}\|u\|_{0,\Omega}. \tag{4.37}$$

It is known in [32, Proposition 2.1] that for $u\in H^2(\Omega)$ and $\mathbf{z}\in(C^1(\overline{\Omega}))^2$, the following identity holds

$$\begin{aligned}2\Re\int_\Omega\Delta u(\mathbf{z}\cdot\nabla\overline{u})=2\Re\int_{\partial\Omega}\frac{\partial u}{\partial\nu}(\mathbf{z}\cdot\nabla\overline{u})-2\Re\int_\Omega\nabla u\cdot(\nabla\overline{u}\cdot\nabla)\mathbf{z}\\+\int_\Omega(\nabla\cdot\mathbf{z})|\nabla u|^2-\int_{\partial\Omega}\mathbf{z}\cdot\mathbf{n}|\nabla u|^2.\end{aligned} \tag{4.38}$$

Take $\mathbf{z}=(x-1,0)$. Since $\Delta u=-f-\kappa^2u$, $2\Re(u\overline{u}_x)=(|u|^2)_x$, by applying integration by parts to the left-hand side of (4.38), we have

$$2\Re\int_\Omega\Delta u(\mathbf{z}\cdot\nabla\overline{u})=-2\Re\int_\Omega(x-1)f\overline{u}_x+\kappa^2(\|u\|^2_{0,\Omega}-\|u\|^2_{0,\Gamma_4}).$$

Note that $\frac{\partial u}{\partial\boldsymbol{\nu}}=\mathrm{i}\kappa u$ on Γ_4. If $\mathcal{B}_1u=u=0$ on Γ_1, then $u_x(x,0)=(u(x,0))_x=0$. Similarly, if $\mathcal{B}_3u=u=0$ on Γ_3, then $u_x(x,1)=(u(x,1))_x=0$. Therefore, we have $2\Re\int_{\partial\Omega}\frac{\partial u}{\partial\boldsymbol{\nu}}(\mathbf{z}\cdot\nabla\overline{u})=2\kappa^2\|u\|^2_{0,\Gamma_4}$ for any $\mathcal{B}_1,\mathcal{B}_3\in\{\mathbf{I}_d,\frac{\partial}{\partial\boldsymbol{\nu}}\}$. Hence, fully expanding (4.38), we have

$$\begin{aligned}-2\Re\int_\Omega(x-1)f\overline{u}_x+\kappa^2(\|u\|^2_{0,\Omega}-\|u\|^2_{0,\Gamma_4})\\=2\kappa^2\|u\|^2_{0,\Gamma_4}-2\|u_x\|^2_{0,\Omega}+\|\nabla u\|^2_{0,\Omega}-\|\nabla u\|^2_{0,\Gamma_4}.\\=2\kappa^2\|u\|^2_{0,\Gamma_4}-\|u_x\|^2_{0,\Omega}+\|u_y\|^2_{0,\Omega}-\|\nabla u\|^2_{0,\Gamma_4},\end{aligned}$$

from which, after using (4.36) to replace $\|u_y\|^2_{0,\Omega}$, we obtain

$$2\|u_x\|^2_{0,\Omega}+\|u_y\|^2_{0,\Gamma_4}=\langle f,u\rangle_\Omega+\mathrm{i}\kappa\|u\|^2_{0,\Gamma_R}+2\Re\langle(x-1)f,u_x\rangle_\Omega+2\kappa^2\|u\|^2_{0,\Gamma_4}. \tag{4.39}$$

Taking the real part of (4.39), we have

$$2\|u_x\|_{0,\Omega}^2 \leqslant \|f\|_{0,\Omega}\|u\|_{0,\Omega} + 2\|f\|_{0,\Omega}\|u_x\|_{0,\Omega} + 2\kappa^2\|u\|_{0,\Gamma_4}^2 \tag{4.40}$$

Suppose $\mathcal{B}_2 = \mathbf{I}_d$. Since $\|u\|_{0,\Omega}^2 \leqslant \|u_x\|_{0,\Omega}^2$ (which is proved by noting that $u(x,y) = -\int_x^1 u_x(s,y)ds$ and then estimating an upper bound), (4.40) and the second inequality of (4.37) yield $\|u_x\|_{0,\Omega} \leqslant \frac{1}{2}(3+2\kappa)\|f\|_{0,\Omega}$. I.e., $\|u\|_{0,\Omega} \leqslant \|u_x\|_{0,\Omega} \leqslant \frac{1}{2}(3+2\kappa)\|f\|_{0,\Omega}$. So, by the first inequality of (4.37), we have

$$\begin{aligned}\|\nabla u\|_{0,\Omega}^2 + \kappa^2\|u\|_{0,\Omega}^2 &\leqslant 2\kappa^2\|u\|_{0,\Omega}^2 + \|f\|_{0,\Omega}\|u\|_{0,\Omega} \\ &\leqslant \left(\tfrac{1}{2}\kappa^2(3+2\kappa)^2 + \tfrac{1}{2}(3+2\kappa)\right)\|f\|_{0,\Omega}^2 \leqslant 15\max\{\kappa^4, 1\}\|f\|_{0,\Omega}^2.\end{aligned}$$

Finally, by using (4.28), we obtain (4.14).

Suppose $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$. We have $\|u\|_{0,\Omega}^2 \leqslant 2(\|u\|_{0,\Gamma_4}^2 + \|u_x\|_{0,\Omega}^2)$ (which is proved by noting that $u(x,y) = u(0,y) + \int_0^x u_x(s,y)ds$ and then estimating an upper bound). By the second inequality of (4.37), we have $\|u\|_{0,\Omega}^2 - 2\kappa^{-1}\|f\|_{0,\Omega}\|u\|_{0,\Omega} - 2\|u_x\|_{0,\Omega}^2 \leqslant 0$. This implies that

$$\|u\|_{0,\Omega} \leqslant \kappa^{-1}\|f\|_{0,\Omega} + \tfrac{1}{2}\sqrt{4\kappa^{-2}\|f\|_{0,\Omega}^2 + 8\|u_x\|_{0,\Omega}^2} \leqslant 2\kappa^{-1}\|f\|_{0,\Omega} + \sqrt{2}\|u_x\|_{0,\Omega}. \tag{4.41}$$

By (4.40), the second inequality of (4.37), and (4.41), we have

$$\begin{aligned}2\|u_x\|_{0,\Omega}^2 &\leqslant (2\kappa+1)\|f\|_{0,\Omega}\|u\|_{0,\Omega} + 2\|f\|_{0,\Omega}\|u_x\|_{0,\Omega} \\ &\leqslant 2(2+\kappa^{-1})\|f\|_{0,\Omega}^2 + (2\sqrt{2}\kappa + \sqrt{2} + 2)\|f\|_{0,\Omega}\|u_x\|_{0,\Omega}.\end{aligned}$$

I.e., $2\|u_x\|_{0,\Omega}^2 - (2\sqrt{2}\kappa + \sqrt{2} + 2)\|f\|_{0,\Omega}\|u_x\|_{0,\Omega} - 2(2+\kappa^{-1})\|f\|_{0,\Omega}^2 \leqslant 0$. So,

$$\begin{aligned}\|u_x\|_{0,\Omega} &\leqslant \tfrac{1}{4}(2\sqrt{2}\kappa + \sqrt{2} + 2)\|f\|_{0,\Omega} + \tfrac{1}{4}\sqrt{(2\sqrt{2}\kappa + \sqrt{2} + 2)^2 + 16(2+\kappa^{-1})}\|f\|_{0,\Omega} \\ &\leqslant (\tfrac{1}{2}(2\sqrt{2}\kappa + \sqrt{2} + 2) + (2+\kappa^{-1})^{1/2})\|f\|_{0,\Omega} \\ &\leqslant (\sqrt{2}\kappa + \tfrac{3}{2}\sqrt{2} + 1 + \kappa^{-1/2})\|f\|_{0,\Omega}.\end{aligned} \tag{4.42}$$

By the first inequality of (4.37), (4.41), and (4.42), we have

$$\begin{aligned}\|\nabla u\|_{0,\Omega}^2 + \kappa^2\|u\|_{0,\Omega}^2 &\leqslant 2\kappa^2\|u\|_{0,\Omega}^2 + \|f\|_{0,\Omega}\|u\|_{0,\Omega} \\ &\leqslant 4\kappa^2(4\kappa^{-2}\|f\|_{0,\Omega}^2 + 2\|u_x\|_{0,\Omega}^2) + 2\kappa^{-1}\|f\|_{0,\Omega}^2 + \sqrt{2}\|f\|_{0,\Omega}\|u_x\|_{0,\Omega} \\ &\leqslant 4\kappa^2(4\kappa^{-2} + 2(\sqrt{2}\kappa + \tfrac{3}{2}\sqrt{2} + 1 + \kappa^{-1/2})^2)\|f\|_{0,\Omega}^2 + 2\kappa^{-1}\|f\|_{0,\Omega}^2 \\ &\quad + \sqrt{2}(\sqrt{2}\kappa + \tfrac{3}{2}\sqrt{2} + 1 + \kappa^{-1/2})\|f\|_{0,\Omega}^2\end{aligned}$$

$$\leqslant (155+82\sqrt{2})\max\{\kappa^4,\kappa^{-1}\}\|f\|_{0,\Omega}^2 \leqslant 271\max\{\kappa^4,\kappa^{-1}\}\|f\|_{0,\Omega}^2.$$

Finally, by using (4.28), we obtain (4.15).

Suppose $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}} - \mathrm{i}\kappa\mathbf{I}_d$. Keeping in mind that the second inequality of (4.37) implies $\kappa\|u\|_{0,\Gamma_4}^2 \leqslant \kappa\|u\|_{0,\Gamma_2\cup\Gamma_4}^2 \leqslant \|f\|_{0,\Omega}\|u\|_{0,\Omega}$, the proof of this case is identical to the case where $\mathcal{B}_2 = \frac{\partial}{\partial \boldsymbol{\nu}}$. □

In order to prove Proposition 4.5 and Theorem 4.6, we first prove two auxiliary results stated in Lemmas 4.8 and 4.9 below.

Lemma 4.8. *Let $\tilde{\mu}_n, n \in \mathbb{N}_0$ be given in (4.18) and define $\tilde{\lambda}_n := \sqrt{|1-\tilde{\mu}_n^2\kappa^{-2}|}$ for $n \in \mathbb{N}_0$. For $\mathcal{B}_1 = \mathcal{B}_3$,*

$$\inf_{j\in\mathbb{Z}} |\kappa\tilde{\lambda}_n - j\pi| \geqslant \tfrac{\frac{1}{8}\pi}{1+2\pi^{-1}\kappa\tilde{\lambda}_n}, \qquad \forall\, n \in \mathbb{N}_0, \tag{4.43}$$

and for $\mathcal{B}_1 \neq \mathcal{B}_3$,

$$\inf_{j\in\mathbb{Z}} |\kappa\tilde{\lambda}_n - (j+\tfrac{1}{2})\pi| \geqslant \tfrac{\frac{1}{8}\pi}{1+2\pi^{-1}\kappa\tilde{\lambda}_n}, \qquad \forall\, n \in \mathbb{N}_0. \tag{4.44}$$

Proof. We first consider $\mathcal{B}_1 = \mathcal{B}_3$. Let j be the unique integer such that $j \leqslant \frac{\kappa\tilde{\lambda}_n}{\pi} < j+1$. Then it is obvious that $\inf_{m\in\mathbb{Z}} |\kappa\tilde{\lambda}_n - m\pi| = \min(|\kappa\tilde{\lambda}_n - j\pi|, |\kappa\tilde{\lambda}_n - (j+1)\pi)|)$. If $d_0 \in [0, \frac{1}{8}\pi^2]$, then $-\frac{1}{4}\pi^2 \pm d_0 \in [-\frac{3}{8}\pi^2, -\frac{1}{8}\pi^2]$ and $\tilde{\mu}_n = (n+\frac{1}{2})\pi$ by (4.18). By the definition of d_0, we have $\kappa^2 = m\pi^2 \pm d_0$ for some $m \in \mathbb{Z}$. By $\tilde{\mu}_n = (n+\frac{1}{2})\pi$ in (4.18) and $(\kappa\tilde{\lambda}_n)^2 = |\kappa^2 - \tilde{\mu}_n^2| = |(m-n^2-n)\pi^2 + (-\frac{1}{4}\pi^2 \pm d_0)|$, we have

$$|\kappa\tilde{\lambda}_n - j\pi| = \tfrac{|(\kappa\tilde{\lambda}_n)^2 - j^2\pi^2|}{\kappa\tilde{\lambda}_n + j\pi} = \tfrac{|(N-j^2)\pi^2 \pm (-\frac{1}{4}\pi^2 \pm d_0)|}{\kappa\tilde{\lambda}_n + j\pi} \geqslant \tfrac{\frac{1}{8}\pi^2}{\kappa\tilde{\lambda}_n + j\pi} \geqslant \tfrac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1},$$

where $N := m - n^2 - n$ for $\kappa^2 \geqslant \tilde{\mu}_n^2$ or $N := n^2 + n - m$ for $\kappa^2 < \tilde{\mu}_n^2$, and

$$|\kappa\tilde{\lambda}_n - (j+1)\pi| = \tfrac{|(\kappa\tilde{\lambda}_n)^2 - (j+1)^2\pi^2|}{\kappa\tilde{\lambda}_n + (j+1)\pi} = \tfrac{|((j+1)^2 - N)\pi^2 \pm (-\frac{1}{4}\pi^2 \pm d_0)|}{\kappa\tilde{\lambda}_n + (j+1)\pi} \geqslant \tfrac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1},$$

where we used $j\pi \leqslant \kappa\tilde{\lambda}_n$ and hence $\kappa\tilde{\lambda}_n + j\pi \leqslant \kappa\tilde{\lambda}_n + (j+1)\pi \leqslant 2\kappa\tilde{\lambda}_n + \pi$.

If $d_0 \notin [0, \frac{1}{8}\pi^2]$, then $d_1 = \frac{1}{2}\pi^2 - d_0 \in [0, \frac{3}{8}\pi^2]$ and hence, $\frac{1}{2}\pi^2 \pm d_1 \in [\frac{1}{8}\pi^2, \frac{7}{8}\pi^2]$. By the definition of d_1, we have $\kappa^2 = (m+\frac{1}{2})\pi^2 \pm d_1$ for some $m \in \mathbb{Z}$. By $\tilde{\mu}_n = n\pi$ in (4.18) and $(\kappa\tilde{\lambda}_n)^2 = |\kappa^2 - \tilde{\mu}_n^2| = |(m-n^2)\pi^2 + (\frac{1}{2}\pi^2 \pm d_1)|$, we have

$$|\kappa\tilde{\lambda}_n - j\pi| = \tfrac{|(\kappa\tilde{\lambda}_n)^2 - j^2\pi^2|}{\kappa\tilde{\lambda}_n + j\pi} = \tfrac{|(N-j^2)\pi^2 \pm (\frac{1}{2}\pi^2 \pm d_1)|}{\kappa\tilde{\lambda}_n + j\pi} \geqslant \tfrac{\frac{1}{8}\pi^2}{\kappa\tilde{\lambda}_n + j\pi} \geqslant \tfrac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1},$$

where $N := m - n^2$ for $\kappa^2 \geqslant \tilde{\mu}_n^2$ or $N := n^2 - m$ for $\kappa^2 < \tilde{\mu}_n^2$, and

$$|\kappa\tilde{\lambda}_n - (j+1)\pi| = \tfrac{|(\kappa\tilde{\lambda}_n)^2 - (j+1)^2\pi^2|}{\kappa\tilde{\lambda}_n + (j+1)\pi} = \tfrac{|((j+1)^2 - N)\pi^2 \pm (\frac{1}{2}\pi^2 \pm d_1)|}{\kappa\tilde{\lambda}_n + (j+1)\pi} \geqslant \tfrac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1}.$$

This proves (4.43) for the case $\mathcal{B}_1 = \mathcal{B}_3$.

We now consider $\mathcal{B}_1 \neq \mathcal{B}_3$. Consider the unique integer j such that $j \leqslant \frac{\kappa\tilde{\lambda}_n}{\pi} + \frac{1}{2} < j+1$. Then it is obvious that $(j-\frac{1}{2})\pi \leqslant \kappa\tilde{\lambda}_n < (j+\frac{1}{2})\pi$ and $\inf_{m\in\mathbb{Z}} |\kappa\tilde{\lambda}_n - (m+\frac{1}{2})\pi| = \min(|\kappa\tilde{\lambda}_n - (j-\frac{1}{2})\pi|, |\kappa\tilde{\lambda}_n - (j+\frac{1}{2})\pi|)$. If $d_0 \in [0, \frac{1}{8}\pi^2] \cup [\frac{3}{8}\pi^2, \frac{1}{2}\pi^2]$, then $-\frac{1}{4}\pi^2 \pm d_0 \in [-\frac{6}{8}\pi^2, -\frac{5}{8}\pi^2] \cup [-\frac{3}{8}\pi^2, -\frac{1}{8}\pi^2] \cup [\frac{1}{8}\pi^2, \frac{2}{8}\pi^2]$. By the definition of d_0, we have $\kappa^2 = m\pi^2 \pm d_0$ for some $m \in \mathbb{Z}$. Therefore, by $\tilde{\mu}_n = n\pi$ in (4.18) and $(\kappa\tilde{\lambda}_n)^2 = |\kappa^2 - \tilde{\mu}_n^2| = |(m-n^2)\pi^2 \pm d_0|$, we have

$$|\kappa\tilde{\lambda}_n - (j-\tfrac{1}{2})\pi| = \frac{|(\kappa\tilde{\lambda}_n)^2 - (j-\frac{1}{2})^2\pi^2|}{\kappa\tilde{\lambda}_n + (j-\frac{1}{2})\pi} = \frac{|(N-j^2+j)\pi^2 + (-\frac{1}{4}\pi^2 \pm d_0)|}{\kappa\tilde{\lambda}_n + (j-\frac{1}{2})\pi} \geqslant \frac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1},$$

where $N := m - n^2$ for $\kappa^2 \geqslant \tilde{\mu}_n^2$ or $N := n^2 - m$ for $\kappa^2 < \tilde{\mu}_n^2$, and

$$|\kappa\tilde{\lambda}_n - (j+\tfrac{1}{2})\pi| = \frac{|(\kappa\tilde{\lambda}_n)^2 - (j+\frac{1}{2})^2\pi^2|}{\kappa\tilde{\lambda}_n + (j+\frac{1}{2})\pi} = \frac{|(j^2+j-N)\pi^2 - (-\frac{1}{4}\pi^2 \pm d_0)|}{\kappa\tilde{\lambda}_n + (j+\frac{1}{2})\pi} \geqslant \frac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1},$$

where we used $(j-\frac{1}{2})\pi \leqslant \kappa\tilde{\lambda}_n$ and hence $\kappa\tilde{\lambda}_n + (j-\frac{1}{2})\pi \leqslant \kappa\tilde{\lambda}_n + (j+\frac{1}{2})\pi \leqslant 2\kappa\tilde{\lambda}_n + \pi$.

If $d_0 \notin [0, \frac{1}{8}\pi^2] \cup [\frac{3}{8}\pi^2, \frac{1}{2}\pi^2]$, then $d_1 = \frac{1}{2}\pi^2 - d_0 \in [\frac{1}{8}\pi^2, \frac{3}{8}\pi^2]$ and consequently, $\pm(\frac{1}{4}\pi^2 \pm d_1) \in [-\frac{5}{8}\pi^2, -\frac{3}{8}\pi^2] \cup [-\frac{1}{8}\pi^2, \frac{1}{8}\pi^2] \cup [\frac{3}{8}\pi^2, \frac{5}{8}\pi^2]$, from which we obtain $\frac{1}{4}\pi^2 \pm (\frac{1}{4}\pi^2 \pm d_1) \in [-\frac{3}{8}\pi^2, -\frac{1}{8}\pi^2] \cup [\frac{1}{8}\pi^2, \frac{3}{8}\pi^2] \cup [\frac{5}{8}\pi^2, \frac{7}{8}\pi^2]$. By the definition of d_1, we have $\kappa^2 = (m+\frac{1}{2})\pi^2 \pm d_1$ for some $m \in \mathbb{Z}$. By $\tilde{\mu}_n = (n+\frac{1}{2})\pi$ in (4.18) and $(\kappa\tilde{\lambda}_n)^2 = |\kappa^2 - \tilde{\mu}_n^2| = |(m-n^2-n)\pi^2 + \frac{1}{4}\pi^2 \pm d_1|$, we have

$$|\kappa\tilde{\lambda}_n - (j-\tfrac{1}{2})\pi| = \frac{|(\kappa\tilde{\lambda}_n)^2 - (j-\frac{1}{2})^2\pi^2|}{\kappa\tilde{\lambda}_n + (j-\frac{1}{2})\pi} = \frac{|(N-j^2+j)\pi^2 - [\frac{1}{4}\pi^2 \pm (\frac{1}{4}\pi^2 \pm d_1)]|}{\kappa\tilde{\lambda}_n + (j+1/2)\pi} \geqslant \frac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1},$$

where $N := m - n^2 - n$ for $\kappa^2 \geqslant \tilde{\mu}_n^2$ or $N := n^2 + n - m$ for $\kappa^2 < \tilde{\mu}_n^2$, and

$$|\kappa\tilde{\lambda}_n - (j+\tfrac{1}{2})\pi| = \frac{|(\kappa\tilde{\lambda}_n)^2 - (j+\frac{1}{2})^2\pi^2|}{\kappa\tilde{\lambda}_n + (j+\frac{1}{2})\pi} = \frac{|(j^2+j-N)\pi^2 + [\frac{1}{4}\pi^2 \pm (\frac{1}{4}\pi^2 \pm d_1)]|}{\kappa\tilde{\lambda}_n + (j+\frac{1}{2})\pi} \geqslant \frac{\frac{1}{8}\pi}{2\pi^{-1}\kappa\tilde{\lambda}_n + 1}.$$

This proves (4.44) for the case $\mathcal{B}_1 \neq \mathcal{B}_3$. □

Lemma 4.9. *Consider the problem (4.20). Define*

$$\tilde{\lambda}_n := \sqrt{\left|1 - \tfrac{\tilde{\mu}_n^2}{\kappa^2}\right|} \quad \text{and} \quad \mathring{\tilde{\lambda}}_n := \begin{cases} \tilde{\lambda}_n & \text{if } \tilde{\mu}_n^2 \leqslant \kappa^2, \\ i\tilde{\lambda}_n & \text{if } \tilde{\mu}_n^2 > \kappa^2, \end{cases} \quad \forall n \in \mathbb{N}_0.$$

where $\tilde{\mu}_n$ for $n \in \mathbb{N}_0$ is given in (4.18).

(1) Suppose that $\mathcal{B}_1\tilde{Y}_n(0) = -\tilde{Y}_n'(0) = 1$ and $\mathcal{B}_3\tilde{Y}_n(1) = \alpha\tilde{Y}_n'(1) + (1-\alpha)\tilde{Y}_n(1) = 0$ with

$\alpha = \{0, 1\}$. *Then, the solutions* $\{\tilde{Y}_n\}_{n\in\mathbb{N}_0}$ *to the problem* (4.20) *satisfy*

$$\tilde{Y}_n(y) = \begin{cases} \frac{-\alpha\cos(\kappa\mathring{\tilde{\lambda}}_n(y-1))-(1-\alpha)\sin(\kappa\mathring{\tilde{\lambda}}_n(y-1))}{\kappa\mathring{\tilde{\lambda}}_n(\alpha\sin(k\mathring{\tilde{\lambda}}_n)+(1-\alpha)\cos(\kappa\mathring{\tilde{\lambda}}_n))}, & \text{if} \quad \tilde{\mu}_n^2 \neq \kappa^2, \\ \alpha(\frac{1}{2}y^2 - y) - (1-\alpha)(y-1), & \text{if} \quad \tilde{\mu}_n^2 = \kappa^2. \end{cases}$$

If $\mathcal{B}_1\tilde{Y}_n(0) = -\alpha\tilde{Y}_n'(0) + (1-\alpha)\tilde{Y}_n(0) = 0$ *and* $\mathcal{B}_3\tilde{Y}_n(1) = \tilde{Y}_n'(1) = 1$ *with* $\alpha = \{0, 1\}$, *then the solutions to* (4.20) *are given above with* y *replaced by* $1 - y$. *Moreover, for both cases, the norms* $\|\tilde{Y}_n\|_{0,\mathcal{I}}$ *and* $\|\tilde{Y}_n'\|_{0,\mathcal{I}}$ *are given by*

$$\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 = \begin{cases} \frac{\kappa\mathring{\tilde{\lambda}}_n+(-1)^{1-\alpha}\sin(\kappa\mathring{\tilde{\lambda}}_n)\cos(\kappa\mathring{\tilde{\lambda}}_n)}{2(\kappa\mathring{\tilde{\lambda}}_n)^3((1-\alpha)\cos^2(\kappa\mathring{\tilde{\lambda}}_n)+\alpha\sin^2(\kappa\mathring{\tilde{\lambda}}_n))}, & \text{if} \quad \tilde{\mu}_n^2 \neq \kappa^2, \\ \frac{1}{3}(1-\alpha) + \frac{2}{15}\alpha, & \text{if} \quad \tilde{\mu}_n^2 = \kappa^2, \end{cases}$$

$$\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 = \begin{cases} \frac{\kappa\mathring{\tilde{\lambda}}_n+(-1)^{\alpha}\sin(\kappa\mathring{\tilde{\lambda}}_n)\cos(\kappa\mathring{\tilde{\lambda}}_n)}{2\kappa\mathring{\tilde{\lambda}}_n((1-\alpha)\cos^2(\kappa\mathring{\tilde{\lambda}}_n)+\alpha\sin^2(\kappa\mathring{\tilde{\lambda}}_n))}, & \text{if} \quad \tilde{\mu}_n^2 \neq \kappa^2, \\ (1-\alpha) + \frac{1}{3}\alpha, & \text{if} \quad \tilde{\mu}_n^2 = \kappa^2. \end{cases}$$

(2) *Suppose that* $\mathcal{B}_1\tilde{Y}_n(0) = \tilde{Y}_n(0) = 1$ *and* $\mathcal{B}_3\tilde{Y}_n(1) = \alpha\tilde{Y}_n'(1) + (1-\alpha)\tilde{Y}_n(1) = 0$ *with* $\alpha = \{0, 1\}$. *Then, the solutions* $\{\tilde{Y}_n\}_{n\in\mathbb{N}_0}$ *to the problem* (4.20) *satisfy*

$$\tilde{Y}_n(y) = \frac{\alpha\cos(\kappa\mathring{\tilde{\lambda}}_n(y-1))+(1-\alpha)\sin(\kappa\mathring{\tilde{\lambda}}_n(1-y))}{\alpha\cos(\kappa\mathring{\tilde{\lambda}}_n)+(1-\alpha)\sin(\kappa\mathring{\tilde{\lambda}}_n)}.$$

If $\mathcal{B}_1\tilde{Y}_n(0) = -\alpha\tilde{Y}_n'(0) + (1-\alpha)\tilde{Y}_n(0) = 0$ *and* $\mathcal{B}_3\tilde{Y}_n(1) = \tilde{Y}_n(1) = 1$ *with* $\alpha = \{0, 1\}$, *then the solutions to* (4.20) *are given above with* y *replaced by* $1 - y$. *Moreover, for both cases, the norms* $\|\tilde{Y}_n\|_{0,\mathcal{I}}$ *and* $\|\tilde{Y}_n'\|_{0,\mathcal{I}}$ *are given by*

$$\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 = \frac{\kappa\mathring{\tilde{\lambda}}_n+(-1)^{1-\alpha}\sin(\kappa\mathring{\tilde{\lambda}}_n)\cos(\kappa\mathring{\tilde{\lambda}}_n)}{2\kappa\mathring{\tilde{\lambda}}_n((1-\alpha)\sin^2(\kappa\mathring{\tilde{\lambda}}_n)+\alpha\cos^2(\kappa\mathring{\tilde{\lambda}}_n))}, \quad \|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 = \frac{\kappa\mathring{\tilde{\lambda}}_n((-1)^{\alpha}\sin(\kappa\mathring{\tilde{\lambda}}_n)\cos(\kappa\mathring{\tilde{\lambda}}_n)+\kappa\mathring{\tilde{\lambda}}_n)}{2((1-\alpha)\sin^2(\kappa\mathring{\tilde{\lambda}}_n)+\alpha\cos^2(\kappa\mathring{\tilde{\lambda}}_n))}.$$

Proof. The above solutions and norms can be obtained from direct calculations (similar to the proof of Lemma 4.7). □

Proof of Proposition 4.5. Given that (4.18) holds, we know by Lemma 4.8 that each solution stated in Lemma 4.9 is well defined and hence each term of $\tilde{u}$ in (4.19) is well defined. It is also straightforward to see that (4.19) satisfies (4.16).

By (4.6), we recall that $\{\tilde{X}_n\}_{n\in\mathbb{N}_0}$, $\{\tilde{X}_n'\}_{n\in\mathbb{N}_0}$, $\{\tilde{X}_n''\}_{n\in\mathbb{N}_0}$ are orthogonal systems in $L_2(\mathcal{I})$. Also, $\|\tilde{X}_n'\|_{0,\mathcal{I}} = \tilde{\mu}_n$ and $\|\tilde{X}_n''\|_{0,\mathcal{I}} = \tilde{\mu}_n^2$. Let $S_M(\tilde{u})$ denote a partial sum of $\tilde{u}$ with the Mth term as its last term. Let $\tilde{u}_x$ be the first partial derivative of $\tilde{u}$ in the x direction obtained by term-by-term differentiation. Define $\tilde{u}_y$, $\tilde{u}_{xx}$, and $\tilde{u}_{xy}$ similarly.

Suppose that $\mathcal{B}_1 = \mathbf{I}_d$ and $\mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$. We can pick $M \in \mathbb{N}$ such that for all $n \geqslant M$, we have $\tilde{\mu}_n^2 > \kappa^2$, $\coth(\kappa\tilde{\lambda}_n) \leqslant 2$, and $\frac{1}{4} \leqslant |1 - \kappa^2\tilde{\mu}_n^{-2}|$ such that item (2) of Lemma 4.9 with $\alpha = 0$ (i.e., $\mathcal{B}_3 = \mathbf{I}_d$) implies

$$\begin{aligned}
\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 &= \tfrac{\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) - \kappa\tilde{\lambda}_n}{2\kappa\tilde{\lambda}_n \sinh^2(\kappa\tilde{\lambda}_n)} \leqslant \tfrac{\coth(\kappa\tilde{\lambda}_n)}{2\kappa\tilde{\lambda}_n} \leqslant (\kappa\tilde{\lambda}_n)^{-1} = \tilde{\mu}_n^{-1}\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{-1/2} \leqslant 2\tilde{\mu}_n^{-1}, \\
\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 &= \tfrac{\kappa\tilde{\lambda}_n(\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) + \kappa\tilde{\lambda}_n)}{2\sinh^2(\kappa\tilde{\lambda}_n)} \leqslant \tfrac{3}{2}\kappa\tilde{\lambda}_n = \tfrac{3}{2}\tilde{\mu}_n\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{1/2} \leqslant \tfrac{3}{2}\tilde{\mu}_n,
\end{aligned}$$

and item (2) of Lemma 4.9 with $\alpha = 1$ (i.e., $\mathcal{B}_3 = \frac{\partial}{\partial \boldsymbol{\nu}}$) implies

$$\begin{aligned}
\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 &= \tfrac{\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) + \kappa\tilde{\lambda}_n}{2\kappa\tilde{\lambda}_n \cosh^2(\kappa\tilde{\lambda}_n)} \leqslant \tfrac{\tanh(\kappa\tilde{\lambda}_n)}{\kappa\tilde{\lambda}_n} \leqslant \tilde{\mu}_n^{-1}\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{-1/2} \leqslant 2\tilde{\mu}_n^{-1}, \\
\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 &= \tfrac{\kappa\tilde{\lambda}_n(\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) - \kappa\tilde{\lambda}_n)}{2\cosh^2(\kappa\tilde{\lambda}_n)} \leqslant \tfrac{1}{2}\kappa\tilde{\lambda}_n = \tfrac{1}{2}\tilde{\mu}_n\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{1/2} \leqslant \tfrac{1}{2}\tilde{\mu}_n.
\end{aligned}$$

That is given $\mathcal{B}_1 = \mathbf{I}_d$ and $\mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$, $\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 \leqslant (\kappa\tilde{\lambda}_n)^{-1} \leqslant 2\tilde{\mu}_n^{-1}$ and $\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 \leqslant \kappa\tilde{\lambda}_n \leqslant \frac{3}{2}\tilde{\mu}_n$ for all $n \geqslant M$. Furthermore,

$$\begin{aligned}
\|\tilde{u}_x - (S_M(\tilde{u}))_x\|_{0,\Omega} &= \sum_{n=M}^{\infty} |\widehat{g}_1(n)\tilde{\mu}_n|^2\|\tilde{Y}_n\|_{0,\mathcal{I}}^2, \quad \|\tilde{u}_{xx} - (S_M(\tilde{u}))_{xx}\|_{0,\Omega} = \sum_{n=M}^{\infty} |\widehat{g}_1(n)\tilde{\mu}_n^2|^2\|\tilde{Y}_n\|_{0,\mathcal{I}}^2, \\
\|\tilde{u}_y - (S_M(\tilde{u}))_y\|_{0,\Omega} &= \sum_{n=M}^{\infty} |\widehat{g}_1(n)|^2\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2, \quad \|\tilde{u}_{xy} - (S_M(\tilde{u}))_{xy}\|_{0,\Omega} = \sum_{n=M}^{\infty} |\widehat{g}_1(n)\tilde{\mu}_n|^2\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2,
\end{aligned}$$

Since $g_1 \in \mathcal{Z}^{3/2}(\Gamma_1)$ for $\mathcal{B}_1 = \mathbf{I}_d$, the above inequalities all tend to zero as $M \to \infty$.

Now suppose that $\mathcal{B}_1 = \frac{\partial}{\partial \boldsymbol{\nu}}$ and $\mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$. We can pick $M \in \mathbb{N}_0$ such that for all $n \geqslant M$, we have $\tilde{\mu}_n^2 > \kappa^2$, $\sin(2\kappa\tilde{\lambda}_n) \leqslant 2(\cosh(2\kappa\tilde{\lambda}_n) - 1)$, and $2^{-2/3} \leqslant |1 - \kappa^2\tilde{\mu}_n^{-2}|$ such that item (1) of Lemma 4.9 with $\alpha = 0$ (i.e., $\mathcal{B}_3 = \mathbf{I}_d$) implies

$$\begin{aligned}
\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 &= \tfrac{\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) - \kappa\tilde{\lambda}_n}{2(\kappa\tilde{\lambda}_n)^3\cosh^2(\kappa\tilde{\lambda}_n)} \leqslant \tfrac{1}{2}(\kappa\tilde{\lambda}_n)^{-3} \leqslant \tfrac{1}{2}\tilde{\mu}_n^{-3}\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{-3/2} \leqslant \tilde{\mu}_n^{-3}, \\
\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 &= \tfrac{\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) + \kappa\tilde{\lambda}_n}{2\kappa\tilde{\lambda}_n\cosh^2(\kappa\tilde{\lambda}_n)} \leqslant \tilde{\mu}_n^{-1}\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{-1/2} \leqslant 2^{1/3}\tilde{\mu}_n^{-1},
\end{aligned}$$

and item (1) of Lemma 4.9 with $\alpha = 1$ (i.e., $\mathcal{B}_3 = \frac{\partial}{\partial \boldsymbol{\nu}}$) implies

$$\begin{aligned}
\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 &= \tfrac{\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) + \kappa\tilde{\lambda}_n}{2(\kappa\tilde{\lambda}_n)^3\sinh^2(\kappa\tilde{\lambda}_n)} \leqslant \tfrac{3}{2}(\kappa\tilde{\lambda}_n)^{-3} \leqslant \tfrac{3}{2}\tilde{\mu}_n^{-3}\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{-3/2} \leqslant 3\tilde{\mu}_n^{-3}, \\
\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 &= \tfrac{\sinh(\kappa\tilde{\lambda}_n)\cosh(\kappa\tilde{\lambda}_n) - \kappa\tilde{\lambda}_n}{2\kappa\tilde{\lambda}_n\sinh^2(\kappa\tilde{\lambda}_n)} \leqslant \tilde{\mu}_n^{-1}\left|1 - \kappa^2\tilde{\mu}_n^{-2}\right|^{-1/2} \leqslant 2^{1/3}\tilde{\mu}_n^{-1}.
\end{aligned}$$

That is given $\mathcal{B}_1 = \frac{\partial}{\partial \boldsymbol{\nu}}$ and $\mathcal{B}_3 \in \{\mathbf{I}_d, \frac{\partial}{\partial \boldsymbol{\nu}}\}$, $\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 \leqslant (\kappa\tilde{\lambda}_n)^{-3} \leqslant 3\tilde{\mu}_n^{-3}$ and $\|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 \leqslant (\kappa\tilde{\lambda}_n)^{-1} \leqslant 2^{1/3}\tilde{\mu}_n^{-1}$ for all $n \geqslant M$. Since $g_1 \in \mathcal{Z}^{1/2}(\Gamma_1)$ for $\mathcal{B}_1 = \frac{\partial}{\partial \boldsymbol{\nu}}$, the above inequalities all tend to zero as $M \to \infty$. Therefore, in both of the previously discussed cases, we have shown that

$\tilde{u}_x \in L^2(\Omega)$ and $(S_M(\tilde{u}))_x$ converges to $\tilde{u}_x$ in $L^2(\Omega)$; the same implications also hold for the other three cases involving $\tilde{u}_{xx}$, $\tilde{u}_y$, and $\tilde{u}_{xy}$.

Clearly, $\tilde{u} \in H^1(\Omega)$. We have $\|\tilde{u}\|_{1/2,\partial\Omega} \leqslant C\|\tilde{u}\|_{1,\Omega} < \infty$ for some constant C by the trace inequality [101, Theorem 3.37]. Also, by the multiplicative trace inequality [62, Theorem 1.5.10 and the last inequality of p. 41], we have $\|\tilde{u}_x\|^2_{0,\partial\Omega} \leqslant C\|\tilde{u}_x\|_{1,\Omega}\|\tilde{u}_x\|_{0,\Omega} < \infty$ for some other constant C. □

Proof of Theorem 4.6. Let $N_p := \max\{n \in \mathbb{N}_0 : \tilde{\mu}_n^2 < \kappa^2\}$, $N_c \in \mathbb{N}$ be such that $\tilde{\mu}_{N_c}^2 = \kappa^2$, and $N_e := \min\{n \in \mathbb{N} : \tilde{\mu}_n^2 > \kappa^2\}$. Recall that $\tilde{\lambda}_n := \sqrt{|1 - \tilde{\mu}_n^2\kappa^{-2}|}$ and observe that $\|\tilde{X}_n'\|_{0,\mathcal{I}} = \tilde{\mu}_n$. By (4.6), since $\{\tilde{X}_n\}_{n\in\mathbb{N}_0}$ and $\{\tilde{X}_n'\}_{n\in\mathbb{N}_0}$ are orthogonal systems in $L_2(\mathcal{I})$, we deduce from (4.19) that

$$\begin{aligned}
&\|\nabla\tilde{u}\|^2_{0,\Omega} + \kappa^2\|\tilde{u}\|^2_{0,\Omega} \\
&\quad = \|\sum_{n=0}^{\infty} \widehat{g_1}(n)\tilde{X}_n'\tilde{Y}_n\|^2_{0,\Omega} + \|\sum_{n=0}^{\infty} \widehat{g_1}(n)\tilde{X}_n\tilde{Y}_n'\|^2_{0,\Omega} + \kappa^2\|\sum_{n=0}^{\infty} \widehat{g_1}(n)\tilde{X}_n\tilde{Y}_n\|^2_{0,\Omega} \\
&\quad \leqslant \sum_{n=0}^{\infty} |\widehat{g_1}(n)\tilde{\mu}_n|^2\|\tilde{Y}_n\|^2_{0,\mathcal{I}} + \sum_{n=0}^{\infty} |\widehat{g_1}(n)|^2\|\tilde{Y}_n'\|^2_{0,\mathcal{I}} + \kappa^2\sum_{n=0}^{\infty} |\widehat{g_1}(n)|^2\|\tilde{Y}_n\|^2_{0,\mathcal{I}}.
\end{aligned} \tag{4.45}$$

Regrouping the terms, we have

$$\sum_{n=0}^{\infty} |\widehat{g_1}(n)|^2(\|\tilde{Y}_n'\|^2_{0,\mathcal{I}} + (\tilde{\mu}_n^2+\kappa^2)\|\tilde{Y}_n\|^2_{0,\mathcal{I}}) \leqslant \max\left\{\max_{0\leqslant n\leqslant N_p} \tilde{\phi}_n, \tilde{\theta}_{N_c}, \max_{n\geqslant N_e} \tilde{\psi}_n\right\} \sum_{n=0}^{\infty} |\widehat{g_1}(n)|^2, \tag{4.46}$$

where $\tilde{\phi}_n$, $\tilde{\theta}_{N_c}$, and $\tilde{\psi}_n$ are defined similar to (4.27) with X_n, x, and μ_n being replaced by $\tilde{Y}_n$, y, and $\tilde{\mu}_n$ respectively.

Item (1): Suppose $\mathcal{B}_1 = \frac{\partial}{\partial\boldsymbol{\nu}}$ and $\mathcal{B}_3 = \frac{\partial}{\partial\boldsymbol{\nu}}$. Define $z_n := \kappa\tilde{\lambda}_n$ for $n \in \mathbb{N}_0$. From (4.43), we observe that if the infimum on the left-hand side of the inequality occurs at $j = 0$, then

$$\tfrac{\pi}{2} > z_n \geqslant \tfrac{\pi}{8(1+2\pi^{-1}z_n)} \geqslant \tfrac{\pi}{16}, \quad \text{that is,} \quad z_n \geqslant \tfrac{\pi}{16} \quad \forall n \in \mathbb{N}_0, \kappa > 0. \tag{4.47}$$

Otherwise, if the infimum on the left-hand side of (4.43) occurs at a nonzero j, then $z_n \geqslant \frac{\pi}{2}$. Using item (i) of Lemma 4.9 with $\alpha = 1$, we obtain

$$\tilde{\phi}_n = \tfrac{\kappa^2+(\kappa^2-z_n^2)\frac{\sin(2z_n)}{2z_n}}{z_n^2\sin^2(z_n)}, \quad \tilde{\psi}_n = \tfrac{2((\kappa^2+z_n^2)\frac{\sinh(2z_n)}{2z_n}+\kappa^2)}{z_n^2(\cosh(2z_n)-1)}.$$

For each $n \in \mathbb{N}_0$, let $\gamma_n = \arg\inf_{j\in\mathbb{Z}} |z_n - j\pi|$. By (4.43), we deduce that for all $z_n \in$

$[(\gamma_n - \frac{1}{2})\pi, (\gamma_n + \frac{1}{2})\pi]$,

$$\sin^2(z_n) = \sin^2(z_n - \gamma_n\pi) \geqslant \tfrac{4}{\pi^2}(z_n - \gamma_n\pi)^2 \geqslant \tfrac{1}{16(1+2\pi^{-1}z_n)^2}. \tag{4.48}$$

For $\kappa > 0$, $n \leqslant N_p$, and $z_n \in (0, \kappa]$, we have

$$\begin{aligned}\tilde{\phi}_n &\leqslant \tfrac{2\kappa^2 - z_n^2}{z_n^2 \sin^2(z_n)} \leqslant 32\tfrac{\kappa^2}{z_n^2}(1 + 2\pi^{-1}z_n)^2 \leqslant 32\kappa^2\left(\tfrac{1}{z_n^2} + \tfrac{4}{z_n\pi} + \tfrac{4}{\pi^2}\right) \\ &\leqslant 32\left(\tfrac{256}{\pi^2} + \tfrac{64}{\pi^2} + \tfrac{4}{\pi^2}\right)\max\{\kappa^2, 1\} \leqslant \tfrac{10368}{\pi^2}\max\{\kappa^2, 1\} \leqslant 1051\max\{\kappa^2, 1\},\end{aligned}$$

where we used (4.48) to arrive at the second inequality, and applied (4.47) to arrive at the fourth inequality. Next for $\kappa > 0$, $n \geqslant N_e$, and $z_n \in (0, \infty)$, we have

$$\tilde{\psi}_n \leqslant 2\left(\tfrac{\kappa^2}{z_n^2} + 1\right)\left(\tfrac{\cosh(2z_n)+1}{\cosh(2z_n)-1}\right) \leqslant 2\left(\tfrac{256}{\pi^2} + 1\right)\left(\tfrac{\cosh(\frac{\pi}{8})+1}{\cosh(\frac{\pi}{8})-1}\right)\max\{\kappa^2, 1\} \leqslant 1434\max\{\kappa^2, 1\},$$

where we used (4.47) to arrive at the second inequality. Consequently,

$$\begin{aligned}&\max\left\{\max_{0\leqslant n\leqslant N_p}\tilde{\phi}_n, \tilde{\theta}_{N_c}, \max_{n\geqslant N_e}\tilde{\psi}_n\right\} \\ &\qquad\leqslant \max\left\{1051\max\{\kappa^2, 1\}, \tfrac{4}{15}\kappa^2 + \tfrac{1}{3}, 1434\max\{\kappa^2, 1\}\right\} = 1434\max\{\kappa^2, 1\}.\end{aligned}$$

Applying the Parseval's identity to (4.46) and finally using (4.28), we have (4.22).

Item (1): suppose $\mathcal{B}_1 = \frac{\partial}{\partial\boldsymbol{\nu}}$ and $\mathcal{B}_3 = \mathbf{I}_d$. Using item (i) of Lemma 4.9 with $\alpha = 0$, we obtain

$$\tilde{\phi}_n = \tfrac{\kappa^2 - (\kappa^2 - z_n^2)\frac{\sin(2z_n)}{2z_n}}{z_n^2\cos^2(z_n)}, \quad \tilde{\psi}_n = \tfrac{\kappa^2(\sinh(2z_n) - 2z_n) + z_n^2\sinh(2z_n)}{z_n^3(\cosh(2z_n)+1)}.$$

To obtain an upper bound for $\tilde{\phi}_n$, we shall list several observations, which are used in its estimation. For each $n \in \mathbb{N}_0$, let $\gamma_n = \arg\inf_{j\in\mathbb{Z}}|z_n - (j + \frac{1}{2})\pi|$. By (4.44), we deduce that for all $z_n \in [\gamma_n\pi, (\gamma_n + 1)\pi]$,

$$\cos^2(z_n) = \sin^2(z_n - (\gamma_n + \tfrac{1}{2})\pi) \geqslant \tfrac{4}{\pi^2}(z_n - (\gamma_n + \tfrac{1}{2})\pi)^2 \geqslant \tfrac{1}{16(1+2\pi^{-1}z_n)^2}. \tag{4.49}$$

Also, for all $z_n \in (0, 1]$, we have

$$\tfrac{d}{dz_n}\left(\tfrac{\left(1 - \frac{\sin(2z_n)}{2z_n}\right)}{z_n^2\cos^2(z_n)}\right) = \tfrac{3\cos(z_n)z_n\left(\frac{\sin(2z_n)}{2z_n} + \frac{2}{3}z_n\tan(z_n) - 1\right)}{z_n^4\cos^3(z_n)} \geqslant \tfrac{3\left(1 - \frac{2}{3}z_n^2 + \frac{2}{3}(z_n^2 + \frac{1}{3}z_n^4) - 1\right)}{z_n^3\cos^2(z_n)} \geqslant 0,$$

where we used (4.30) and $z_n^2 + \frac{1}{3}z_n^4 \leqslant z_n\tan(z_n)$ for all $z_n \in (0, 1]$. Now, for $\kappa > 0$, $n \leqslant N_p$,

and $z_n \in (0, \kappa]$, we have

$$\begin{aligned}
\tilde{\phi}_n &\leqslant \left(\tfrac{1-\frac{\sin(2z_n)}{2z_n}}{z_n^2\cos^2(z_n)}\kappa^2 + \tfrac{\sin(2z_n)}{2z_n\cos^2(z_n)} \right) \mathbb{1}_{\{z_n \leqslant 1\}} + \tfrac{2}{z_n^2\cos^2(z_n)}\kappa^2 \mathbb{1}_{\{1<z_n\leqslant\kappa\}} \\
&\leqslant \left(\tfrac{1-\frac{\sin(2)}{2}}{\cos^2(1)}\kappa^2 + \tfrac{1}{\cos^2(1)} \right) \mathbb{1}_{\{z_n \leqslant 1\}} + \tfrac{32(1+2\pi^{-1}z_n)^2}{z_n^2}\kappa^2 \mathbb{1}_{\{1<z_n\leqslant\kappa\}} \\
&\leqslant \tfrac{4-\sin(2)}{2\cos^2(1)} \max\{\kappa^2, 1\} \mathbb{1}_{\{z_n \leqslant 1\}} + 32(1+2\pi^{-1})^2 \max\{\kappa^2, 1\} \mathbb{1}_{\{1<z_n\leqslant\kappa\}} \\
&\leqslant 86 \max\{\kappa^2, 1\},
\end{aligned}$$

where we used (4.49) to arrive at the second term of the first inequality. Next, note that for all $z_n > 0$

$$\begin{aligned}
\tfrac{d}{dz_n}\left(\tfrac{(\sinh(2z_n)-2z_n)}{2z_n^3(\cosh(2z_n)+1)} \right) &= \tfrac{6\sinh(z_n)\cosh^3(z_n)}{z_n^4(\cosh(2z_n)+1)^2} \left(-1 + \tfrac{2z_n^2}{3\cosh^2(z_n)} + \tfrac{2z_n}{\sinh(2z_n)} \right) \\
&\leqslant \tfrac{6\sinh(z_n)\cosh^3(z_n)}{z_n^4(\cosh(2z_n)+1)^2} \Big((-1 + \tfrac{2}{3}(z_n^2 - z_n^4 + \tfrac{2}{3}z_n^6) + \big(1 - \tfrac{2}{3}z_n^2 + \tfrac{14}{45}z_n^4\big)) \mathbb{1}_{\{z_n \leqslant \frac{2}{\sqrt{5}}\}} \\
&\quad + (-\tfrac{2}{3} + \tfrac{4}{\sqrt{5}}(\sinh(\tfrac{4}{\sqrt{5}}))^{-1}) \mathbb{1}_{\{z_n > \frac{2}{\sqrt{5}}\}} \Big) \leqslant 0,
\end{aligned}$$

where we used $\frac{z_n^2}{\cosh^2(z_n)} \leqslant z_n^2 - z_n^4 + \frac{2}{3}z_n^6$ and $\frac{2z_n}{\sinh(2z_n)} \leqslant 1 - \frac{2}{3}z_n^2 + \frac{14}{45}z_n^4$ for all $z_n \in (0, \frac{2}{\sqrt{5}}]$. Now, for $\kappa > 0$, $n \geqslant N_e$, and $z_n \in (0, \infty)$, we have

$$\tilde{\psi}_n \leqslant \lim_{z_n \to 0} \left(\tfrac{\sinh(2z_n)-2z_n}{z_n^3(\cosh(2z_n)+1)}\kappa^2 + \tfrac{\sinh(2z_n)}{z_n(\cosh(2z_n)+1)} \right) = \tfrac{2}{3}\kappa^2 + 1 \leqslant 2\max\{\kappa^2, 1\}.$$

Consequently,

$$\begin{aligned}
\max\left\{ \max_{0\leqslant n\leqslant N_p} \tilde{\phi}_n, \tilde{\theta}_{N_c}, \max_{n\geqslant N_e} \tilde{\psi}_n \right\} &\leqslant \max\left\{ 86\max\{\kappa^2, 1\}, \tfrac{2}{3}\kappa^2 + 1, 2\max\{\kappa^2, 1\} \right\} \\
&= 86\max\{\kappa^2, 1\}.
\end{aligned}$$

Applying the Parseval's identity to (4.46), and finally using (4.28), we have (4.22).

Item (2): suppose $\mathcal{B}_1 = \mathbf{I}_d$ and $\mathcal{B}_3 = \frac{\partial}{\partial \boldsymbol{\nu}}$. Continuing from (4.45), we have

$$\begin{aligned}
&\sum_{n=0}^{\infty} |\widehat{g_1}(n)\tilde{\mu}_n|^2 \|\tilde{Y}_n\|_{0,\mathcal{I}}^2 + \sum_{n=0}^{\infty} |\widehat{g_1}(n)|^2 \|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 + \kappa^2 \sum_{n=0}^{\infty} |\widehat{g_1}(n)|^2 \|\tilde{Y}_n\|_{0,\mathcal{I}}^2 \\
&= \sum_{n=N_e}^{\infty} k^2 |\widehat{g_1}(n)|^2 \|\tilde{Y}_n\|_{0,\mathcal{I}}^2 + \sum_{n=0}^{N_e-1} |\widehat{g_1}(n)|^2 \left((\tilde{\mu}_n^2 + \kappa^2)\|\tilde{Y}_n\|_{0,\mathcal{I}}^2 + \|\tilde{Y}_n'\|_{0,\mathcal{I}}^2 \right) \\
&\quad + \sum_{n=N_e}^{\infty} |\widehat{g_1}(n)|^2 \tilde{\mu}_n \left(\tfrac{\tilde{\mu}_n^2 \|\tilde{Y}_n\|_{0,\mathcal{I}}^2 + \|\tilde{Y}_n'\|_{0,\mathcal{I}}^2}{\tilde{\mu}_n} \right)
\end{aligned}$$

$$\leqslant \max\left\{\max_{0\leqslant n\leqslant N_p}\tilde{\phi}_n, \tilde{\theta}_{N_c}, \kappa^2 \max_{n\geqslant N_e}\|\tilde{Y}_n\|^2_{0,\mathcal{I}}\right\}\sum_{n=0}^{\infty}|\widehat{g_1}(n)|^2$$
$$+\max_{n\geqslant N_e}\left(\frac{\tilde{\mu}_n^2\|\tilde{Y}_n\|^2_{0,\mathcal{I}}+\|\tilde{Y}_n'\|^2_{0,\mathcal{I}}}{\tilde{\mu}_n}\right)\sum_{n=0}^{\infty}|\widehat{g_1}(n)|^2\tilde{\mu}_n. \tag{4.50}$$

By item (ii) of Lemma 4.9 with $\alpha = 1$, we have for $\kappa > 0$, $0 \leqslant n \leqslant N_p$, and $z_n \in (0, \kappa]$

$$\begin{aligned}\tilde{\phi}_n &= \frac{\kappa^2+(\kappa^2-z_n^2)\frac{\sin(2z_n)}{2z_n}}{\cos^2(z_n)} \leqslant \frac{2}{\cos^2(1)}\max\{\kappa^2,1\}\mathbb{1}_{\{z_n\leqslant 1\}} + 32\kappa^2(1+2\pi^{-1}z_n)^2\mathbb{1}_{\{1<z_n\leqslant\kappa\}}\\
&\leqslant \frac{2}{\cos^2(1)}\max\{\kappa^2,1\}\mathbb{1}_{\{z_n\leqslant 1\}} + 32\kappa^2(1+2\pi^{-1}\kappa)^2\mathbb{1}_{\{1<z_n\leqslant\kappa\}}\\
&\leqslant \frac{2}{\cos^2(1)}\max\{\kappa^2,1\}\mathbb{1}_{\{z_n\leqslant 1\}} + 32(1+2\pi^{-1})^2\max\{\kappa^4,1\}\mathbb{1}_{\{1<z_n\leqslant\kappa\}}\\
&\leqslant \max\left\{\frac{2}{\cos^2(1)}, 32(1+2\pi^{-1})^2\right\}\max\{\kappa^4,1\} \leqslant 86\max\{\kappa^4,1\},\end{aligned}$$

where we used (4.49) to arrive at the second term of the first inequality. Next, for $\kappa > 0$, $n \geqslant N_e$, and $z_n \in (0,\infty)$, we have $\kappa^2\|\tilde{Y}_n\|^2_{0,\mathcal{I}} = \frac{\frac{\sinh(2z_n)}{2z_n}+1}{\cosh(2z_n)+1}\kappa^2 \leqslant \kappa^2$. Consequently,

$$\max\left\{\max_{0\leqslant n\leqslant N_p}\tilde{\phi}_n, \tilde{\theta}_{N_c}, \kappa^2\max_{n\geqslant N_e}\|\tilde{Y}_n\|^2_{0,\mathcal{I}}\right\} \leqslant \max\left\{86\max\{\kappa^4,1\}, 2\kappa^2, \kappa^2\right\} = 86\max\{\kappa^4,1\}. \tag{4.51}$$

Applying the Parseval's identity and using (4.28), we have the first term of the right-hand side of (4.23). Next, we recall that since $\{\tilde{\mu}_n = n\pi\}_{n\in\mathbb{N}_0}$ or $\{\tilde{\mu}_n = (n+\frac{1}{2})\pi\}_{n\in\mathbb{N}_0}$, the lower bound in (4.32) still holds with μ_n, λ_n replaced by $\tilde{\mu}_n, \tilde{\lambda}_n$ respectively. We also note that for all $\kappa > 0$ and $z_n \in (0,\infty)$

$$\begin{aligned}\frac{(z_n^2+\kappa^2)^{\frac{1}{2}}\sinh(z_n)}{z_n\cosh(z_n)} &\leqslant \left(1+\kappa^2 z_n^{-2}\right)^{\frac{1}{2}}\tanh(z_n)\mathbb{1}_{\{\kappa\geqslant 1, z_n>1\}\cup\{\kappa<1,z_n>0\}}\\
&\quad + \left(z_n^2\kappa^{-2}+1\right)^{\frac{1}{2}}\left(\tanh(z_n)z_n^{-1}\right)\kappa\mathbb{1}_{\{\kappa\geqslant 1, z_n\leqslant 1\}}\\
&\leqslant \sqrt{2}\kappa\mathbb{1}_{\{\kappa\geqslant 1,z_n>1\}} + (1+\eta^{-2})^{\frac{1}{2}}\mathbb{1}_{\{\kappa<1,z_n>0\}} + \sqrt{2}\kappa\mathbb{1}_{\{\kappa\geqslant 1,z_n\leqslant 1\}}\\
&\leqslant \max\left\{\sqrt{2}, (1+\eta^{-2})^{\frac{1}{2}}\right\}\max\{\kappa,1\} \leqslant 2\max\{\kappa,1\},\end{aligned}$$

where we used the lower bound in (4.32) to arrive at the second term of the second inequality. Now, for $\kappa > 0$ and $n \geqslant N_e$, we have

$$\begin{aligned}\frac{\tilde{\mu}_n^2\|\tilde{Y}_n\|^2_{0,\mathcal{I}}+\|\tilde{Y}_n'\|^2_{0,\mathcal{I}}}{\tilde{\mu}_n} &= \frac{(2z_n^2+\kappa^2)\sinh(2z_n)+2\kappa^2 z_n}{2z_n(z_n^2+\kappa^2)^{\frac{1}{2}}(\cosh(2z_n)+1)}\\
&\leqslant \frac{(z_n^2+\kappa^2)^{\frac{1}{2}}\sinh(z_n)}{z_n\cosh(z_n)} + \frac{\kappa^2}{(z_n^2+\kappa^2)^{\frac{1}{2}}(\cosh(2z_n)+1)} \leqslant 3\max\{\kappa,1\}.\end{aligned} \tag{4.52}$$

Recall that we assumed $g_1 \in \mathcal{Z}^{1/2}(\Gamma_1)$. Plugging (4.51)-(4.52) into (4.50) and using (4.28),

we have the second term of the right-hand side of (4.23).

Item (2): suppose $\mathcal{B}_1 = \mathbf{I}_d$ and $\mathcal{B}_3 = \mathbf{I}_d$. Now, by item (ii) of Lemma 4.9 with $\alpha = 0$, we have for $\kappa > 0$, $0 \leqslant n \leqslant N_p$, and $z_n \in (0, \kappa]$

$$\begin{aligned}\tilde{\phi}_n &= \tfrac{\kappa^2 - (\kappa^2 - z_n^2)\frac{\sin(2z_n)}{2z_n}}{\sin^2(z_n)} \leqslant \tfrac{2}{\sin^2(z_n)}\kappa^2 \leqslant 32\kappa^2(1 + 2\pi^{-1}z_n)^2 \leqslant 32\kappa^2(1 + 2\pi^{-1}\kappa)^2 \\ &\leqslant 32(1 + 2\pi^{-1})^2 \max\{\kappa^4, 1\} \leqslant 86 \max\{\kappa^4, 1\},\end{aligned}$$

where we used (4.48) to arrive at the second inequality. Next for $n \geqslant N_e$ and $z_n > 0$, we have $\kappa^2 \|\tilde{Y}_n\|_{0,\mathcal{I}}^2 \leqslant \frac{(\sinh(2z_n) - 2z_n)}{2z_n(\cosh(2z_n) - 1)}\kappa^2 \leqslant \kappa^2$. Consequently,

$$\max\Big\{ \max_{0 \leqslant n \leqslant N_p} \tilde{\phi}_n, \tilde{\theta}_{N_c}, \max_{n \geqslant N_e} \tilde{\psi}_n \Big\} \leqslant \max\big\{86 \max\{\kappa^4, 1\}, \tfrac{2}{3}\kappa^2 + 1, \kappa^2\big\} = 86 \max\{\kappa^4, 1\}. \tag{4.53}$$

Applying the Parseval's identity and using (4.28), we have the first term of the right-hand side of (4.23). As before, since $\{\tilde{\mu}_n = n\pi\}_{n \in \mathbb{N}_0}$ or $\{\tilde{\mu}_n = (n + \frac{1}{2})\pi\}_{n \in \mathbb{N}_0}$, the lower bound in (4.32) still holds with μ_n, λ_n replaced by $\tilde{\mu}_n, \tilde{\lambda}_n$ respectively. Now, for $\kappa > 0$ and $n \geqslant N_e$, we have

$$\begin{aligned}\tfrac{\tilde{\mu}_n^2 \|\tilde{Y}_n\|_{0,\mathcal{I}}^2 + \|\tilde{Y}_n'\|_{0,\mathcal{I}}^2}{\tilde{\mu}_n} &= \tfrac{(2z_n^2 + \kappa^2)\sinh(2z_n) - 2\kappa^2 z_n}{2z_n(z_n^2 + \kappa^2)^{\frac{1}{2}}(\cosh(2z_n) - 1)} \leqslant \tfrac{\left(1 + \kappa^2 z_n^{-2}\right)^{1/2}\sinh(2z_n)}{(\cosh(2z_n) - 1)} \mathbb{1}_{\{\kappa \geqslant 1, z_n > 1\}} \\ &+ \tfrac{2\kappa\left(1 + z_n^2\kappa^{-2}\right)^{1/2}\sinh(2z_n)}{2z_n(\cosh(2z_n) - 1)} \mathbb{1}_{\{\kappa \geqslant 1, z_n \leqslant 1\}} + \tfrac{\left(1 + \kappa^2 z_n^{-2}\right)^{1/2}\sinh(2z_n)}{(\cosh(2z_n) - 1)} \mathbb{1}_{\{\kappa < 1\}} \\ &\leqslant \tfrac{\sqrt{2}\sinh(2)}{(\cosh(2) - 1)} \kappa \mathbb{1}_{\{\kappa \geqslant 1, z_n > 1\}} + \tfrac{2\sqrt{2}\cosh(\frac{\pi}{8})}{(\cosh(\frac{\pi}{8}) - 1)} \kappa \mathbb{1}_{\{\kappa \geqslant 1, z_n \leqslant 1\}} + \tfrac{\left(1 + \eta^{-2}\right)^{1/2}\sinh(2\eta)}{(\cosh(2\eta) - 1)} \mathbb{1}_{\{\kappa < 1\}} \\ &\leqslant \max\Big\{ \tfrac{\sqrt{2}\sinh(2)}{(\cosh(2) - 1)}, \tfrac{2\sqrt{2}\cosh(\frac{\pi}{8})}{(\cosh(\frac{\pi}{8}) - 1)}, \tfrac{\left(1 + \eta^{-2}\right)^{1/2}\sinh(2\eta)}{(\cosh(2\eta) - 1)} \Big\} \max\{\kappa, 1\} \leqslant 40 \max\{\kappa, 1\},\end{aligned} \tag{4.54}$$

where we used (4.47) and (4.32) to arrive at the second and third terms of the second inequality. Recall that we assumed $g_1 \in \mathcal{Z}^{1/2}(\Gamma_1)$. Plugging (4.53)-(4.54) into (4.50) and using (4.28), we have the second term of the right-hand side of (4.23). □

Chapter 5

Construction of Wavelets on a Bounded Interval

Before we present our wavelet Galerkin method, we need to address the critical issue on the construction of wavelets on a bounded interval, which is the main focus of this chapter.

We first provide some outlines and road maps of the classical and direct approaches for constructing compactly supported biorthogonal wavelets on $[0,\infty)$ from an arbitrarily given compactly supported biorthogonal wavelet on the real line in Section 5.1 before delving into the technical details and proofs. In Section 5.2, we study some basic properties of wavelets on the interval $[0,\infty)$ such as their Bessel properties and vanishing moments. In Section 5.3, we generalize the classical approach from scalar wavelets to multiwavelets for constructing compactly supported biorthogonal wavelets on the interval $[0,\infty)$. Additionally, we discuss the construction of orthogonal (multi)wavelets on $[0,\infty)$ in Algorithm 5.1. In Section 5.4, we present the direct approach for constructing all possible compactly supported biorthogonal wavelets on $[0,\infty)$ from any given compactly supported biorthogonal (multi)wavelets on the real line. Additionally, we discuss how to further improve the classical approach by the direct approach. In Section 5.5, we address stationary and nonstationary (multi)wavelets on $[0,\infty)$ satisfying any prescribed general homogeneous boundary conditions including Robin boundary conditions. In Section 5.6, we discuss how to construct wavelets on the interval $[0,N]$ with $N\in\mathbb{N}$ from wavelets on $[0,\infty)$. Using the classical approach and the direct approach, we present in Section 5.7 a few examples of orthogonal and biorthogonal wavelets on the interval $[0,1]$ such that the boundary wavelets have high vanishing moments and prescribed homogeneous boundary conditions. For improved readability, some technical proofs are postponed to Section 5.8.

Results in this chapter are based on [75].

5.1 Road maps

According to Theorem 5.1 in Section 5.2, a compactly supported biorthogonal wavelet $(\{\tilde{\phi};\tilde{\psi}\},\{\phi;\psi\})$ in $L_2(\mathbb{R})$ must satisfy

$$\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot -k), \quad \psi = 2\sum_{k\in\mathbb{Z}} b(k)\phi(2\cdot -k),$$
$$\tilde{\phi} = 2\sum_{k\in\mathbb{Z}} \tilde{a}(k)\tilde{\phi}(2\cdot -k), \quad \tilde{\psi} = 2\sum_{k\in\mathbb{Z}} \tilde{b}(k)\tilde{\phi}(2\cdot -k),$$

where $a,b,\tilde{a},\tilde{b}\in(l_0(\mathbb{Z}))^{r\times r}$ and by $(l_0(\mathbb{Z}))^{r\times r}$ we denote the space of all finitely supported sequences $u=\{u(k)\}_{k\in\mathbb{Z}}:\mathbb{Z}\to\mathbb{C}^{r\times r}$. The above multiscale relations (called the refinable structures in this paper) are well known to play the key role for a fast multiwavelet transform. By $\mathbb{P}_{m-1}$ we denote the space of all polynomials of degree less than m. Define $m:=\mathrm{vm}(\tilde{\psi})$ and $\tilde{m}:=\mathrm{vm}(\psi)$ for vanishing moments. Also, define a modified system $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ on $[0,\infty)$ adapted from $\mathsf{AS}_J(\phi;\psi)$ on $\mathbb{R}$ as follows

$$\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)} := \{2^{J/2}\varphi(2^J\cdot)\ :\ \varphi\in\Phi\}\cup\{2^{j/2}\eta(2^j\cdot)\ :\ j\geqslant J,\eta\in\Psi\}, \qquad J\in\mathbb{Z} \quad (5.1)$$

with

$$\Phi := \{\phi^L\}\cup\{\phi(\cdot-k)\ :\ k\geqslant n_\phi\}, \qquad \Psi := \{\psi^L\}\cup\{\psi(\cdot-k)\ :\ k\geqslant n_\psi\}, \quad (5.2)$$

where the boundary elements ϕ^L and ψ^L are vectors/sets of compactly supported functions in $L_2([0,\infty))$ and the integers n_ϕ,n_ψ are chosen so that all elements in $\{\phi(\cdot-k)\ :\ k\geqslant n_\phi\}\cup\{\psi(\cdot-k)\ :\ k\geqslant n_\psi\}$ are supported inside $[0,\infty)$ and hence are interior elements. From a given compactly supported biorthogonal wavelet $(\{\tilde{\phi};\tilde{\psi}\},\{\phi;\psi\})$ in $L_2(\mathbb{R})$, we are interested in deriving a compactly supported Riesz basis $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ in $L_2([0,\infty))$ satisfying

$$\phi^L = 2A_L\phi^L(2\cdot)+2\sum_{k=n_\phi}^{\infty} A(k)\phi(2\cdot-k), \qquad \psi^L = 2B_L\phi^L(2\cdot)+2\sum_{k=n_\phi}^{\infty} B(k)\phi(2\cdot-k), \quad (5.3)$$

for some matrices A_L,B_L and finitely supported sequences A,B, where $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is defined in (5.1) for Φ,Ψ in (5.2) with compactly supported boundary vector functions ϕ^L,ψ^L.

We shall prove in Theorem 5.2 that the unique dual Riesz basis of $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ must be given by $\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}\subseteq L_2([0,\infty))$, defined similarly as in (5.2) with $\tilde{\Phi}=\{\tilde{\phi}^L\}\cup\{\tilde{\phi}(\cdot-k)\ :\ k\geqslant n_{\tilde{\phi}}\}$ and $\tilde{\Psi}=\{\tilde{\psi}^L\}\cup\{\tilde{\psi}(\cdot-k)\ :\ k\geqslant n_{\tilde{\psi}}\}$, such that $\tilde{\phi}^L$ and $\tilde{\psi}^L$ must have *compact*

support and satisfy

$$\tilde{\phi}^L = 2\tilde{A}_L\tilde{\phi}^L(2\cdot) + 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{A}(k)\tilde{\phi}(2\cdot - k), \qquad \tilde{\psi}^L = 2\tilde{B}_L\tilde{\phi}^L(2\cdot) + 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{B}(k)\tilde{\phi}(2\cdot - k), \quad (5.4)$$

for some matrices $\tilde{A}_L, \tilde{B}_L$ and finitely supported sequences $\tilde{A}, \tilde{B}$. I.e., $(\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}, \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ forms a compactly supported biorthogonal wavelet in $L_2([0,\infty))$ such that all boundary elements $\phi^L, \psi^L, \tilde{\phi}^L, \tilde{\psi}^L$ have compact support and satisfy the refinable structures in (5.3) and (5.4). As stated in Theorem 5.7, the classical approach described in Section 5.3 for constructing a compactly supported biorthogonal wavelet $(\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}, \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ in $L_2([0,\infty))$ has four major steps:

(S1) Apply Proposition 5.4 and Section 5.3.1 to construct a compactly supported vector function ϕ^L in $L_2([0,\infty))$ satisfying the first identity in (5.3). To have polynomial reproduction, every $\mathsf{p}\chi_{[0,\infty)}$ with $\mathsf{p} \in \mathbb{P}_{m-1}$ should be an infinite linear combination of elements in Φ.

(S2) Use Algorithm 5.2 to construct a compactly supported vector function $\tilde{\phi}^L$ in $L_2([0,\infty))$ such that the first identity in (5.4) holds and $\tilde{\Phi}$ is biorthogonal to Φ. To have vanishing moments $\mathrm{vm}(\psi^L) = \tilde{m}$ with $\tilde{m} := \mathrm{vm}(\psi)$, every $\mathsf{p}\chi_{[0,\infty)}$ with $\mathsf{p} \in \mathbb{P}_{\tilde{m}-1}$ must necessarily be an infinite linear combination of elements in $\tilde{\Phi}$, see Lemma 5.3 for details.

(S3) Employ Proposition 5.11 to construct a compactly supported boundary primal wavelet ψ^L such that the second identity in (5.3) holds and Ψ is perpendicular to $\tilde{\Phi}$.

(S4) Employ Proposition 5.12 to construct a compactly supported boundary dual wavelet $\tilde{\psi}^L$ such that the second identity in (5.4) holds, $\tilde{\Psi}$ is perpendicular to Φ, and $\tilde{\Psi}$ is biorthogonal to Ψ.

The classical approach for constructing special vector functions ϕ^L in (S1) is quite simple, because each entry of ϕ^L is either some $\phi(\cdot - k)\chi_{[0,\infty)}, k \in \mathbb{Z}$ or their linear combinations. Once (S1) and (S2) are done, based on two simple observations in Theorem 5.5 and Lemma 5.8, ψ^L in (S3) can be easily constructed by Proposition 5.11. Though ψ^L itself is not unique, the remark after Proposition 5.11 shows that the finite-dimensional space generated by ψ^L modulated by the space spanned by $\{\psi(\cdot - k) \ : \ k \geqslant n_\psi\}$ is uniquely determined by Φ and $\tilde{\Phi}$. Once (S1)–(S3) are given, $\tilde{\psi}^L$ in (S4) can be easily constructed through Proposition 5.12 and both $\tilde{\psi}^L$ and $\tilde{\Psi}$ are uniquely determined by $\Phi, \tilde{\Phi}$ and Ψ. The Bessel property for the stability of $\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ and $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is guaranteed by Theorem 5.6. To apply wavelet-based methods for numerically solving boundary value problems, all the elements in the Riesz basis

$\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ must satisfy prescribed homogeneous boundary conditions. This can be easily done by applying Proposition 5.18 to the constructed ϕ^L and Φ in (S1).

For a given orthogonal (multi)wavelet $\{\phi;\psi\}$ in $L_2(\mathbb{R})$, because $\{\phi;\psi\}$ is biorthogonal to itself, (S2) can be avoided. Hence, adapting an orthogonal (multi)wavelet from the real line to $[0,\infty)$ becomes quite simple, because (S1) for constructing ϕ^L and (S3) for constructing ψ^L are fairly easy, see Algorithm 5.1 for more details. However, by Corollary 5.9 and Theorem 5.17, their boundary wavelets ψ^L cannot possess high vanishing moments and satisfy prescribed homogeneous boundary conditions simultaneously. By Theorem 5.17, this also holds for nonstationary orthonormal wavelets on $[0,\infty)$.

The main complexity/difficulty for the classical approach is (S2) in Algorithm 5.2 for constructing vector functions $\tilde{\phi}^L$, whose entries are finite linear combinations of $\tilde{\phi}(2^j\cdot -k)\chi_{[0,\infty)}, k\in\mathbb{Z}$ with $j\in\{0,1\}$. The complexity of (S2) is largely due to two facts: (1) The support of $\tilde{\phi}$ is often much longer than that of ϕ. Therefore, there are many more elements $\tilde{\phi}(\cdot-k)$ essentially touch the endpoint 0. (2) Because $\tilde{\Phi}$ is biorthogonal to Φ, we must have $\#\tilde{\phi}^L=\#\phi^L+(n_{\tilde{\phi}}-n_\phi)(\#\phi)$ and consequently, we do not have any freedom about the length of $\tilde{\phi}^L$. Therefore, it is no longer that easy or simple to construct even particular $\tilde{\phi}^L$ such that the first identity in (5.4) holds and $\tilde{\Phi}$ is biorthogonal to Φ.

The difficulty in (S2) for the classical approach motivates us to propose a direct approach, which is more general but simpler than the classical approach. The direct approach constructs ϕ^L and ψ^L in Theorems 5.13 and 5.14 without explicitly involving $\tilde{\phi}^L$ and $\tilde{\psi}^L$. Though the particularly constructed ϕ^L in (S1) by the classical approach can be reused, the direct approach in Theorem 5.13 constructs all possible general vector functions ϕ^L in (S1) by directly employing the first identity in (5.3) under the condition $\rho(A_L)<2^{-1/2}$, i.e., the spectral radius of A_L is less than $2^{-1/2}$. Without explicitly constructing $\tilde{\Phi}$ and $\tilde{\Psi}$, inspired by Lemma 5.8, the direct approach constructs ψ^L in Theorem 5.14 through the second identity in (5.3) under some necessary and sufficient conditions stated in Theorem 5.14. Now we can easily derive from (5.3) matrices $\tilde{A}_L,\tilde{B}_L$ and finitely supported sequences $\tilde{A},\tilde{B}$ from (5.3). Then we only need to check the condition $\rho(\tilde{A}_L)<2^{-1/2}$ to obtain a compactly supported biorthogonal wavelet $(\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)},\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ in $L_2([0,\infty))$, where $\tilde{\phi}^L$ and $\tilde{\psi}^L$ are defined in (5.4). The proof of Theorem 5.14 builds on Theorem 5.5, Theorem 5.6 for stability, and convergence property of non-standard vector cascade algorithms (which are closely linked to nonstandard vector subdivision schemes). In addition, the direct approach can also improve the classical approach by constructing all possible general vector functions $\tilde{\phi}^L$ in (S2) through Theorem 5.15 by only requiring that $\rho(\tilde{A}_L)<2^{-1/2}$ and $A_L,\tilde{A}_L,A,\tilde{A}$ in (5.3) and (5.4) should satisfy the identity in (5.64).

The procedure stated in Theorem 5.19 is well known (but without a proof) in the litera-

ture (e.g., see [31]) for adapting a compactly supported biorthogonal wavelet $(\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}, \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ in $L_2([0,\infty))$ to a bounded interval $[0,N]$ with $N \in \mathbb{N}$. We shall provide a rigorous proof for Theorem 5.19 in this paper. The main idea of Theorem 5.19 is quite simple: one constructs a closely related biorthogonal wavelet on the interval $(-\infty, N]$ whose interior elements are still given by $\psi_{j,k} := 2^{j/2}\psi(2^j\cdot - k)$ for some $j \in \mathbb{N}_0$ and $k \in \mathbb{Z}$. To obtain a locally supported biorthogonal wavelet on $[0,N]$, these two biorthogonal wavelets on $[0,\infty)$ and $(-\infty,N]$ are fused together in a straightforward way such that their boundary elements and all the common interior elements are kept. The main steps in Theorem 5.19 for obtaining a biorthogonal wavelet on $(-\infty,N]$ are as follows. (1) Flip functions about the origin, that is, we define

$$\mathring{\phi} := \phi(-\cdot), \quad \mathring{\psi} := \psi(-\cdot), \quad \mathring{\tilde{\phi}} := \tilde{\phi}(-\cdot) \quad \text{and} \quad \mathring{\tilde{\psi}} := \tilde{\psi}(-\cdot).$$

(2) Construct a biorthogonal wavelet $(\mathsf{AS}_0(\mathring{\tilde{\Phi}};\mathring{\tilde{\Psi}})_{[0,\infty)}, \mathsf{AS}_0(\mathring{\Phi};\mathring{\Psi})_{[0,\infty)})$ in $L_2([0,\infty))$ from the flipped biorthogonal wavelet $(\{\mathring{\tilde{\phi}};\mathring{\tilde{\psi}}\},\{\mathring{\phi};\mathring{\psi}\})$ in $L_2(\mathbb{R})$. (3) Then $\{\tilde{h}(N-\cdot)\}_{\tilde{h}\in\mathsf{AS}_0(\mathring{\tilde{\Phi}};\mathring{\tilde{\Psi}})_{[0,\infty)}}$ and $\{h(N-\cdot)\}_{h\in\mathsf{AS}_0(\mathring{\Phi};\mathring{\Psi})_{[0,\infty)}}$ form a biorthogonal wavelet in $L_2((-\infty,N])$. If all the vector functions in $\phi,\psi,\tilde{\phi},\tilde{\psi}$ possess symmetry, then a biorthogonal wavelet on $(-\infty,N]$ can be directly obtained from the constructed biorthogonal wavelet $(\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}, \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ in $L_2([0,\infty))$, see the remark after Theorem 5.19 for details.

5.2 Properties of biorthogonal wavelets on the interval $[0,\infty)$

In this section, we shall first recall some results on biorthogonal (multi)wavelets on the real line. Then we shall study some properties of biorthogonal wavelets on the interval $[0,\infty)$ which are derived from a compactly supported biorthogonal wavelet on the real line $\mathbb{R}$. Throughout the paper, for simplicity, wavelets stand for both scalar wavelets and multi-wavelets.

5.2.1 Biorthogonal wavelets on the real line

To recall some results on biorthogonal wavelets on the real line, let us first recall some definitions. The Fourier transform used in this paper is defined to be $\widehat{f}(\xi) := \int_{\mathbb{R}} f(x)e^{-\mathrm{i}x\xi}dx, \xi \in \mathbb{R}$ for $f \in L_1(\mathbb{R})$ and is naturally extended to square integrable functions in $L_2(\mathbb{R})$. By $(l_0(\mathbb{Z}))^{r\times s}$ we denote the set of all finitely supported sequences $u = \{u(k)\}_{k\in\mathbb{Z}} : \mathbb{Z} \to \mathbb{C}^{r\times s}$.

For $u = \{u(k)\}_{k\in\mathbb{Z}} \in (l_0(\mathbb{Z}))^{r\times s}$, its Fourier series is defined to be

$$\widehat{u}(\xi) := \sum_{k\in\mathbb{Z}} u(k)e^{-\mathrm{i}k\xi} \qquad \text{for} \quad \xi \in \mathbb{R},$$

which is an $r \times s$ matrix of 2π-periodic trigonometric polynomials. An element in $(l_0(\mathbb{Z}))^{r\times s}$ is often called a (matrix-valued) mask or filter in the literature. By $\boldsymbol{\delta}$ we denote the Dirac sequence such that $\boldsymbol{\delta}(0) = 1$ and $\boldsymbol{\delta}(k) = 0$ for all $k \in \mathbb{Z}\backslash\{0\}$. Note that $\widehat{\boldsymbol{\delta}} = 1$. By $f \in (L_2(\mathbb{R}))^{r\times s}$ we mean that f is an $r \times s$ matrix of functions in $L_2(\mathbb{R})$ and we define

$$\langle f, g\rangle := \int_{\mathbb{R}} f(x)\overline{g(x)}^{\mathsf{T}} dx, \qquad f \in (L_2(\mathbb{R}))^{r\times t}, g \in (L_2(\mathbb{R}))^{s\times t}.$$

According to [71, Theorem 4.5.1] and [70, Theorem 7], any biorthogonal wavelet $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ in $L_2(\mathbb{R})$ must be intrinsically derived from refinable vector functions and biorthogonal wavelet filter banks. For simplicity, we only state the following result for compactly supported biorthogonal wavelets $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ in $L_2(\mathbb{R})$.

Theorem 5.1. *([71, Theorem 4.5.1] and [70, Theorem 7]) Let $\phi, \tilde{\phi}$ be $r \times 1$ vectors of compactly supported distributions and $\psi, \tilde{\psi}$ be $s \times 1$ vectors of compactly supported distributions on $\mathbb{R}$. Then $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ is a biorthogonal wavelet in $L_2(\mathbb{R})$ if and only if the following are satisfied*

(1) $\phi, \tilde{\phi} \in (L_2(\mathbb{R}))^r$ and $\overline{\widehat{\phi}(0)}^{\mathsf{T}}\widehat{\tilde{\phi}}(0) = 1$.

(2) ϕ and $\tilde{\phi}$ are biorthogonal to each other: $\langle \phi, \tilde{\phi}(\cdot - k)\rangle = \boldsymbol{\delta}(k)I_r$ for all $k \in \mathbb{Z}$.

(3) There exist low-pass filters $a, \tilde{a} \in (l_0(\mathbb{Z}))^{r\times r}$ and high-pass filters $b, \tilde{b} \in (l_0(\mathbb{Z}))^{s\times r}$ such that

$$\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot -k), \qquad \psi = 2\sum_{k\in\mathbb{Z}} b(k)\phi(2\cdot -k), \tag{5.5}$$

$$\tilde{\phi} = 2\sum_{k\in\mathbb{Z}} \tilde{a}(k)\tilde{\phi}(2\cdot -k), \qquad \tilde{\psi} = 2\sum_{k\in\mathbb{Z}} \tilde{b}(k)\tilde{\phi}(2\cdot -k), \tag{5.6}$$

and $(\{\tilde{a}; \tilde{b}\}, \{a; b\})$ is a biorthogonal wavelet filter bank, i.e., $s = r$ and

$$\begin{bmatrix} \widehat{\tilde{a}}(\xi) & \widehat{\tilde{a}}(\xi+\pi) \\ \widehat{\tilde{b}}(\xi) & \widehat{\tilde{b}}(\xi+\pi) \end{bmatrix} \begin{bmatrix} \overline{\widehat{a}(\xi)}^{\mathsf{T}} & \overline{\widehat{b}(\xi)}^{\mathsf{T}} \\ \overline{\widehat{a}(\xi+\pi)}^{\mathsf{T}} & \overline{\widehat{b}(\xi+\pi)}^{\mathsf{T}} \end{bmatrix} = I_{2r}, \qquad \xi \in \mathbb{R}. \tag{5.7}$$

(4) Both $\mathsf{AS}_0(\phi; \psi)$ and $\mathsf{AS}_0(\tilde{\phi}; \tilde{\psi})$ are Bessel sequences in $L_2(\mathbb{R})$, that is, there exists a

positive constant C such that

$$\sum_{h\in \mathsf{AS}_0(\phi;\psi)} |\langle f,h\rangle|^2 \leqslant C\|f\|^2_{L_2(\mathbb{R})} \quad \textit{and} \quad \sum_{\tilde{h}\in \mathsf{AS}_0(\tilde{\phi};\tilde{\psi})} |\langle f,\tilde{h}\rangle|^2 \leqslant C\|f\|^2_{L_2(\mathbb{R})}, \qquad \forall\, f\in L_2(\mathbb{R}).$$

A vector function ϕ satisfying (5.5) is called *a refinable vector function* with a refinement filter/mask $a \in (l_0(\mathbb{Z}))^{r\times r}$. For a vector function ϕ we also regard ϕ as an ordered set and vice versa. We define $\#\phi$ to be the number of entries in ϕ, that is, the cardinality of the set/vector ϕ. For $r = 1$, a refinable vector function is often called a (scalar) refinable function. By [71, Theorems 4.6.5 and 6.4.6] or [67, Theorem 2.3], item (4) of Theorem 5.1 can be replaced by

(4') both ψ and $\tilde{\psi}$ have at least one vanishing moment, i.e., $\int_{\mathbb{R}} \psi(x)dx = \int_{\mathbb{R}} \tilde{\psi}(x)dx = 0$.

Orthogonal and biorthogonal wavelets on the real line which are derived from refinable functions have been extensively studied, for example, see [24, 29, 38] for scalar wavelets and [30, 58, 59, 65, 71, 72, 78, 79, 89, 91, 93, 95] and references therein for multiwavelets. It is well known in these papers that the study and construction of multiwavelets and refinable vector functions are often much more involved and complicated than their scalar counterparts, largely because the refinable vector function ϕ and multiwavelet ψ in (5.5) are vector functions with matrix-valued filters a and b.

5.2.2 The dual of a Riesz basis $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ on $[0,\infty)$

To improve readability and to reduce confusion later about some notations, in the following let us first state our conventions on some notations. If not explicitly stated, $(\{\tilde{\phi};\tilde{\psi}\},\{\phi;\psi\})$ in this paper is always a compactly supported biorthogonal (multi)wavelet in $L_2(\mathbb{R})$ satisfying all items (1)–(4) of Theorem 5.1, which necessarily implies $\#\phi = \#\psi$. For a compactly supported (vector) function ϕ (or a finitely supported filter $a \in (l_0(\mathbb{Z}))^{r\times s}$), we define $\mathrm{fsupp}(\phi)$ (or $\mathrm{fsupp}(a)$) to be the shortest interval with *integer endpoints* such that ϕ (or a) vanishes outside $\mathrm{fsupp}(\phi)$ (or $\mathrm{fsupp}(a)$). Throughout the paper we always define

$$[l_\phi,h_\phi] := \mathrm{fsupp}(\phi), \quad [l_\psi,h_\psi] := \mathrm{fsupp}(\psi), \quad [l_a,h_a] := \mathrm{fsupp}(a), \quad [l_b,h_b] := \mathrm{fsupp}(b), \tag{5.8}$$

$$[l_{\tilde{\phi}},h_{\tilde{\phi}}] := \mathrm{fsupp}(\tilde{\phi}), \quad [l_{\tilde{\psi}},h_{\tilde{\psi}}] := \mathrm{fsupp}(\tilde{\psi}), \quad [l_{\tilde{a}},h_{\tilde{a}}] := \mathrm{fsupp}(\tilde{a}), \quad [l_{\tilde{b}},h_{\tilde{b}}] := \mathrm{fsupp}(\tilde{b}). \tag{5.9}$$

One can easily deduce from (5.5) (called the refinable structure in this paper) that

$$[l_\phi, h_\phi] \subseteq [l_a, h_a] \quad \text{and} \quad [l_\psi, h_\psi] \subseteq [\lfloor \tfrac{l_b+l_\phi}{2} \rfloor, \lceil \tfrac{h_b+h_\phi}{2} \rceil]. \tag{5.10}$$

For the scalar case $r = 1$, both $\subseteq$ in (5.10) become identities. But strict $\subseteq$ in (5.10) can happen for the multiwavelet case $r > 1$. Note that $\text{fsupp}(\phi(\cdot - k)) = [k + l_\phi, k + h_\phi]$. Hence, $\text{supp}(\phi(\cdot - k)) \subseteq (-\infty, 0]$ for all integers $k \leqslant -h_\phi$ and $\text{supp}(\phi(\cdot - k)) \subseteq [0, \infty)$ for all integers $k \geqslant -l_\phi$. In other words, the point 0 is an interior point of $\text{fsupp}(\phi(\cdot - k))$ if and only if $1 - h_\phi \leqslant k \leqslant -1 - l_\phi$. On the other hand, we deduce from the refinable structure in (5.5) that

$$\phi(\cdot - k_0) = 2 \sum_{k=l_a+2k_0}^{h_a+2k_0} a(k - 2k_0)\phi(2 \cdot -k), \qquad k_0 \in \mathbb{Z}, \tag{5.11}$$

$$\psi(\cdot - k_0) = 2 \sum_{k=l_b+2k_0}^{h_b+2k_0} b(k - 2k_0)\phi(2 \cdot -k), \qquad k_0 \in \mathbb{Z}. \tag{5.12}$$

For any integer n_ϕ satisfying $n_\phi \geqslant \max(-l_\phi, -l_a)$, we have $l_a + 2k_0 \geqslant l_a + 2n_\phi \geqslant n_\phi$ for all $k_0 \geqslant n_\phi$ and consequently, we deduce from (5.11) that

$$\text{fsupp}(\phi(\cdot - k_0)) \subseteq [0, \infty) \quad \text{and} \quad \phi(\cdot - k_0) = 2 \sum_{k=n_\phi}^{\infty} a(k - 2k_0)\phi(2 \cdot -k), \qquad \forall\, k_0 \geqslant n_\phi. \tag{5.13}$$

Similarly, for any integer n_ψ satisfying $n_\psi \geqslant \max(-l_\psi, \frac{n_\phi - l_b}{2})$, we have $l_b + 2k_0 \geqslant l_b + 2n_\psi \geqslant n_\phi$ for all $k_0 \geqslant n_\psi$ and hence we deduce from (5.12) that

$$\text{fsupp}(\psi(\cdot - k_0)) \subseteq [0, \infty) \quad \text{and} \quad \psi(\cdot - k_0) = 2 \sum_{k=n_\phi}^{\infty} b(k - 2k_0)\phi(2 \cdot -k), \qquad \forall\, k_0 \geqslant n_\psi. \tag{5.14}$$

Throughout the paper, the integers n_ϕ and n_ψ are always chosen (not necessary to be the smallest) such that (5.13) and (5.14) hold. We make the same convention for $n_{\tilde{\phi}}$ and $n_{\tilde{\psi}}$ similarly. Let $\tilde{\phi}^L$ and $\tilde{\psi}^L$ be vector functions in $L_2([0, \infty))$. Similarly to Φ and Ψ in (5.2), we define

$$\tilde{\Phi} := \{\tilde{\phi}^L\} \cup \{\tilde{\phi}(\cdot - k) \ : \ k \geqslant n_{\tilde{\phi}}\}, \qquad \tilde{\Psi} := \{\tilde{\psi}^L\} \cup \{\tilde{\psi}(\cdot - k) \ : \ k \geqslant n_{\tilde{\psi}}\}. \tag{5.15}$$

Under the following conditions for matching cardinality between $\Phi \cup \Psi$ and $\tilde{\Phi} \cup \tilde{\Psi}$:

$$\#\tilde{\phi}^L - \#\phi^L = (n_{\tilde{\phi}} - n_\phi)(\#\phi) \quad \text{and} \quad \#\tilde{\psi}^L - \#\psi^L = (n_{\tilde{\psi}} - n_\psi)(\#\psi), \tag{5.16}$$

throughout the paper, the mapping $\sim\colon \Phi \to \tilde{\Phi}$ with $h \mapsto \tilde{h}$ is always the default bijection between Φ and $\tilde{\Phi}$ such that $\phi(\cdot - k)$ corresponds to $\tilde{\phi}(\cdot - k)$ for all $k \geqslant \max(n_\phi, n_{\tilde{\phi}})$, and the bijection $\sim$ for other elements is determined by their corresponding positions in the ordered sets/vectors Φ and $\tilde{\Phi}$. The bijection $\sim\colon \Psi \to \tilde{\Psi}$ is defined similarly by mapping $\psi(\cdot - k)$ to $\tilde{\psi}(\cdot - k)$ for all $k \geqslant \max(n_\psi, n_{\tilde{\psi}})$.

As explained in Section 5.1, the pair $(\mathsf{AS}_J(\tilde{\phi}; \tilde{\psi}), \mathsf{AS}_J(\phi; \psi))$ on the real line will be modified into a pair $(\mathsf{AS}_J(\tilde{\Phi}; \tilde{\Psi})_{[0,\infty)}, \mathsf{AS}_J(\Phi; \Psi)_{[0,\infty)})$ of biorthogonal systems in $L_2([0,\infty))$ by keeping their elements supported inside $[0,\infty)$ as interior elements and by modifying their elements essentially touching the endpoint 0 into boundary elements. By the definition in (5.1) and a simple scaling argument as in [70], it is straightforward to see that $\mathsf{AS}_J(\Phi; \Psi)_{[0,\infty)}$ is a Riesz (or orthonormal) basis of $L_2([0,\infty))$ for all $J \in \mathbb{Z}$ if and only if $\mathsf{AS}_0(\Phi; \Psi)_{[0,\infty)}$ is a Riesz (or orthonormal) basis of $L_2([0,\infty))$.

We now study the structure of compactly supported Riesz wavelets on $[0,\infty)$ in the following result, whose proof is presented in Section 5.8 and which plays a key role in our study of locally supported biorthogonal wavelets on intervals.

Theorem 5.2. *Let $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a biorthogonal wavelet filter bank $(\{\tilde{a}\ \tilde{b}\}, \{a; b\})$ satisfying items (1)–(4) of Theorem 5.1. Let ϕ^L and ψ^L be vectors of compactly supported functions in $L_2([0,\infty))$. Define l_ϕ, h_ϕ, l_a, h_a and $l_{\tilde{\phi}}, h_{\tilde{\phi}}, l_{\tilde{a}}, h_{\tilde{a}}$ as in (5.8) and (5.9). Define Φ, Ψ as in (5.2) with integers $n_\phi \geqslant \max(-l_\phi, -l_a)$ and $n_\psi \geqslant \max(-l_\psi, \frac{n_\phi - l_b}{2})$. If $\mathsf{AS}_0(\Phi; \Psi)_{[0,\infty)}$ in (5.1) is a Riesz basis of $L_2([0,\infty))$ and satisfies*

$$\phi^L = 2A_L \phi^L(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A(k)\phi(2\cdot - k), \tag{5.17}$$

$$\psi^L = 2B_L \phi^L(2\cdot) + 2\sum_{k=n_\phi}^{\infty} B(k)\phi(2\cdot - k), \tag{5.18}$$

for some matrices A_L, B_L and finitely supported sequences A, B of matrices, then

(1) there must exist compactly supported vector functions $\tilde{\phi}^L, \tilde{\psi}^L$ in $L_2([0,\infty))$ and integers $n_{\tilde{\phi}} \geqslant \max(-l_{\tilde{\phi}}, -l_{\tilde{a}}, n_\phi)$ and $n_{\tilde{\psi}} \geqslant \max(-l_{\tilde{\psi}}, \frac{n_{\tilde{\phi}} - l_{\tilde{b}}}{2}, n_\psi)$ satisfying (5.16) such that $\mathsf{AS}_0(\tilde{\Phi}; \tilde{\Psi})_{[0,\infty)}$ is the dual Riesz basis of $\mathsf{AS}_0(\Phi; \Psi)_{[0,\infty)}$ in $L_2([0,\infty))$, where

$$\mathsf{AS}_0(\tilde{\Phi}; \tilde{\Psi})_{[0,\infty)} := \tilde{\Phi} \cup \{2^{j/2}\tilde{\eta}(2^j\cdot)\ :\ j \in \mathbb{N} \cup \{0\}, \tilde{\eta} \in \tilde{\Psi}\}$$

and $\tilde{\Phi}, \tilde{\Psi}$ are defined in (5.15);

(2) there exist matrices $\tilde{A}_L, \tilde{B}_L$ and finitely supported sequences $\tilde{A}, \tilde{B}$ of matrices such that

$$\tilde{\phi}^L = 2\tilde{A}_L\tilde{\phi}^L(2\cdot) + 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{A}(k)\tilde{\phi}(2\cdot - k), \tag{5.19}$$

$$\tilde{\psi}^L = 2\tilde{B}_L\tilde{\phi}^L(2\cdot) + 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{B}(k)\tilde{\phi}(2\cdot - k), \tag{5.20}$$

and

$$\operatorname{fsupp}(\tilde{\phi}(\cdot - k_0)) \subseteq [0,\infty), \quad \tilde{\phi}(\cdot - k_0) = 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{a}(k-2k_0)\tilde{\phi}(2\cdot - k), \qquad \forall\, k_0 \geqslant n_{\tilde{\phi}}, \tag{5.21}$$

$$\operatorname{fsupp}(\tilde{\psi}(\cdot - k_0)) \subseteq [0,\infty), \quad \tilde{\psi}(\cdot - k_0) = 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{b}(k-2k_0)\tilde{\phi}(2\cdot - k), \qquad \forall\, k_0 \geqslant n_{\tilde{\psi}}; \tag{5.22}$$

(3) Every element in $\Phi(2\cdot) := \{\phi^L(2\cdot)\} \cup \{\phi(2\cdot - k) \,:\, k \geqslant n_\phi\}$ can be uniquely written as a finite linear combination of elements in $\Phi \cup \Psi$.

Basically, Theorem 5.2 says that a compactly supported Riesz basis $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ of $L_2([0,\infty))$ satisfying (5.17) and (5.18) must have a dual *compactly supported* Riesz basis $\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ of $L_2([0,\infty))$ satisfying (5.19) and (5.20). Theorem 5.2 serves as our foundation for developing the classical approach through items (1) and (2) of Theorem 5.2 and the direct approach through item (3) of Theorem 5.2 for deriving wavelets on intervals from biorthogonal multiwavelets in $L_2(\mathbb{R})$.

5.2.3 Vanishing moments of biorthogonal wavelets on $[0,\infty)$

Recall that ψ has m vanishing moments if $\int_{\mathbb{R}} x^j\psi(x)dx = 0$ for all $j = 0,\ldots,m-1$. In particular, we define $\operatorname{vm}(\psi) := m$ with m being the largest such nonnegative integer. By $\mathbb{P}_{m-1}$ we denote the space of all polynomials of degree less than m. Define $\mathbb{N}_0 := \mathbb{N}\cup\{0\}$. Let us now discuss the known relation between vanishing moments and polynomial reproduction for biorthogonal wavelets on the interval $[0,\infty)$.

Lemma 5.3. *Let $\phi,\psi,\tilde{\phi},\tilde{\psi}$ be vectors of compactly supported functions in $L_2(\mathbb{R})$. Let $\phi^L,\psi^L,\tilde{\phi}^L,\tilde{\psi}^L$ be vectors of compactly supported functions in $L_2([0,\infty))$. Suppose that $\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ and $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ form a pair of biorthogonal Riesz bases in $L_2([0,\infty))$, where*

Φ, Ψ *and* $\tilde\Phi, \tilde\Psi$ *are defined in* (5.2) *and* (5.15)*, respectively. Then* $\min(\mathrm{vm}(\psi^L), \mathrm{vm}(\psi)) \geqslant m$ *if and only if every polynomial* $\mathsf{p}\chi_{[0,\infty)}$ *on* $[0,\infty)$ *with* $\mathsf{p} \in \mathbb{P}_{m-1}$ *can be written as an infinite linear combination of elements in* $\tilde\Phi$*.*

Proof. Necessity. Suppose that $\min(\mathrm{vm}(\psi^L), \mathrm{vm}(\psi)) \geqslant m$. Since all functions in $\phi \cup \psi \cup \tilde\phi \cup \tilde\psi$ and $\phi^L \cup \psi^L \cup \tilde\phi^L \cup \tilde\psi^L$ have compact support, we assume that they are supported inside $[-N_0, N_0]$ for some $N_0 \in \mathbb{N}$. For every polynomial $\mathsf{p} \in \mathbb{P}_{m-1}$ and $N \in \mathbb{N}$, we have $\mathsf{p}\chi_{[0,N)} \in L_2(\mathbb{R})$ and hence

$$\mathsf{p}\chi_{[0,N)} = \sum_{h\in\Phi} \langle \mathsf{p}\chi_{[0,N)}, h\rangle \tilde h + \sum_{j=0}^{\infty} \sum_{h\in\Psi} \langle \mathsf{p}\chi_{[0,N)}, 2^{j/2} h(2^j\cdot)\rangle 2^{j/2} \tilde h(2^j\cdot).$$

Note that $\tilde\psi_{j;k}$ and $\psi_{j;k}$ are supported inside $[2^{-j}(k-N_0), 2^{-j}(k+N_0)]$. Since $\langle \mathsf{p}, h(2^j\cdot)\rangle = 0$ for all $h \in \Psi$ and $j \in \mathbb{N}_0$, we observe that $\langle \mathsf{p}\chi_{[0,N)}, h(2^j\cdot)\rangle \tilde h(2^j x) = 0$ a.e. $x \in [0, N-2N_0)$ for all $h \in \Psi$ and $j \in \mathbb{N}_0$. Therefore, we conclude from the above identity that

$$\mathsf{p}(x)\chi_{[0,\infty)}(x) = \mathsf{p}(x)\chi_{[0,N)}(x) = \sum_{h\in\Phi} \langle \mathsf{p}\chi_{[0,N)}, h\rangle \tilde h(x) = \sum_{h\in\Phi} \langle \mathsf{p}, h\rangle \tilde h(x), \qquad a.e.\, x \in [0, N-2N_0).$$

Taking $N \to \infty$ in the above identity, we conclude from the above identity that $\mathsf{p}\chi_{[0,\infty)}$ with $\mathsf{p} \in \mathbb{P}_{m-1}$ can be written as an infinite linear combination of elements in $\tilde\Phi$.

Sufficiency. Suppose that $\mathsf{p}\chi_{[0,\infty)}$ with $\mathsf{p} \in \mathbb{P}_{m-1}$ can be written as $\mathsf{p}\chi_{[0,\infty)} = \sum_{h\in\Phi} c_h \tilde h$. Since every element in Ψ is perpendicular to $\tilde\Phi$ and all elements in $\Psi \cup \tilde\Phi$ have compact support, we have $\langle \mathsf{p}\chi_{[0,\infty)}, g\rangle = \sum_{h\in\Phi} c_h \langle \tilde h, g\rangle = 0$ for all $g \in \Psi$. This proves $\min(\mathrm{vm}(\psi^L), \mathrm{vm}(\psi)) = \mathrm{vm}(\Psi) \geqslant m$. □

The above same argument in Lemma 5.3 can be also applied to biorthogonal wavelets on the real line $\mathbb{R}$ or intervals $[0, N]$ with $N \in \mathbb{N}$. For a biorthogonal wavelet $(\{\tilde\phi; \tilde\psi\}, \{\phi; \psi\})$ in $L_2(\mathbb{R})$, that every polynomial in $\mathbb{P}_{m-1}$ can be written as an infinite linear combination of $\phi(\cdot - k), k \in \mathbb{Z}$ if and only if $\mathrm{vm}(\tilde\psi) \geqslant m$. We say that a (matrix-valued) filter $a \in (l_0(\mathbb{Z}))^{r\times r}$ has *order* m *sum rules with a (moment) matching filter* $v \in (l_0(\mathbb{Z}))^{1\times r}$ if $\widehat v(0)\widehat\phi(0) = 1$ and

$$[\widehat v(2\cdot)\widehat a]^{(j)}(0) = \widehat v^{(j)}(0) \quad \text{and} \quad [\widehat v(2\cdot)\widehat a(\cdot+\pi)]^{(j)}(0) = 0, \qquad \forall\, j = 0, \ldots, m-1. \tag{5.23}$$

In particular, we define $\mathrm{sr}(a) = m$ with m being the largest such nonnegative integer. Let $(\{\tilde\phi; \tilde\psi\}, \{\phi; \psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a finitely supported biorthogonal wavelet filter bank $(\{\tilde a; \tilde b\}, \{a; b\})$ in Theorem 5.1. Then $\mathrm{vm}(\tilde\psi) \geqslant m$ if and only if $\mathrm{sr}(a) \geqslant m$. That is, $\mathrm{vm}(\tilde\psi) = \mathrm{sr}(a)$ and $\mathrm{vm}(\psi) = \mathrm{sr}(\tilde a)$. Moreover, from (5.23), we further have $[\widehat v\widehat\phi]^{(j)}(2\pi k) = \boldsymbol{\delta}(k)\boldsymbol{\delta}(j)$ for all $j = 0, \ldots, m-1$ and $k \in \mathbb{Z}$ (see [71, (5.6.6)]

and [65, 66]) and consequently, for all $\mathsf{p} \in \mathbb{P}_{m-1}$,

$$\mathsf{p} = \sum_{k\in\mathbb{Z}} [\mathsf{p} * \upsilon](k)\phi(\cdot - k) = \sum_{k\in\mathbb{Z}} \mathsf{p}_\upsilon(k)\phi(\cdot - k) \quad \text{with} \quad \mathsf{p}_\upsilon := \mathsf{p} * \upsilon = \sum_{j=0}^{\infty} \frac{(-\mathsf{i})^j}{j!} \mathsf{p}^{(j)}(\cdot)\widehat{\upsilon}^{(j)}(0), \tag{5.24}$$

since $\mathsf{p} * \upsilon := \sum_{n\in\mathbb{Z}} \mathsf{p}(\cdot - n)\upsilon(n) = \mathsf{p}_\upsilon \in \mathbb{P}_{m-1}$ (see [71, Lemma 1.2.1 and Theorem 5.5.1]). Moreover, one can easily deduce from (5.23) that the quantities $\widehat{\upsilon}^{(j)}(0), j = 0, 1, \ldots, m-1$ are determined (see [71, (5.6.10)]) through $\widehat{\upsilon}(0)\widehat{a}(0) = \widehat{\upsilon}(0)$ with $\widehat{\upsilon}(0)\widehat{\phi}(0) = 1$, and the following recursive formula

$$\widehat{\upsilon}^{(j)}(0) = \sum_{k=0}^{j-1} \frac{2^k j!}{k!(j-k)!} \widehat{\upsilon}^{(k)}(0)\widehat{a}^{(j-k)}(0)[I_r - 2^j\widehat{a}(0)]^{-1}, \qquad j = 1, \ldots, m-1 \tag{5.25}$$

provided that 2^{-j} is not an eigenvalue of $\widehat{a}(0)$ for all $j = 1, \ldots, m-1$. For $r = 1$ and a scalar filter $a \in l_0(\mathbb{Z})$ with $\widehat{a}(0) = 1$, a scalar filter/mask a has order m sum rules if and only if $(1 + e^{-\mathsf{i}\xi})^m \mid \widehat{a}(\xi)$, which is equivalent to $\widehat{a}^{(j)}(\pi) = 0$ for all $j = 0, \ldots, m-1$. For the scalar case $r = 1$ and $\widehat{a}(0) = 1$, because $\widehat{\phi}(\xi) := \prod_{j=1}^{\infty} \widehat{a}(2^{-j}\xi)$ is well defined, we must have $\widehat{\upsilon}^{(j)}(0) = [1/\widehat{\phi}]^{(j)}(0)$ for all $j \in \mathbb{N}_0$, which can be computed via (5.25) by starting with $\widehat{\upsilon}(0) = 1$. The sum rules in (5.23) for matrix-valued filters in (5.23) make it more involved to study refinable vector functions and matrix-valued filters than their scalar counterparts. Biorthogonal multiwavelets in $L_2(\mathbb{R})$ with high vanishing moments can be easily constructed by a coset by coset (CBC) algorithm in [65, Theorem 3.4] or [71, Algorithm 6.5.2]. Moreover, the values $\widehat{\tilde{\upsilon}}^{(j)}(0), j \in \mathbb{N}_0$ of the matching filter $\tilde{\upsilon}$ for the dual mask $\tilde{a}$ are uniquely determined by the primal mask a as given in [71, Theorem 6.5.1] or [65, Theorem 3.1] through the following identities: $\widehat{\tilde{\upsilon}}(0)\overline{\widehat{a}(0)}^{\mathsf{T}} = \widehat{\tilde{\upsilon}}(0)$ with $\widehat{\tilde{\upsilon}}(0)\widehat{\tilde{\phi}}(0) = 1$, and the following recursive formula:

$$\widehat{\tilde{\upsilon}}^{(j)}(0) = \sum_{k=0}^{j-1} \frac{j!}{k!(j-k)!} \widehat{\tilde{\upsilon}}^{(k)}(0)\overline{\widehat{a}^{(j-k)}(0)}^{\mathsf{T}}[2^j I_r - \overline{\widehat{a}(0)}^{\mathsf{T}}]^{-1}, \qquad j \in \mathbb{N}$$

provided that 2^j is not an eigenvalue of $\widehat{a}(0)$ for all $j \in \mathbb{N}$.

The following result constructs special ϕ^L satisfying (5.17) with polynomial reproduction property.

Proposition 5.4. *Let ϕ be an $r \times 1$ vector of compactly supported functions in $L_2(\mathbb{R})$ such that $\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot - k)$ for some finitely supported sequence $a \in (l_0(\mathbb{Z}))^{r\times r}$. Define $[l_\phi, h_\phi] := \mathrm{fsupp}(\phi)$ and $[l_a, h_a] := \mathrm{fsupp}(a)$. For any integer $n_\phi \in \mathbb{Z}$ satisfying $n_\phi \geqslant \max(-l_\phi, -l_a)$, then*

(i) the column vector function $\phi^c := (\phi(\cdot - k)\chi_{[0,\infty)})_{1-h_\phi \leqslant k \leqslant n_\phi - 1}$ *(note that* ϕ^c *contains interior elements* $\phi(\cdot - k), -l_\phi \leqslant k \leqslant n_\phi - 1$*) satisfies the following refinement equation:*

$$\phi^c = 2A_{L_c}\phi^c(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A_c(k)\phi(2\cdot - k), \tag{5.26}$$

where $A_{L_c} = [a(k_0 - 2n)]_{1-h_\phi \leqslant n, k_0 \leqslant n_\phi - 1}$ *and* A_c *is a finitely supported sequence given by* $A_c(k) := [a(k - 2n)]_{1-h_\phi \leqslant n \leqslant n_\phi - 1}$ *for* $k \geqslant n_\phi$*, where* n *is row index and* k_0 *is column index;*

(ii) if in addition the filter a *has order* m *sum rules in (5.23) with a matching filter* $\upsilon \in (l_0(\mathbb{Z}))^{1\times r}$ *satisfying* $\widehat{\upsilon}(0)\widehat{\phi}(0) = 1$*, then for any* $j_0, \ldots, j_\ell \in \{0, \ldots, m-1\}$*, the vector function* ϕ^{p}*, whose entries are linear combinations of elements in* ϕ^c*, is given by*

$$\phi^{\mathsf{p}} := \sum_{k=1-h_\phi}^{n_\phi - 1}\sum_{j=0}^{m-1} \frac{(-i)^j}{j!}\mathsf{p}^{(j)}(k)\widehat{\upsilon}^{(j)}(0)\phi(\cdot - k)\chi_{[0,\infty)} \quad \text{with} \quad \mathsf{p}(x) := (x^{j_0}, \ldots, x^{j_\ell})^{\mathsf{T}} \tag{5.27}$$

must satisfy (5.17) with ϕ^L *being replaced by* ϕ^{p}*, more precisely,*

$$\phi^{\mathsf{p}} = 2A_{L_{\mathsf{p}}}\phi^{\mathsf{p}}(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A_{\mathsf{p}}(k)\phi(2\cdot - k) \quad \text{with} \quad A_{L_{\mathsf{p}}} := \mathrm{diag}(2^{-1-j_0}, \ldots, 2^{-1-j_\ell}), \tag{5.28}$$

where $A_{\mathsf{p}} = \{A_{\mathsf{p}}(k)\}_{k=n_\phi}^{\infty}$ *is a finitely supported sequence given by* $A_{\mathsf{p}}(k) = 0$ *for* $k < n_\phi$ *and*

$$A_{\mathsf{p}}(k) := A_{L_{\mathsf{p}}}\mathsf{p}_\upsilon(k) - \sum_{n=n_\phi}^{\infty} \mathsf{p}_\upsilon(n)a(k - 2n) \quad \text{with} \quad \mathsf{p}_\upsilon(k) := \sum_{j=0}^{m-1} \frac{(-i)^j}{j!}\mathsf{p}^{(j)}(k)\widehat{\upsilon}^{(j)}(0), \quad k \geqslant n_\phi. \tag{5.29}$$

In fact, $A_{\mathsf{p}}(k) = 0$ *for all* $k \geqslant l_a + 2n_\phi$*, where* $[l_a, h_a] := \mathrm{fsupp}(a)$*.*

Proof. By the refinement equation $\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot - k)$, we deduce that

$$\phi(\cdot - n)\chi_{[0,\infty)} = 2\sum_{k\in\mathbb{Z}} a(k - 2n)\phi(2\cdot - k)\chi_{[0,\infty)} = 2\sum_{k\in\mathbb{Z}} a(k - 2n)\Big(\phi(\cdot - k)\chi_{[0,\infty)}\Big)(2\cdot).$$

Note that $\phi(\cdot - n)\chi_{[0,\infty)} = 0$ for all $n \leqslant -h_\phi$ and $\phi(\cdot - n)\chi_{[0,\infty)} = \phi(\cdot - n)$ for all $n \geqslant -l_\phi$.

For $1-h_\phi \leqslant n \leqslant n_\phi - 1$, by $n_\phi \geqslant -l_\phi$, we have

$$\phi(\cdot - n)\chi_{[0,\infty)} = 2\sum_{k=1-h_\phi}^{n_\phi-1} a(k-2n)\Big(\phi(\cdot-k)\chi_{[0,\infty)}\Big)(2\cdot) + 2\sum_{k=n_\phi}^{\infty} a(k-2n)\phi(2\cdot - k). \quad (5.30)$$

Therefore, (5.26) holds and we proved item (i).

To prove item (ii), by (5.24) we have $\mathsf{p}(x) = \phi^{\mathsf{p}}(x) + \sum_{k=n_\phi}^{\infty} \mathsf{p}_\upsilon(k)\phi(x-k)$ for all $x \in [0,\infty)$. Because $\mathsf{p}(x) = 2A_{L_\mathsf{p}}\mathsf{p}(2x)$ trivially holds for $x \in [0,\infty)$, we have $\mathsf{p} = 2A_{L_\mathsf{p}}\mathsf{p}(2\cdot) = 2A_{L_\mathsf{p}}\phi^{\mathsf{p}}(2\cdot) + \sum_{k=n_\phi}^{\infty} 2A_{L_\mathsf{p}}\mathsf{p}_\upsilon(k)\phi(2\cdot - k)$. By $n_\phi \geqslant \max(-l_\phi, -l_a)$, (5.13) must hold and then on $[0,\infty)$ we have

$$\begin{aligned}
\phi^{\mathsf{p}} &= \mathsf{p} - \sum_{n=n_\phi}^{\infty} \mathsf{p}_\upsilon(n)\phi(\cdot - n) = 2A_{L_\mathsf{p}}\phi^{\mathsf{p}}(2\cdot) + \sum_{k=n_\phi}^{\infty} 2A_{L_\mathsf{p}}\mathsf{p}_\upsilon(k)\phi(2\cdot - k) - \sum_{n=n_\phi}^{\infty} \mathsf{p}_\upsilon(n)\phi(\cdot - n) \\
&= 2A_{L_\mathsf{p}}\phi^{\mathsf{p}}(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A_{L_\mathsf{p}}\mathsf{p}_\upsilon(k)\phi(2\cdot - k) - 2\sum_{n=n_\phi}^{\infty}\sum_{k=n_\phi}^{\infty} \mathsf{p}_\upsilon(n)a(k-2n)\phi(2\cdot - k) \\
&= 2A_{L_\mathsf{p}}\phi^{\mathsf{p}}(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A_{\mathsf{p}}(k)\phi(2\cdot - k).
\end{aligned}$$

We now prove $A_\mathsf{p}(k) = 0$ for all $k \geqslant l_a + 2n_\phi$. Define the subdivision operator

$$[\mathcal{S}_a v](n) := 2\sum_{k\in\mathbb{Z}} v(k)a(n-2k) \quad \text{for} \quad n \in \mathbb{Z}.$$

We conclude from the definition of A_p in (5.29) that $A_\mathsf{p}(k) = A_{L_\mathsf{p}}\mathsf{p}_\upsilon(k) - 2^{-1}\mathcal{S}_a\mathsf{p}_\upsilon(k)$ for all $k \geqslant l_a + 2n_\phi$. Define $\widehat{u}(\xi) := \widehat{\upsilon}(2\xi)\widehat{a}(\xi)$. We deduce from (5.23) that

$$\widehat{u}^{(j)}(0) = \widehat{\upsilon}^{(j)}(0) \quad \text{and} \quad \widehat{u}^{(j)}(\pi) = 0, \qquad \forall\, j = 0,\ldots,m-1. \quad (5.31)$$

Since $\mathsf{p}_\upsilon = \mathsf{p} * \upsilon$, by [71, Theorem 1.2.4 and Lemma 1.2.1], we conclude from (5.31) that

$$\mathcal{S}_a\mathsf{p}_\upsilon = \mathcal{S}_u\mathsf{p} = (\mathsf{p}(2^{-1})) * u = 2A_{L_\mathsf{p}}(\mathsf{p} * u) = 2A_{L_\mathsf{p}}(\mathsf{p} * \upsilon) = 2A_{L_\mathsf{p}}\mathsf{p}_\upsilon.$$

This proves $A_\mathsf{p}(k) = A_{L_\mathsf{p}}\mathsf{p}_\upsilon(k) - 2^{-1}\mathcal{S}_a\mathsf{p}_\upsilon(k) = 0$ for all $k \geqslant l_a + 2n_\phi$. Therefore, item (ii) holds. □

5.2.4 Stability and construction of biorthogonal wavelets on $[0,\infty)$

We shall adopt the following notation:

$$\mathtt{S}_j(H) := \overline{\operatorname{span}\{f(2^j\cdot) \ : \ f \in H\}}, \qquad j \in \mathbb{Z}, H \subseteq L_2(\mathbb{R}), \tag{5.32}$$

where the overhead bar refers to closure in $L_2(\mathbb{R})$. For a countable subset H of $L_2(\mathbb{R})$ or $L_2([0,\infty))$, we define $\ell_2(H)$ to be the linear space of all sequences $\{c_h\}_{h\in H}$ satisfying $\sum_{h\in H}|c_h|^2<\infty$. For a Bessel sequence H in $L_2(\mathbb{R})$, we say that H is *ℓ_2-linearly independent* if $\sum_{h\in H} c_h h = 0$ for some $\{c_h\}_{h\in H}\in\ell_2(H)$, then we must have $c_h = 0$ for all $h\in H$.

In Theorem 5.2, it is necessary that Φ,Ψ in (5.2) and $\tilde\Phi,\tilde\Psi$ in (5.15) must be Riesz sequences in $L_2([0,\infty))$. We need the following result later for constructing wavelets on $[0,\infty)$.

Theorem 5.5. *Let $(\{\tilde\phi;\tilde\psi\},\{\phi;\psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ satisfying items (1)–(4) of Theorem 5.1. Let $\Phi := \{\phi^L\}\cup\{\phi(\cdot-k) \ : \ k\geqslant n_\phi\}\subseteq L_2([0,\infty))$ as in (5.2) with ϕ^L having compact support. Then the following statements are equivalent:*

(1) Φ is a Riesz sequence in $L_2([0,\infty))$, i.e., there exists a positive constant C such that

$$C^{-1}\sum_{h\in\Phi}|c_h|^2 \leqslant \Big\|\sum_{h\in\Phi}c_h h\Big\|^2_{L_2(\mathbb{R})} \leqslant C\sum_{h\in\Phi}|c_h|^2, \qquad \forall\,\{c_h\}_{h\in\Phi}\in\ell_2(\Phi). \tag{5.33}$$

(2) Φ is ℓ_2-linearly independent, i.e., $c_h=0$ for all $h\in\Phi$ if $\sum_{h\in\Phi}c_h h=0$ with $\{c_h\}_{h\in\Phi}\in\ell_2(\Phi)$.

(3) $\{\phi^L\}\cup\{\phi(\cdot-k) \ : \ n_\phi\leqslant k<N_\phi\}$ is linearly independent, where $N_\phi := \max(n_\phi, h_{\phi^L}-l_{\tilde\phi})$ with $[l_{\phi^L},h_{\phi^L}] := \operatorname{fsupp}(\phi^L)$ and $[l_{\tilde\phi},h_{\tilde\phi}] := \operatorname{fsupp}(\tilde\phi)$.

(4) There exists $\tilde H := \{\tilde\eta^L\}\cup\{\tilde\phi(\cdot-k) \ : \ k\geqslant N_\phi\}\subset L_2([0,\infty))$ such that $\tilde\eta^L$ has compact support, $\#\tilde\eta^L = \#\phi^L + (N_\phi - n_\phi)(\#\phi)$ and $\tilde H$ is biorthogonal to Φ, where N_ϕ is defined in item (3).

Moreover, for $\Phi=\{\phi^L\}\cup\{\phi(\cdot-k) \ : \ k\geqslant n_\phi\}$ such that item (3) fails, perform the following procedure:

(S1) Initially take $E := \{\phi(\cdot-k) \ : \ n_\phi\leqslant k<N_\phi\}$, which must be linearly independent.

(S2) Visit all elements $\eta\in\phi^L$ one by one: replace E by $E\cup\{\eta\}$ if $E\cup\{\eta\}$ is linearly independent; otherwise, delete η from ϕ^L.

Update $\phi^L := E\backslash\{\phi(\cdot - k) \ : \ n_\phi \leqslant k < N_\phi\}$, *i.e., the updated* ϕ^L *is obtained by removing redundant elements in the original* ϕ^L. *Then the new* Φ *is a Riesz sequence in* $L_2([0,\infty))$ *and preserves* $\mathtt{S}_0(\Phi)$.

Proof. (1)$\Longrightarrow$(2)$\Longrightarrow$(3) is obvious. Using a standard argument, we now prove (4)$\Longrightarrow$(1). Since ϕ^L and $\{\phi(\cdot - k) \ : \ k \in \mathbb{Z}\}$ are obviously Bessel sequences in $L_2(\mathbb{R})$, we conclude that Φ is a Bessel sequence in $L_2([0,\infty))$, i.e., there exists a positive constant C such that

$$\sum_{h\in\Phi} |\langle f, h\rangle|^2 \leqslant C\|f\|^2_{L_2(\mathbb{R})}, \qquad \forall\ f \in L_2(\mathbb{R}), \tag{5.34}$$

which is well known to be equivalent to the second inequality in (5.33). Similarly, $\tilde{H}$ must be a Bessel sequence in $L_2([0,\infty))$, i.e., (5.34) holds with Φ being replaced by $\tilde{H}$ (probably with a different constant C). For $\{c_h\}_{h\in\Phi} \in \ell_2(\Phi)$, define $f := \sum_{h\in\Phi} c_h h$. Using the biorthogonality in item (4), we have $c_h = \langle f, \tilde{h}\rangle$, where $\tilde{h}$ is the corresponding element of h in $\tilde{H}$. Now it follows from (5.34) with Φ being replaced by $\tilde{H}$ that the first inequality in (5.33) holds. This proves (4)$\Longrightarrow$(1).

To complete the proof, let us prove the key step (3)$\Longrightarrow$(4). Note that $\mathrm{fsupp}(\tilde{\phi}(\cdot - k)) = [k + l_{\tilde{\phi}}, k + h_{\tilde{\phi}}]$ for $k \in \mathbb{Z}$. Hence, for $k \geqslant h_{\phi^L} - l_{\tilde{\phi}}$, we have $k + l_{\tilde{\phi}} \geqslant h_{\phi^L}$ and trivially, $\langle \phi^L, \tilde{\phi}(\cdot - k)\rangle = 0$ due to their essentially disjoint supports. Let $\eta \in \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ n_\phi \leqslant k < N_\phi\}$. Then $\mathrm{fsupp}(\eta) \subseteq [0, N]$ with $N := \max(h_{\phi^L}, N_\phi + h_\phi - 1, 1)$. We now prove that there exists $d(\eta) \in L_2([0, N])$ such that

$$\langle d(\eta), \eta\rangle = 1 \quad \text{and} \quad \langle d(\eta), g\rangle = 0 \quad \forall\, g \in \Phi\backslash\{\eta\}. \tag{5.35}$$

Define $S := \{f \in L_2([0, N]) \ : \ \langle f, g\rangle = 0\ \forall\, g \in \Phi\backslash\{\eta\}\}$. For all integers $k \geqslant N - l_\phi$, we observe that $\mathrm{fsupp}(\phi(\cdot - k))$ is contained inside $[N, \infty)$ and consequently, $\phi(\cdot - k) \perp L_2([0, N])$ for all $k \geqslant N - l_\phi$. Therefore, we conclude that

$$S = \{f \in L_2([0, N]) \ : \ \langle f, g\rangle = 0\ \forall\, g \in \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ n_\phi \leqslant k < N - l_\phi\}, g \neq \eta\}.$$

Since $L_2([0, N])$ has infinite dimension, the above identity forces that S must have infinite dimension. In particular, S is not empty. We now prove that there must exist $d(\eta) \in S$ such that $\langle d(\eta), \eta\rangle = 1$. Suppose not. Then $\eta \perp S$. By the definition of the space S, we must have $\Phi \perp L_2([0, N])$, which forces $\eta = 0$ by $\eta \in \Phi \cap L_2([0, N])$. This contradicts our assumption in item (3). Hence, we proved the existence of $d(\eta) \in S$ such that $\langle d(\eta), \eta\rangle = 1$. Now it is straightforward to check that (5.35) holds. Let $\tilde{\eta}^L$ be the vector/set of all elements $d(\eta)$ for $\eta \in \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ n_\phi \leqslant k < N_\phi\}$. Then the derived $\tilde{H}$ must be biorthogonal

to Φ. This proves (3)$\Longrightarrow$(4).

Suppose that $\vec{c}\phi^L + \sum_{k=n_\phi}^{N_\phi-1} c_k\phi(\cdot-k) = 0$. If $\vec{c} = 0$, then $\sum_{k=n_\phi}^{N_\phi-1} c_k\phi(\cdot-k) = 0$ and by biorthogonality of integer shifts of ϕ and $\tilde{\phi}$, we conclude that $c_k = 0$ for all $n_\phi \leqslant k < N_\phi$. Hence, if item (3) fails, then $\vec{c} \neq 0$ and we can remove the redundant entry in ϕ^L corresponding to a nonzero entry in $\vec{c}$. In this way, the new Φ with fewer elements in ϕ^L can satisfy item (3) and preserves $\mathsf{S}_0(\Phi)$. □

Theorem 5.5 can be applied to all Φ, Ψ in (5.2) and $\tilde{\Phi}, \tilde{\Psi}$ in (5.15). If any of $\Phi, \Psi, \tilde{\Phi}, \tilde{\Psi}$ is not ℓ_2-linearly independent, by Theorem 5.5, then we can always remove the redundant elements in $\phi^L, \psi^L, \tilde{\phi}^L, \tilde{\psi}^L$ to turn $\Phi, \Psi, \tilde{\Phi}, \tilde{\Psi}$ into Riesz sequences in $L_2([0,\infty))$.

To study the Bessel property of $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$, we need the definition of the Sobolev space $H^\tau(\mathbb{R})$ with $\tau \in \mathbb{R}$. Recall that the Sobolev space $H^\tau(\mathbb{R})$ with $\tau \in \mathbb{R}$ consists of all tempered distributions f on $\mathbb{R}$ such that $\int_{\mathbb{R}} |\widehat{f}(\xi)|^2(1+|\xi|^2)^\tau d\xi < \infty$.

Theorem 5.6. *Let ϕ be an $r \times 1$ vector of compactly supported functions in $L_2(\mathbb{R})$ such that $\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot-k)$ for some finitely supported sequence $a \in (l_0(\mathbb{Z}))^{r\times r}$. Let ψ be a vector of compactly supported functions in $L_2(\mathbb{R})$. Define $[l_\phi, h_\phi] := \mathrm{fsupp}(\phi)$ and $[l_\psi, h_\psi] := \mathrm{fsupp}(\psi)$. Let $n_\phi \geqslant \max(-l_\phi, -l_a)$ and $n_\psi \geqslant -l_\psi$. Define Φ, Ψ as in (5.2) with finite subsets $\phi^L \cup \psi^L$ of compactly supported functions in $L_2([0,\infty))$. If ψ has at least one vanishing moment (i.e., $\widehat{\psi}(0) = \int_{\mathbb{R}} \psi(x)dx = 0$) and $\psi \cup \phi^L \cup \psi^L \subseteq H^\tau(\mathbb{R})$ for some $\tau > 0$ (this latter technical condition always holds if each element in $\psi \cup \phi^L \cup \psi^L$ is a finite linear combination of $\phi(2^j\cdot-k)\chi_{[0,\infty)}$ with $j,k \in \mathbb{Z}$), then $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ must be a Bessel sequence in $L_2([0,\infty))$ for all $J \in \mathbb{Z}$, that is, there exists a positive constant C, which is independent of J, such that*

$$\sum_{h\in\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}} |\langle f, h\rangle|^2 \leqslant C\|f\|^2_{L_2([0,\infty))}, \qquad \forall\, f \in L_2([0,\infty)). \tag{5.36}$$

Proof. Let ϕ^c be defined as in Proposition 5.4. Let $\mathring{\phi}$ be a vector function obtained by appending ϕ to ϕ^c. By the refinement equation $\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot-k)$ and (5.26), it is straightforward to see that the vector function $\mathring{\phi}$ is a compactly supported refinable vector function with a finitely supported matrix-valued filter. Because all entries in $\mathring{\phi}$ belong to $L_2(\mathbb{R})$ and have compact support, we conclude by [71, Corollary 5.8.2 or Corollary 6.3.4] (also see [67, Theorem 2.2]) that there exists $\tau > 0$ such that every entry in $\mathring{\phi}$ belongs to $H^\tau(\mathbb{R})$. In particular, we conclude that all the entries in $\phi \cup \phi^c$ belong to $H^\tau(\mathbb{R})$. Note that $\phi(\cdot-n)\chi_{[0,\infty)} = 0$ for all $n \leqslant -h_\phi$ and $\phi(\cdot-n)\chi_{[0,\infty)} = \phi(\cdot-n)$ for all $n \geqslant -l_\phi$. Consequently, by $n_\phi \geqslant -l_a$, $\phi(\cdot-n)\chi_{[0,\infty)} \in H^\tau(\mathbb{R})$ for all $n \in \mathbb{Z}$. Hence, all elements of $\phi(2^j\cdot-k)\chi_{[0,\infty)}$ with $j,k \in \mathbb{Z}$ must belong to $H^\tau(\mathbb{R})$. In particular, if each element in

$\psi \cup \phi^L \cup \psi^L$ is a finite linear combination of $\phi(2^j \cdot -k)\chi_{[0,\infty)}$ with $j, k \in \mathbb{Z}$, then we must have $\phi \cup \psi \cup \phi^L \cup \psi^L \subseteq H^\tau(\mathbb{R})$. Since $\phi \cup \psi \subseteq H^\tau(\mathbb{R})$ with $\tau > 0$ and ψ has at least one vanishing moment, by [71, Corollary 4.6.6] or [67, Theorem 2.3], $\mathsf{AS}_0(\phi;\psi)$ must be a Bessel sequence in $L_2(\mathbb{R})$, that is, there exists a positive constant C such that $\sum_{h\in\mathsf{AS}_0(\phi;\psi)} |\langle f, h\rangle|^2 \leqslant C\|f\|^2_{L_2(\mathbb{R})}$ for all $f \in L_2(\mathbb{R})$. Since all the elements in $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ but not in $\mathsf{AS}_0(\phi;\psi)$ are ϕ^L and $\psi^L_{j;0} := 2^{j/2}\psi^L(2^j\cdot)$ for $j \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, to prove (5.36) with $J = 0$ and C being replaced by $2C$, it suffices to prove

$$\|\langle f, \phi^L\rangle\|^2_{l_2} + \sum_{j=0}^{\infty} \|\langle f, \psi^L_{j;0}\rangle\|^2_{l_2} \leqslant C\|f\|^2_{L_2([0,\infty))}, \qquad \forall\, f \in L_2([0,\infty)). \tag{5.37}$$

Define $\phi^S := \phi^L - \phi^L(-\cdot)$ and $\psi^S := \psi^L - \psi^L(-\cdot)$. Since $\phi^L \cup \psi^L \subseteq H^\tau(\mathbb{R})$ with $\tau > 0$, all elements in $\phi^S \cup \psi^S$ belong to $H^\tau(\mathbb{R})$ and have compact support with one vanishing moment. Now we conclude from [71, Corollary 4.6.6] or [67, Theorem 2.3] that there exists a positive constant C such that

$$\sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}} \|\langle f, \phi^S_{j;k}\rangle\|^2_{l_2} + \sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}} \|\langle f, \psi^S_{j;k}\rangle\|^2_{l_2} \leqslant C\|f\|^2_{L_2(\mathbb{R})}, \qquad \forall\, f \in L_2(\mathbb{R}),$$

where $\psi^S_{j;k} := 2^{j/2}\psi^S(2^j\cdot -k)$. For $f \in L_2([0,\infty))$, we have $\mathrm{fsupp}(f) \subseteq [0,\infty)$ and trivially $\langle f, \psi^S_{j;0}\rangle = \langle f, \psi^L_{j;0}\rangle$. Consequently, it follows directly from the above inequality that (5.37) must hold and hence (5.36) holds for $J = 0$. By a simple scaling argument (e.g., see [70, Proposition 4 and (2.6)] and [71, Theorem 4.3.3]), (5.36) must hold for all $J \in \mathbb{Z}$ with the same constant C. □

Based on Theorems 5.1, 5.2 and 5.6, we now present the following result, whose proof is given in Section 5.8, for constructing compactly supported biorthogonal wavelets on $[0,\infty)$.

Theorem 5.7. *Let $(\{\tilde\phi;\tilde\psi\},\{\phi;\psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a biorthogonal wavelet filter bank $(\{\tilde a;\tilde b\},\{a;b\})$ satisfying items (1)–(4) of Theorem 5.1. Define integers l_ϕ, l_ψ, l_a, h_a as in (5.8) and $l_{\tilde\phi}, l_{\tilde\psi}, l_{\tilde a}, l_{\tilde b}$ as in (5.9). Let $\phi^L, \psi^L, \tilde\phi^L, \tilde\psi^L$ be finite sets of compactly supported functions in $L_2([0,\infty))$. Let $n_\phi, n_\psi, n_{\tilde\phi}, n_{\tilde\psi}$ be integers satisfying (5.16). Define Φ, Ψ as in (5.2) and $\tilde\Phi, \tilde\Psi$ as in (5.15). Assume that $\phi^L \cup \psi^L \cup \tilde\phi^L \cup \tilde\psi^L \subseteq H^\tau(\mathbb{R})$ for some $\tau > 0$ and*

- *(i) $\Phi \subset L_2([0,\infty))$ satisfies both (5.13) and (5.17) for some matrix A_L and some finitely supported sequence A of matrices.*
- *(ii) $\tilde\Phi \subset L_2([0,\infty))$ is biorthogonal to Φ, and $\tilde\Phi$ satisfies both (5.19) and (5.21) for some matrix $\tilde A_L$ and some finitely supported sequence $\tilde A$ of matrices.*

(iii) $\mathsf{S}_0(\Psi) = \mathsf{S}_1(\Phi) \cap (\mathsf{S}_0(\tilde{\Phi}))^{\perp}$ *and* Ψ *satisfies both* (5.14) *and* (5.18) *for some matrix* B_L *and some finitely supported sequence* B *of matrices.*

(iv) $\mathsf{S}_0(\tilde{\Psi}) = \mathsf{S}_1(\tilde{\Phi}) \cap (\mathsf{S}_0(\Phi))^{\perp}$, $\tilde{\Psi}$ *is biorthogonal to* Ψ, *and* $\tilde{\Psi}$ *satisfies* (5.20) *and* (5.22) *for some matrix* $\tilde{B}_L$ *and some finitely supported sequence* $\tilde{B}$ *of matrices.*

Then $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ *and* $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$, *as defined in* (5.1), *form a pair of compactly supported biorthogonal Riesz bases in* $L_2([0,\infty))$ *for every* $J \in \mathbb{Z}$.

5.3 Classical approach for constructing biorthogonal wavelets on $[0,\infty)$

The main ingredients of the classical approach for constructing (bi)orthogonal wavelets on intervals are outlined in items (i)–(iv) of Theorem 5.7. Most papers in the current literature (e.g., [4, 5, 6, 18, 19, 20, 25, 31, 33, 34, 72, 80, 88, 99, 102, 113, 114]) employ (variants of) the classical approach to construct particular wavelets on intervals from special (bi)orthogonal wavelets on the real line such as Daubechies orthogonal wavelets in [38] and spline biorthogonal scalar wavelets in [29]. From any arbitrarily given compactly supported biorthogonal (multi)wavelets on the real line, the main goal of this section is to follow the classical approach outlined in Theorem 5.7 for constructing all possible compactly supported biorthogonal wavelets $(\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}, \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ on the interval $[0,\infty)$ with or without vanishing moments and polynomial reproduction, under the restriction that every boundary element in Φ (or $\tilde{\Phi}$) is a finite linear combination of $\phi(\cdot-k)\chi_{[0,\infty)}$ (or $\tilde{\phi}(\cdot-k)\chi_{[0,\infty)}$) with $k \in \mathbb{Z}$. As we shall see in this section, though adapting orthogonal (multi)wavelets from the real line to $[0,\infty)$ is easy, constructing general compactly supported biorthogonal (multi)wavelets on $[0,\infty)$ is often much more involved and complicated than their orthogonal counterparts. The complexity of the classical approach in this section also motivates us to propose a direct approach in Section 5.4 to construct all possible biorthogonal (multi)wavelets on $[0,\infty)$ without explicitly constructing the dual refinable functions $\tilde{\Phi}$ and dual wavelets $\tilde{\Psi}$, while removing the restrictions on the boundary elements in Φ and $\tilde{\Phi}$.

5.3.1 Construct refinable Φ satisfying item (i) of Theorem 5.7

Though elements in ϕ^L in Theorem 5.2 could be any compactly supported functions in $L_2([0,\infty))$, in this section we only consider particular ϕ^L. The general case $\phi^L \subset L_2([0,\infty))$ will be addressed later in Theorem 5.13.

To satisfy item (i) of Theorem 5.7, we have only two conditions (5.13) and (5.17) on $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \,:\, k \geqslant n_\phi\}$. As we discussed before, (5.13) is trivially true by choosing any integer n_ϕ satisfying $n_\phi \geqslant \max(-l_\phi, -l_a)$. So, the main task is to construct ϕ^L to satisfy (5.17). By Proposition 5.4, there are three straightforward choices of ϕ^L satisfying (5.17):

(C1) $\phi^L = \phi^c$ in item (i) of Proposition 5.4 satisfies (5.17) with $A_L = A_{L_c}$ and $A = A_c$ in (5.26);

(C2) $\phi^L = \phi^{\mathsf{p}}$ in (5.27) in item (ii) of Proposition 5.4 satisfies (5.17) with $A_L = A_{L_{\mathsf{p}}}$ and $A = A_{\mathsf{p}}$, where $\mathsf{p}(x) = (x^{j_0}, \dots, x^{j_\ell})^{\mathsf{T}}$ such that $\{j_0, \dots, j_\ell\} \subseteq \{0, \dots, m-1\}$ with $m := \mathrm{sr}(a)$;

(C3) $\phi^L := 2\sum_{k=n_\phi}^{\infty} A(k)\phi(2\cdot - k)$ with a finitely supported sequence A satisfies (5.17) with $A_L = 0$.

Through the following breaking and merging steps, many new ϕ^L satisfying (5.17) can be obtained from the finite-dimensional space generated by known/given ϕ^L such as in (C1)–(C3):

(BS) Breaking Step: For ϕ^L satisfying (5.17), write $A_L = C^{-1}\,\mathrm{diag}(J_1, \dots, J_N)C$ in its Jordan normal form and define $(\phi^{L_{J_\ell}})_{1\leqslant \ell \leqslant N} := C\phi^L$ with $\#\phi^{L_{J_\ell}} = N_{J_\ell} \in \mathbb{N}$, where J_ℓ is an $N_{J_\ell} \times N_{J_\ell}$ (basic Jordan block) matrix given by

$$J_\ell := \begin{bmatrix} \lambda_\ell & 1 & 0 & \cdots & 0 \\ 0 & \lambda_\ell & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_\ell & 1 \\ 0 & 0 & 0 & \cdots & \lambda_\ell \end{bmatrix}, \qquad \lambda_\ell \in \mathbb{C}. \tag{5.38}$$

Then every $\phi^{L_{J_\ell}}$ and its truncated vector functions by throwing away the first n entries of $\phi^{L_{J_\ell}}$ with $1 \leqslant n < N_{J_\ell}$ satisfy (5.17) and $\mathrm{span}(\phi^L) = \mathrm{span}(\cup_{\ell=1}^N \phi^{L_{J_\ell}})$, since

$$\begin{bmatrix} \phi^{L_{J_1}} \\ \vdots \\ \phi^{L_{J_N}} \end{bmatrix} = 2\begin{bmatrix} J_1\phi^{L_{J_1}}(2\cdot) \\ \vdots \\ J_N\phi^{L_{J_N}}(2\cdot) \end{bmatrix} + 2\sum_{k=n_\phi}^{\infty} CA(k)\phi(\cdot - k).$$

(MS) Merging Step: For vector functions ϕ^{L_1} and ϕ^{L_2} satisfying (5.17), i.e.,

$$\phi^{L_1} = 2A_{L_1}\phi^{L_1}(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A_1(k)\phi(2\cdot - k)$$

and

$$\phi^{L_2} = 2A_{L_2}\phi^{L_2}(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A_2(k)\phi(2\cdot - k),$$

then $\phi^L := \phi^{L_1} \cup \phi^{L_2}$ satisfies (5.17) with $A_L := \operatorname{diag}(A_{L_1}, A_{L_2})$ and $A(k) = [A_1(k)^\mathsf{T}, A_2(k)^\mathsf{T}]^\mathsf{T}$ for all $k \geqslant n_\phi$.

Note that we can always add ϕ^p to ϕ^L by the merging step (MS) for polynomial reproduction. $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\}$ satisfying item (i) of Theorem 5.7 is not necessarily a Riesz sequence in $L_2([0,\infty))$. However, we can always perform the procedure in Theorem 5.5 to remove redundant elements in ϕ^L so that the new Φ is a Riesz sequence and still satisfies item (i) of Theorem 5.7. The particular choice of $\phi^L = \phi^\mathsf{p}$ in item (ii) of Proposition 5.4 with $\mathsf{p}(x) = (1, x, \ldots, x^{m-1})^\mathsf{T}$ and $m := \operatorname{sr}(a)$ was first considered in [31] for Daubechies orthogonal wavelets in [38] and used in [34, 100] for spline biorthogonal scalar wavelets. The particular choice $\phi^L = \phi^c$ in item (i) of Proposition 5.4 was originally employed in [102] for Daubechies orthogonal wavelets and used in [19, 114] for spline biorthogonal scalar wavelets with improved condition numbers.

We further study some properties of Φ satisfying item (i) of Theorem 5.7 in the following result, which is useful later for us to construct Ψ satisfying item (iii) of Theorem 5.7.

Lemma 5.8. *Let $(\{\tilde\phi; \tilde\psi\}, \{\phi; \psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a finitely supported biorthogonal wavelet filter bank $(\{\tilde a; \tilde b\}, \{a; b\})$ satisfying items (1)–(4) of Theorem 5.1. Suppose that $\Phi := \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\} \subset L_2([0,\infty))$ with ϕ^L having compact support satisfies item (i) of Theorem 5.7. For any integer $n_\psi \in \mathbb{Z}$ satisfying (5.14), define*

$$m_\phi := \max(2n_\phi + h_{\tilde a}, 2n_\psi + h_{\tilde b}), \tag{5.39}$$

where $[l_{\tilde a}, h_{\tilde a}] := \operatorname{fsupp}(\tilde a)$ and $[l_{\tilde b}, h_{\tilde b}] := \operatorname{fsupp}(\tilde b)$, and define

$$H := \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\} \cup \{\psi(\cdot - k) \ : \ k \geqslant n_\psi\}, \tag{5.40}$$

then

$$\phi(2\cdot - k_0) \text{ is a finite linear combination of elements in } H \text{ for all } k_0 \geqslant m_\phi \tag{5.41}$$

and the finite-dimensional quotient space $\mathsf{S}_1(\Phi)/\mathsf{S}_0(H)$ has a basis, which can be selected from $\{\phi^L(2\cdot)\} \cup \{\phi(2\cdot - k) \ : \ n_\phi \leqslant k < m_\phi\}$.

Proof. Using (1.2) with $f = \phi(2\cdot - k_0)$ and $J = 0$, we deduce from (5.6) that for all $k_0 \in \mathbb{Z}$,

$$\phi(2\cdot - k_0) = \sum_{k=\lceil \frac{k_0 - h_{\tilde{a}}}{2} \rceil}^{\lfloor \frac{k_0 - l_{\tilde{a}}}{2} \rfloor} \overline{\tilde{a}(k_0 - 2k)}^{\mathsf{T}} \phi(\cdot - k) + \sum_{k=\lceil \frac{k_0 - h_{\tilde{b}}}{2} \rceil}^{\lfloor \frac{k_0 - l_{\tilde{b}}}{2} \rfloor} \overline{\tilde{b}(k_0 - 2k)}^{\mathsf{T}} \psi(\cdot - k). \tag{5.42}$$

By the definition of m_ϕ in (5.39), we have $\frac{k_0 - h_{\tilde{a}}}{2} \geqslant n_\phi$ and $\frac{k_0 - h_{\tilde{b}}}{2} \geqslant n_\psi$ for all $k_0 \geqslant m_\phi$. Hence, (5.42) implies (5.41). By (5.13) and (5.17) in item (i) of Theorem 5.7, we have $H \subseteq \mathtt{S}_1(\Phi)$ and hence $\mathtt{S}_0(H) \subseteq \mathtt{S}_1(\Phi)$. So, $\mathtt{S}_1(\Phi)/\mathtt{S}_0(H)$ is well defined. By (5.41), $\mathtt{S}_0(H) + \mathtt{S}_0(\{\phi^L(2\cdot)\} \cup \{\phi(2\cdot - k)\}_{k=n_\phi}^{m_\phi - 1}) = \mathtt{S}_1(\Phi)$. Hence, the dimension of $\mathtt{S}_1(\Phi)/\mathtt{S}(H)$ is no more than $(\#\phi^L) + (m_\phi - n_\phi)(\#\phi) < \infty$. □

5.3.2 Construction of orthogonal wavelets on $[0, \infty)$

We call $\{\phi; \psi\}$ *an orthogonal wavelet in* $L_2(\mathbb{R})$ if $(\{\phi; \psi\}, \{\phi; \psi\})$ is a biorthogonal wavelet in $L_2(\mathbb{R})$, i.e., $\mathsf{AS}_0(\phi; \psi)$ is an orthonormal basis of $L_2(\mathbb{R})$. Similarly, we call $\{a; b\}$ *an orthogonal wavelet filter bank* if $(\{a; b\}, \{a; b\})$ is a biorthogonal wavelet filter bank. As a direct consequence of Lemma 5.8 and Theorems 5.5 and 5.7, we have

Algorithm 5.1. *Let $\{\phi; \psi\}$ be a compactly supported orthogonal wavelet in $L_2(\mathbb{R})$ associated with a finitely supported orthogonal wavelet filter bank $\{a; b\}$.*

(S1) Construct $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\} \subseteq L_2([0, \infty))$ (e.g., by Section 5.3.1 or Theorem 5.13) such that item (i) of Theorem 5.7 holds but Φ is not necessarily a Riesz sequence in $L_2([0, \infty))$.

(S2) Apply the following Gram-Schmidt orthonormalization procedure to Φ:

- *(1) Initially take $E := \{\phi(\cdot - k) \ : \ n_\phi \leqslant k < N_\phi\}$ with $N_\phi := \max(n_\phi, h_{\phi^L} - l_\phi)$, where $[l_{\phi^L}, h_{\phi^L}] := \mathrm{fsupp}(\phi^L)$ and $[l_\phi, h_\phi] := \mathrm{fsupp}(\phi)$;*
- *(2) Visit all elements $\eta \in \phi^L$ one by one: replace E by $E \cup \{\mathring{\eta}/\|\mathring{\eta}\|_{L_2(\mathbb{R})}\}$ if $\|\mathring{\eta}\|_{L_2(\mathbb{R})} \neq 0$, where $\mathring{\eta} := \eta - \sum_{h \in E} \langle \eta, h \rangle h$; otherwise, delete η from ϕ^L;*
- *(3) Update/redefine $\phi^L := E \backslash \{\phi(\cdot - k) \ : \ n_\phi \leqslant k < N_\phi\}$.*

Then Φ with the updated boundary vector function ϕ^L is an orthonormal system in $L_2([0, \infty))$.

(S3) Select an integer n_ψ satisfying (5.14) and $\langle \phi^L, \psi(\cdot - k) \rangle = 0$ for all $k \geqslant n_\psi$. For example, we can choose any integer n_ψ satisfying $n_\psi \geqslant \max(-l_\psi, \frac{n_\phi - l_b}{2}, h_{\phi^L})$ with $[l_\psi, h_\psi] :=$

$\mathrm{fsupp}(\psi)$ *and* $[l_b, h_b] := \mathrm{fsupp}(b)$. *Define* $\mathring{\psi}^L := \{\phi^L(2\cdot)\} \cup \{\phi(2\cdot - k) \ : \ n_\phi \leqslant k < m_\phi\}$ *with* $m_\phi := \max(2n_\phi + h_a, 2n_\psi + h_b)$. *Define* $M_\phi := \lceil \max(\frac{h_{\phi^L}}{2}, \frac{h_\phi + m_\phi - 1}{2}) \rceil - l_\phi$ *and calculate*

$$\psi^L := \mathring{\psi}^L - \langle \mathring{\psi}^L, \phi^L \rangle \phi^L - \sum_{k=n_\phi}^{M_\phi - 1} \langle \mathring{\psi}^L, \phi(\cdot - k) \rangle \phi(\cdot - k).$$

(S4) Update ψ^L *by applying the similar Gram-Schmidt orthonormalization procedure in (S2) to* $\{\psi^L\} \cup \{\psi(\cdot - k) \ : \ n_\psi \leqslant k < N_\psi\}$ *with* N_ϕ *being replaced by* $N_\psi := \max(n_\psi, h_{\psi^L} - l_{\tilde{\psi}})$.

Then $\mathsf{AS}_J(\Phi; \Psi)_{[0,\infty)}$ *in* (5.1) *with* $\Psi := \{\psi^L\} \cup \{\psi(\cdot - k) \ : \ k \geqslant n_\psi\}$ *is an orthonormal basis of* $L_2([0,\infty))$ *for all* $J \in \mathbb{Z}$. *Moreover, if we append* ϕ^{p} *with* $\mathsf{p}(x) := (1, x, \ldots, x^{\mathrm{sr}(a)-1})^{\mathsf{T}}$ *in item (ii) of Proposition* 5.4 *to* ϕ^L *in (S1), then all elements in* Ψ *must have* $\mathrm{sr}(a)$ *vanishing moments.*

The choice of M_ϕ for defining ψ^L in (S3) of Algorithm 5.1 guarantees that $\mathrm{supp}(\mathring{\psi}^L) \cap \mathrm{fsupp}(\phi(\cdot - k))$ is at most a singleton and hence $\langle \mathring{\psi}^L, \phi(\cdot - k) \rangle = 0$ for all $k \geqslant M_\phi$. See Theorem 5.10 below for calculating inner products in the Gram-Schmidt orthonormalization procedure in (S2) and (S4) of Algorithm 5.1. Though biorthogonal wavelets on the real line are flexible to design ([65, 71]) and important in many applications, as we shall see later, the construction of biorthogonal wavelets on $[0,\infty)$ is much more involved than their orthogonal counterparts.

As we shall discuss in Section 5.5, if an orthogonal (multi)wavelet on $[0,\infty)$ satisfies homogeneous boundary conditions, then some of its boundary elements often can have no or low order of vanishing moments. The following result is a special case of Theorem 5.17.

Corollary 5.9. *Suppose that* $\mathsf{AS}_0(\Phi; \Psi)_{[0,\infty)}$ *is an orthonormal basis of* $L_2([0,\infty))$ *such that all elements of* $\Phi \cup \Psi$ *are compactly supported and are continuous near* 0. *If* $\eta(0) = 0$ *for all* $\eta \in \Psi$, *then there must exist some* $\eta \in \Psi$ *such that* η *does not have any vanishing moments, i.e.,* $\int_0^\infty \eta(x)dx \neq 0$.

To illustrate the complexity of wavelets on intervals, let us present an "abnormal" example here.

Example 5.1. Let $N \in \mathbb{N}$ be an arbitrary integer. Let $\phi = \chi_{[0,1]}$ and $\psi = \chi_{(0,1/2]} - \chi_{(1/2,1]}$. Then $\{\phi; \psi\}$ is the well-known Haar orthogonal wavelet in $L_2(\mathbb{R})$. Define $\phi^L := \emptyset$ and $n_\phi := 2N$. Then $\Phi = \{\phi(\cdot - k) \ : \ k \geqslant 2N\}$ obviously satisfies item (i) of Theorem 5.7. Define $\psi^L = \{\phi(\cdot - k) \ : \ N \leqslant k < 2N\}$ and $n_\psi := N$. For every $J \in \mathbb{Z}$, then $\mathsf{AS}_J(\Phi; \Psi)_{[0,\infty)}$ with $\Psi := \{\psi^L\} \cup \{\psi(\cdot - k) \ : \ k \geqslant N\}$ must be an orthonormal basis in $L_2([0,\infty))$ such that all elements in $\Phi \cup \Psi$ are supported inside $[N,\infty)$ and the boundary wavelet ψ^L does not

have any vanishing moments. This appears weird but not surprising at all. For simplicity, we present a self-contained proof here only for $J = 0$ and the general case follows by a simple scaling argument. It is easy to directly check that $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is an orthonormal system in $L_2([0,\infty))$ by noting that Ψ is perpendicular to Φ and

$$\mathsf{S}_0(\Phi) \oplus \mathsf{S}_0(\Psi) = \mathsf{S}_0(\Phi \cup \psi^L) \oplus \mathsf{S}_0(\{\psi(\cdot - k) \ : \ k \geqslant N\}) = \mathsf{S}_1(\Phi). \tag{5.43}$$

To prove that $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is dense in $L_2([0,\infty))$, using (5.43), we observe that for any $n \in \mathbb{N}$,

$$\mathsf{S}_n(\Phi) = \mathsf{S}_0(\Phi) \oplus \mathsf{S}_0(\Psi) \oplus \mathsf{S}_1(\Psi) \oplus \cdots \oplus \mathsf{S}_{n-1}(\Psi) \subseteq \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}.$$

Note that $\mathsf{S}_n(\Phi) = \mathsf{S}_0(\Phi(2^n\cdot)) = \mathsf{S}_0(\{\phi(2^n(\cdot - 2^{-n}k)) \ : \ k \geqslant 2N\})$. Because $\lim_{n\to\infty} 2^{-n}N = 0$, we now conclude that $\cup_{n=1}^{\infty}\mathsf{S}_n(\Phi)$ is indeed dense in $L_2([0,\infty))$. Hence, we proved that $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is dense in $L_2([0,\infty))$ and thus is an orthonormal basis in $L_2([0,\infty))$. Similar examples can be constructed from any compactly supported orthogonal (multi)wavelets using Algorithm 5.1.

To perform the Gram-Schmidt orthonormalization procedure in (S2) and (S4) of Algorithm 5.1, we have the following result (see Section 5.8 for its proof) to compute $\int_0^1 \tilde{\phi}(x-m)\overline{\phi(x-n)}^{\mathsf{T}}\,dx$ for all $m, n \in \mathbb{Z}$ from any two arbitrary compactly supported refinable vector functions.

Theorem 5.10. *Let $\phi, \tilde{\phi}$ be two $r \times 1$ vectors of compactly supported functions in $L_2(\mathbb{R})$ such that $\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot - k)$ and $\tilde{\phi} = 2\sum_{k\in\mathbb{Z}} \tilde{a}(k)\tilde{\phi}(2\cdot - k)$ for some finitely supported filters $a, \tilde{a} \in (l_0(\mathbb{Z}))^{r\times r}$. Assume that $\widehat{\phi}(0) \neq 0$ and $\widehat{\tilde{\phi}}(0) \neq 0$. Define $[l_\phi, h_\phi] := \mathrm{fsupp}(\phi)$ and $[l_{\tilde{\phi}}, h_{\tilde{\phi}}] := \mathrm{fsupp}(\tilde{\phi})$.*

(S1) Define two vector functions by $\vec{\phi} := [\phi(\cdot - 1 + h_\phi)\chi_{[0,1]}, \ldots, \phi(\cdot + l_\phi)\chi_{[0,1]}]^{\mathsf{T}}$ and $\vec{\tilde{\phi}} := [\tilde{\phi}(\cdot - 1 + h_{\tilde{\phi}})\chi_{[0,1]}, \ldots, \tilde{\phi}(\cdot + l_{\tilde{\phi}})\chi_{[0,1]}]^{\mathsf{T}}$. Then

$$\vec{\phi} = 2A_0\vec{\phi}(2\cdot) + 2A_1\vec{\phi}(2\cdot - 1) \quad \text{and} \quad \vec{\tilde{\phi}} = 2\tilde{A}_0\vec{\tilde{\phi}}(2\cdot) + 2\tilde{A}_1\vec{\tilde{\phi}}(2\cdot - 1) \tag{5.44}$$

with $A_\gamma := [a(k+\gamma-2j)]_{1-h_\phi\leqslant j,k\leqslant -l_\phi}$ and $\tilde{A}_\gamma := [\tilde{a}(k+\gamma-2j)]_{1-h_{\tilde{\phi}}\leqslant j,k\leqslant -l_{\tilde{\phi}}}$ for $\gamma = 0, 1$, where j is for the row index and k is for the column index.

(S2) If all the entries in $\vec{\phi}$ are not linearly independent on $[0,1]$, then we delete as many entries as possible from $\vec{\phi}$ so that all the deleted entries are linear combinations of entries kept. Do the same for $\vec{\tilde{\phi}}$. Then (5.44) still holds with $A_0, A_1, \tilde{A}_0$ and $\tilde{A}_1$ being appropriately modified.

(S3) Define $M := \langle \vec{\tilde{\phi}}, \vec{\phi} \rangle := \int_0^1 \vec{\tilde{\phi}}(x)\overline{\vec{\phi}(x)}^{\mathsf{T}}\,dx$. *Then the matrix* M *is uniquely determined by the system of linear equations given by*

$$M = 2\tilde{A}_0 M \overline{A_0}^{\mathsf{T}} + 2\tilde{A}_1 M \overline{A_1}^{\mathsf{T}} \tag{5.45}$$

under the normalization condition

$$\vec{\tilde{v}} M \overline{\vec{v}}^{\mathsf{T}} = 1, \tag{5.46}$$

where $\vec{v}$ *is the unique row vector satisfying* $\vec{v}(A_0 + A_1) = \vec{v}$ *and* $\vec{v}\widehat{\vec{\phi}}(0) = 1$, *while similarly* $\vec{\tilde{v}}$ *is the unique row vector satisfying* $\vec{\tilde{v}}(\tilde{A}_0 + \tilde{A}_1) = \vec{\tilde{v}}$ *and* $\vec{\tilde{v}}\widehat{\vec{\tilde{\phi}}}(0) = 1$.

Define $M_{\tilde{\phi},\phi} := (\int_0^1 \tilde{\phi}(x-j)\overline{\phi(x-k)}^{\mathsf{T}}\,dx)_{1-h_{\tilde{\phi}} \leqslant j \leqslant -l_{\tilde{\phi}}, 1-h_\phi \leqslant k \leqslant -l_\phi}$. *If (S2) is not performed, then* $M_{\tilde{\phi},\phi}$ *agrees with* M *as in (S3); Otherwise, we obtain* $M_{\tilde{\phi},\phi}$ *from* M *using the linear combinations in (S2). Hence, all integrals* $\int_0^1 \tilde{\phi}(x-j)\overline{\phi(x-k)}^{\mathsf{T}}\,dx$ *for* $j, k \in \mathbb{Z}$ *can be obtained from* $M_{\tilde{\phi},\phi}$.

Suppose that a has order one sum rule with a matching filter $\upsilon \in (l_0(\mathbb{Z}))^{1\times r}$ and $\widehat{\upsilon}(0)\widehat{\phi}(0) = 1$ and $\tilde{a}$ has order one sum rule with a matching filter $\tilde{\upsilon} \in (l_0(\mathbb{Z}))^{1\times r}$ and $\widehat{\tilde{\upsilon}}(0)\widehat{\tilde{\phi}}(0) = 1$. Then by [71, Proposition 5.6.2], we have $\widehat{\upsilon}(0)\widehat{\phi}(2\pi k) = 0$ for all $k \in \mathbb{Z}\backslash\{0\}$. By Poisson summation formula, we must have $\widehat{\upsilon}(0)\sum_{k\in\mathbb{Z}} \phi(\cdot - k) = 1$. Consequently, we conclude that $\widehat{\upsilon}(0)\phi(x-1+h_\phi)+\cdots+\widehat{\upsilon}(0)\phi(x+l_\phi) = 1$ for almost every $x \in [0,1]$. Similarly, we have $\widehat{\tilde{\upsilon}}(0)\sum_{k\in\mathbb{Z}} \tilde{\phi}(\cdot - k) = 1$ and $\widehat{\tilde{\upsilon}}(0)\tilde{\phi}(x-1+h_{\tilde{\phi}}) + \cdots + \widehat{\tilde{\upsilon}}(0)\tilde{\phi}(x+l_{\tilde{\phi}}) = 1$ for almost every $x \in [0,1]$. If (S2) is not performed, then $M_{\phi,\tilde{\phi}} = M$ and it is easy to directly check that $\vec{v} = [\widehat{\upsilon}(0), \ldots, \widehat{\upsilon}(0)]$, $\vec{\tilde{v}} = [\widehat{\tilde{\upsilon}}(0), \ldots, \widehat{\tilde{\upsilon}}(0)]$, and hence, (5.46) becomes

$$[\widehat{\tilde{\upsilon}}(0), \ldots, \widehat{\tilde{\upsilon}}(0)] M_{\tilde{\phi},\phi} [\overline{\widehat{\upsilon}(0)}, \ldots, \overline{\widehat{\upsilon}(0)}]^{\mathsf{T}} = 1. \tag{5.47}$$

A particular case of Theorem 5.10 is $\tilde{\phi} = \eta$ with $\eta(x) := (1, x, \ldots, x^m)^{\mathsf{T}}\chi_{[0,1]}$, which is a refinable vector function. This allows us to compute $\int_k^{k+1} x^j \phi(x)dx$ for all $j \in \mathbb{N}_0$ and $k \in \mathbb{Z}$. Note that η is refinable: $\eta(x) = C_0\eta(2x) + C_1\eta(2x-1)$ with $C_0 := \mathrm{diag}(1, 2^{-1}, \ldots, 2^{-m})$ and $C_1 := [2^{-m}\binom{j}{\ell}]_{0\leqslant j\leqslant m, 0\leqslant \ell\leqslant j}$.

5.3.3 Construct refinable $\tilde{\Phi}$ satisfying item (ii) of Theorem 5.7

Though elements in $\tilde{\phi}^L$ in Theorem 5.2 could be any compactly supported functions in $L_2([0,\infty))$, in this section we present an algorithm to construct a particular $\tilde{\Phi} = \{\tilde{\phi}^L\} \cup \{\tilde{\phi}(\cdot - k) \,:\, k \geqslant n_{\tilde{\phi}}\}$ satisfying item (ii) of Theorem 5.7. The general case $\tilde{\phi}^L \subseteq L_2([0,\infty))$

will be addressed in Theorem 5.15.

Algorithm 5.2. *Let $(\{\tilde\phi;\tilde\psi\},\{\phi;\psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a biorthogonal wavelet filter bank $(\{\tilde a;\tilde b\},\{a;b\})$ satisfying items (1)–(4) of Theorem 5.1. Let $0\leqslant m\leqslant \mathrm{sr}(a)$ and $0\leqslant \tilde m\leqslant \mathrm{sr}(\tilde a)$. Assume that $\Phi=\{\phi^L\}\cup\{\phi(\cdot-k)\ :\ k\geqslant n_\phi\}\subseteq L_2([0,\infty))$, with ϕ^L having compact support, satisfies item (i) of Theorem 5.7 (e.g., Φ is obtained by Section 5.3.1 or Theorem 5.13) and Φ is a Riesz sequence in $L_2([0,\infty))$. Define $[l_{\tilde\phi},h_{\tilde\phi}]:=\mathrm{fsupp}(\tilde\phi)$ and $[l_{\tilde a},h_{\tilde a}]:=\mathrm{fsupp}(\tilde a)$.*

(S1) Choose $n_{\tilde\phi}\geqslant\max(-l_{\tilde\phi},-l_{\tilde a},n_\phi)$ such that $n_{\tilde\phi}$ is the smallest integer satisfying $\langle\tilde\phi(\cdot-k),\phi^L\rangle=0$ for all $k\geqslant n_{\tilde\phi}$, e.g., we can take any $n_{\tilde\phi}\geqslant\max(-l_{\tilde\phi},-l_{\tilde a},n_\phi,h_{\phi^L})$ with $[l_{\phi^L},h_{\phi^L}]:=\mathrm{fsupp}(\phi^L)$.

(S2) Since $n_{\tilde\phi}\geqslant n_\phi$, we define a vector function $\mathring\phi^L:=\{\phi^L\}\cup\{\phi(\cdot-k)\ :\ n_\phi\leqslant k<n_{\tilde\phi}\}$.

(S3) Define a vector function $\breve\phi^L:=\tilde\phi^c\cup\tilde\phi^h$, where $\tilde\phi^c:=\{\tilde\phi(\cdot-k)\chi_{[0,\infty)}\}_{1-h_{\tilde\phi}\leqslant k\leqslant n_{\tilde\phi}-1}$ and

$$\tilde\phi^h:=2\sum_{k=n_{\tilde\phi}}^{n_h-1}\tilde A(k)\tilde\phi(2\cdot-k)\quad\text{and}\quad\langle\tilde\phi^h,\phi(\cdot-k)\rangle=0,\qquad\forall\,k\geqslant n_{\tilde\phi},\tag{5.48}$$

where $n_h:=2\max(n_{\phi^L},h_\phi+n_{\tilde\phi})-l_{\tilde\phi}$ with $\{\tilde A(k)\}_{k=n_{\tilde\phi}}^{n_h-1}$ to be determined. Note that $\breve\phi^L=2\breve A_L\breve\phi^L(2\cdot)+2\sum_{k=n_{\tilde\phi}}^{\infty}\breve A(k)\tilde\phi(2\cdot-k)$ for some matrix $\breve A_L$ and finitely supported sequence $\breve A$.

(S4) Apply item (BS) of Section 5.3.1 to break $\breve\phi^L$ into short vector functions $\breve\phi_1,\dots,\breve\phi_N$ with each satisfying (5.19). Initially define $\eta^L=\emptyset$. We add/merge $\breve\phi_\ell$ into η^L if $\langle\eta^L\cup\breve\phi_\ell,\mathring\phi^L\rangle$ has full rank. Repeat this procedure until $\#\eta^L=\#\mathring\phi^L$. Then $\tilde\phi^L:=\langle\eta^L,\mathring\phi^L\rangle^{-1}\eta^L$ is a well-defined vector function, because the square matrix $\langle\eta^L,\mathring\phi^L\rangle$ is invertible.

(S4') Assume that $\{\breve\phi^L\}\cup\{\tilde\phi(\cdot-k)\ :\ k\geqslant n_{\tilde\phi}\}$ is a Riesz sequence; otherwise, remove redundant elements from $\breve\phi^L$. Instead of (S4), we can alternatively obtain $\tilde\phi^L:=C\breve\phi^L$, where the unknown $(\#\mathring\phi)\times(\#\breve\phi^L)$ matrix C is determined by solving $C\langle\breve\phi^L,\mathring\phi^L\rangle=I_{\#\mathring\phi^L}$ and $C\breve A_L=C\breve A_L\langle\breve\phi^L,\mathring\phi^L\rangle C$.

Then $\tilde\Phi:=\{\tilde\phi^L\}\cup\{\tilde\phi(\cdot-k)\ :\ k\geqslant n_{\tilde\phi}\}$ satisfies item (ii) of Theorem 5.7.

Proof. By the choice of $n_{\tilde\phi}$ in (S1), for every $k\geqslant n_{\tilde\phi}$, we have $\langle\tilde\phi(\cdot-k),\eta\rangle=0$ for all $\eta\in\Phi\backslash\{\phi(\cdot-k)\}$. Note that the integer n_h in (S3) is chosen so that $\mathrm{fsupp}(\mathring\phi^L)$ is essentially disjoint with $\mathrm{supp}(\tilde\phi(2\cdot-k))$ for all $k\geqslant n_h$ and hence $\langle\tilde\phi(2\cdot-k),\mathring\phi^L\rangle=0$ for all $k\geqslant n_h$. By the definition of $\tilde\phi^c$, it is trivial to see that $\langle\tilde\phi^c,\phi(\cdot-k)\rangle=0$ for all $k\geqslant n_{\tilde\phi}$. This and (5.48)

imply that $\langle \tilde{\phi}^L, \phi(\cdot - k)\rangle = 0$ for all $k \geqslant n_{\tilde{\phi}}$. Now the claim holds trivially for the choice of $\tilde{\phi}^L$ in (S4).

The condition $C\langle \tilde{\phi}^c, \mathring{\phi}^L\rangle = I_{\#\mathring{\phi}^L}$ in (S4') is obviously equivalent to the biorthogonality condition $\langle \tilde{\phi}^L, \mathring{\phi}^L\rangle = I_{\#\mathring{\phi}^L}$. For the choice of $\tilde{\phi}^L$ in (S4'), we have

$$\tilde{\phi}^L = C\breve{\phi}^L = 2C\breve{A}_L\breve{\phi}^L(2\cdot) + 2\sum_{k=n_{\tilde{\phi}}}^{\infty} C\breve{A}(k)\tilde{\phi}(2\cdot - k).$$

Since $\{\breve{\phi}^L\} \cup \{\tilde{\phi}(\cdot - k) \ : \ k \geqslant n_{\tilde{\phi}}\}$ is a Riesz sequence, (5.19) holds if and only if $C\breve{A}_L = \tilde{A}_L C$. Due to the biorthogonality between $\tilde{\phi}^L$ and $\mathring{\phi}^L$, we must have $\tilde{A}_L = \langle \tilde{\phi}^L, \mathring{\phi}^L(2\cdot)\rangle = C\langle \breve{\phi}^L, \mathring{\phi}^L(2\cdot)\rangle = C\breve{A}_L\langle \breve{\phi}^L, \mathring{\phi}^L\rangle$. Now the condition $C\breve{A}_L = C\breve{A}_L\langle \breve{\phi}^L, \mathring{\phi}^L\rangle C$ in (S4') guarantees that $\tilde{\phi}^L$ satisfies (5.19) with $\tilde{A}_L = C\breve{A}_L\langle \breve{\phi}^L, \mathring{\phi}^L\rangle$. Hence, the claim holds for the choice of $\tilde{\phi}^L$ in (S4'). □

For spline biorthogonal scalar wavelets $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ in [29] with $\phi = B_m$ in (1.1) such that $\tilde{m} := \mathrm{sr}(\tilde{a}) \geqslant m$ and $m + \tilde{m}$ is an even integer, [34] considered the particular choice $\phi^L = \phi^{\mathsf{p}}$ and $\tilde{\phi}^L = \langle \tilde{\mathsf{p}}, \mathring{\phi}^L\rangle^{-1}\tilde{\phi}^{\tilde{\mathsf{p}}}$ in (S4) with $\mathsf{p}(x) = (1, x, \ldots, x^{m-1})^{\mathsf{T}}$ and $\tilde{\mathsf{p}}(x) = (1, x, \ldots, x^{\tilde{m}-1})^{\mathsf{T}}$ as in Proposition 5.4. [19, 114] instead took the particular choice $\phi^L = \phi^c$ in Proposition 5.4, which has more boundary elements than [34]. Then [114, (4.3)] and [19, (32)] proposed to take $\eta^L = \tilde{\phi}^{\tilde{\mathsf{p}}} \cup \tilde{\phi}^h$ in (S4) by properly choosing $\tilde{A}$ in (5.48).

5.3.4 Construct wavelets Ψ and $\tilde{\Psi}$ satisfying items (iii) and (iv) of Theorem 5.7

Assume that Φ and $\tilde{\Phi}$ satisfy items (i) and (ii) of Theorem 5.7 but without restricting that ϕ^L and $\tilde{\phi}^L$ are special choices as discussed in Sections 5.3.1 and 5.3.3. We now address how to construct Ψ and $\tilde{\Psi}$ satisfying items (iii) and (iv) of Theorem 5.7.

From Φ and $\tilde{\Phi}$ satisfying items (i) and (ii) of Theorem 5.7, we now construct Ψ satisfying item (iii) of Theorem 5.7 as follows.

Proposition 5.11. *Let $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ satisfying items (1)–(4) of Theorem 5.1. Suppose that Φ and $\tilde{\Phi}$ as defined in (5.2) and (5.15), consisting of compactly supported functions in $L_2([0,\infty))$, satisfy items (i) and (ii) of Theorem 5.7. Define $[l_{\phi^L}, h_{\phi^L}] := \mathrm{fsupp}(\phi^L)$ and $[l_{\tilde{\phi}^L}, h_{\tilde{\phi}^L}] := \mathrm{fsupp}(\tilde{\phi}^L)$. Take $n_\psi \in \mathbb{Z}$ to be the smallest integer satisfying*

$$n_\psi \geqslant \max(-l_\psi, \tfrac{n_\phi - l_b}{2}, h_{\tilde{\phi}^L} - l_\psi) \tag{5.49}$$

and define $m_\phi := \max(2n_\phi + h_{\tilde{a}}, 2n_\psi + h_{\tilde{b}})$ *as in* (5.39). *Let* $\mathring{\psi}^L := \{\phi^L(2\cdot)\} \cup \{\phi(2\cdot - k) \ : \ n_\phi \leqslant k < m_\phi\}$. *Calculate a compactly supported vector function* ψ^L *by*

$$\psi^L := \mathring{\psi}^L - \langle \mathring{\psi}^L, \tilde{\phi}^L \rangle \phi^L - \sum_{k=n_\phi}^{M_\phi - 1} \langle \mathring{\psi}^L, \tilde{\phi}(\cdot - k) \rangle \phi(\cdot - k), \tag{5.50}$$

where $M_\phi := \lceil \max(\frac{h_{\phi^L}}{2}, \frac{h_\phi + m_\phi - 1}{2}) \rceil - l_{\tilde{\phi}}$. *Then* $\Psi := \{\psi^L\} \cup \{\psi(\cdot - k) \ : \ k \geqslant n_\psi\}$ *satisfies item (iii) of Theorem* 5.7. *Moreover, if we apply item (3) of Theorem* 5.5 *with* Φ *being replaced by* Ψ *to remove the redundant elements of* ψ^L *in* $\{\psi^L\} \cup \{\psi(\cdot - k) \ : \ n_\psi \leqslant k < N_\psi\}$ *with* $N_\psi := \max(n_\psi, h_{\psi^L} - l_{\tilde{\psi}})$, *then* Ψ *with updated* ψ^L *satisfies item (iii) of Theorem* 5.7 *and* Ψ *is a Riesz sequence in* $L_2([0, \infty))$.

Proof. By the definition of n_ψ in (5.49), (5.14) holds. Note that $\operatorname{fsupp}(\psi(\cdot - k)) = [k + l_\psi, k + h_\psi]$ for $k \in \mathbb{Z}$. Hence, for $k \geqslant n_{\tilde{\psi}}$, we deduce from (5.49) that $n_\psi \geqslant h_{\tilde{\phi}^L} - l_\psi$. Therefore, due to essentially disjoint support, we trivially have $\langle \psi(\cdot - k), \tilde{\phi}^L \rangle = 0$ and $\psi(\cdot - k) \perp \tilde{\Phi}$ for all $k \geqslant n_{\tilde{\psi}}$.

Due to essentially disjoint support, we have $\langle \mathring{\psi}^L, \tilde{\phi}(\cdot - k) \rangle = 0$ for all $k \geqslant M_\phi$. Since $\tilde{\Phi}$ is biorthogonal to Φ by item (ii) of Theorem 5.7, we trivially deduce from the definition of ψ^L in (5.50) that $\psi^L \perp \tilde{\Phi}$. Therefore, Ψ is perpendicular to $\tilde{\Phi}$ and $\mathtt{S}_0(\Psi) \perp \mathtt{S}_0(\tilde{\Phi})$. By (5.41) in Lemma 5.8, we have $\mathtt{S}_1(\Phi) = \mathtt{S}_0(\Phi) + \mathtt{S}_0(\{\mathring{\psi}^L\} \cup \{\psi(\cdot - k) \ : \ k \geqslant n_\psi\})$ and hence $\mathtt{S}_1(\Phi) = \mathtt{S}_0(\Phi) + \mathtt{S}_0(\Psi)$. Therefore, we must have $\mathtt{S}_0(\Psi) = \mathtt{S}_1(\Phi) \cap (\mathtt{S}_0(\tilde{\Phi}))^\perp$. Because all functions in $\Phi \cup \tilde{\Phi}$ are compactly supported, ψ^L defined in (5.50) obviously has compact support. We can also directly check (5.18) using the definition of ψ^L in (5.50) and the relations in (5.17) and (5.13). Hence, Ψ satisfies item (iii) of Theorem 5.7. □

The integer n_ψ satisfying (5.49) can be replaced by the smallest $n_\psi \in \mathbb{Z}$ such that $\psi(\cdot - k) \in \mathtt{S}_1(\Phi)$ and $\psi(\cdot - k) \perp \tilde{\Phi}$ for all $k \geqslant n_\psi$. The choice of M_ϕ for defining ψ^L in (5.50) guarantees that $\operatorname{supp}(\mathring{\psi}^L) \cap \operatorname{fsupp}(\tilde{\phi}(\cdot - k))$ is at most a singleton and hence $\langle \mathring{\psi}^L, \tilde{\phi}(\cdot - k) \rangle = 0$ for all $k \geqslant M_\phi$. To guarantee that Ψ is a Riesz sequence in Proposition 5.11, we can avoid using Theorem 5.5 to reduce the redundant elements in ψ^L by replacing $\mathring{\psi}^L$ with a suitable subset of $\mathring{\psi}^L$, which forms a basis of the quotient space $\mathtt{S}_1(\Phi)/\mathtt{S}_0(\Phi \cup \{\psi(\cdot - k) \ : \ k \geqslant n_\psi\})$. See the remark after Theorem 5.14 about how to find such desired subset of $\mathring{\psi}^L$. Though ψ^L itself is not unique, the finite-dimensional space $\mathtt{S}_0(\Psi)/\mathtt{S}_0(\{\psi(\cdot - k) \ : \ k \geqslant n_\psi\})$ (or equivalently, $\operatorname{span}(\psi^L) \mod \operatorname{span}\{\psi(\cdot - k) \ : \ k \geqslant n_\psi\}$) is uniquely determined by Φ and $\tilde{\Phi}$ satisfying items (i) and (ii) of Theorem 5.7.

For $\Phi, \tilde{\Phi}$ and Ψ satisfying items (i)–(iii) of Theorem 5.7, it is easy to construct the dual wavelet $\tilde{\Psi}$ satisfying item (iv) of Theorem 5.7, mainly due to the uniqueness of $\tilde{\Psi}$.

Proposition 5.12. *Let $(\{\tilde{\phi};\tilde{\psi}\},\{\phi;\psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ satisfying items (1)–(4) of Theorem 5.1. Suppose that $\Phi,\tilde{\Phi}$ and Ψ as defined in (5.2) and (5.15), consisting of compactly supported functions in $L_2([0,\infty))$, satisfy items (i)–(iii) of Theorem 5.7. Assume that Ψ is a Riesz sequence in $L_2([0,\infty))$. Define*

$$[l_{\phi^L},h_{\phi^L}]:=\operatorname{fsupp}(\phi^L),\quad [l_{\psi^L},h_{\psi^L}]:=\operatorname{fsupp}(\psi^L),\quad [l_{\tilde{\phi}^L},h_{\tilde{\phi}^L}]:=\operatorname{fsupp}(\tilde{\phi}^L).$$

Take $n_{\tilde{\psi}}\in\mathbb{Z}$ to be the smallest integer satisfying

$$n_{\tilde{\psi}}\geqslant\max(-l_{\tilde{\psi}},\tfrac{n_{\tilde{\phi}}-l_{\tilde{b}}}{2},h_{\phi^L}-l_{\tilde{\psi}},h_{\psi^L}-l_{\tilde{\psi}},n_{\psi}) \tag{5.51}$$

and $m_{\tilde{\phi}}:=\max(2n_{\tilde{\phi}}+h_{\tilde{a}},2n_{\tilde{\psi}}+h_{\tilde{b}})$. For each element $\eta\in\{\psi^L\}\cup\{\psi(\cdot-k)\ :\ n_\psi\leqslant k<n_{\tilde{\psi}}\}$, there exists a unique sequence $\{c_\eta(h)\}_{h\in\Phi}\in\ell_2(\Phi)$ such that

$$\langle d_\eta,g\rangle=\begin{cases}1 & \text{if } g=\eta,\\ 0 & \text{if } g\in(\Phi\cup\Psi)\backslash\{\eta\}\end{cases}\qquad \text{with}\qquad d_\eta:=\sqrt{2}\sum_{h\in\Phi}c_\eta(h)\tilde{h}(2\cdot)\in\mathtt{S}_1(\tilde{\Phi}) \tag{5.52}$$

and

$$c_\eta(\phi(\cdot-k))=0\qquad\forall\ k\geqslant m_{\tilde{\phi}}. \tag{5.53}$$

Then $\tilde{\Psi}:=\{\tilde{\psi}^L\}\cup\{\tilde{\psi}(\cdot-k)\ :\ k\geqslant n_{\tilde{\psi}}\}$ with $\tilde{\psi}^L:=\{d_\eta\ :\ \eta\in\{\psi^L\}\cup\{\psi(\cdot-k)\ :\ n_\psi\leqslant k<n_{\tilde{\psi}}\}\}$ satisfies item (iv) of Theorem 5.7 and all functions in $\tilde{\psi}^L$ have compact support.

Proof. Note that all $\Phi,\tilde{\Phi}$ and Ψ are Riesz sequences. We first prove that $\Phi\cup\Psi$ is a Riesz sequence. Suppose not. Then the lower Riesz bound of $\Phi\cup\Psi$ is zero and there exists a sequence $\{c_n\}_{n=1}^\infty$ in $\ell_2(\Phi\cup\Psi)$ such that $\sum_{h\in\Phi\cup\Psi}|c_n(h)|^2=1$ and $\lim_{n\to\infty}\|F_n\|_{L_2(\mathbb{R})}=0$, where $F_n:=f_n+g_n$ with $f_n:=\sum_{h\in\Phi}c_n(h)h$ and $g_n:=\sum_{h\in\Psi}c_n(h)h$. Because $\tilde{\Phi}$ is biorthogonal to Φ and $\tilde{\Phi}\perp\Psi$, for $\tilde{h}\in\tilde{\Phi}$, we have $\langle F_n,\tilde{h}\rangle=\langle f_n,\tilde{h}\rangle=c_n(h)$, where h is the corresponding element of $\tilde{h}$ in Φ. Then

$$\lim_{n\to\infty}\sum_{h\in\Phi}|c_n(h)|^2=\lim_{n\to\infty}\sum_{h\in\Phi}|\langle F_n,\tilde{h}\rangle|^2=\lim_{n\to\infty}\sum_{\tilde{h}\in\tilde{\Phi}}|\langle F_n,\tilde{h}\rangle|^2=0,$$

because $\tilde{\Phi}$ is a Riesz sequence and $\lim_{n\to\infty}\|F_n\|_{L_2(\mathbb{R})}=0$. Then $\lim_{n\to\infty}\|f_n\|_{L_2(\mathbb{R})}=0$ since Φ is a Riesz sequence. For any $\varepsilon>0$, there exists $N\in\mathbb{N}$ such that $\sum_{h\in\Phi}|c_n(h)|^2\leqslant\varepsilon$, $\|F_n\|_{L_2(\mathbb{R})}\leqslant\varepsilon$, and $\|f_n\|_{L_2(\mathbb{R})}\leqslant\varepsilon$ for all $n\geqslant N$. Thus,

$$\sum_{h\in\Psi}|c_n(h)|^2=1-\sum_{h\in\Phi}|c_n(h)|^2\geqslant 1-\varepsilon$$

and $\|g_n\|_{L_2(\mathbb{R})} = \|F_n - f_n\|_{L_2(\mathbb{R})} \leqslant 2\varepsilon$. This shows that the lower Riesz bound of Ψ cannot be larger than $\frac{2\varepsilon}{\sqrt{1-\varepsilon}}$ for all $0 < \varepsilon < 1$, which contradicts that Ψ is a Riesz sequence. Hence, $\Phi \cup \Psi$ must be a Riesz sequence.

Because $\Phi \cup \Psi$ is a Riesz sequence and $\mathtt{S}_0(\Phi \cup \Psi) = \mathtt{S}_1(\Phi)$ by item (iii) of Theorem 5.7, $\Phi \cup \Psi$ is a Riesz basis of $\mathtt{S}_1(\Phi)$. Since $\Phi(2\cdot)$ is also a Riesz basis of $\mathtt{S}_1(\Phi)$, for every $g \in \Phi \cup \Psi$, we define $w_g \in \ell_2(\Phi)$ to be the unique sequence satisfying $g = \sum_{h\in\Phi} w_g(h)h(2\cdot)$. Let $\eta \in \Psi$ and define W_η to be the closed linear span of $w_g, g \in \Phi \cup \Psi \backslash \{\eta\}$. Then there exists a unique element $v_\eta \in \ell_2(\Phi)$ such that $w_\eta - v_\eta \in W_\eta$ and $v_\eta \perp W_\eta$. Because $\Phi \cup \Psi$ is a Riesz basis of $\mathtt{S}_1(\Phi)$, we must have $v_\eta \neq 0$ and $\langle v_\eta, w_\eta\rangle_{\ell_2(\Phi)} \neq 0$. Define $f := \sum_{h\in\Phi} v_\eta(h)\tilde{h}(2\cdot)$. Then $f \in \mathtt{S}_1(\tilde{\Phi})$ and $f \neq 0$ by $v_\eta \neq 0$. Because $\tilde{\Phi}$ is biorthogonal to Φ, we must have $\langle f, \eta\rangle \neq 0$ by $\langle v_\eta, w_\eta\rangle_{\ell_2(\Phi)} \neq 0$ and $f \perp (\Phi \cup \Psi)\backslash\{\eta\}$ by $v_\eta \perp W_\eta$. Obviously, $d_\eta := \langle f, \eta\rangle^{-1} f \in \mathtt{S}_1(\tilde{\Phi})$ must satisfy (5.52). Suppose that both $d_\eta, d'_\eta \in \mathtt{S}_1(\tilde{\Phi})$ satisfy (5.52). Then $d_\eta - d'_\eta \perp \Phi \cup \Psi$. Hence, $d_\eta - d'_\eta \perp \mathtt{S}_0(\Phi \cup \Psi) = \mathtt{S}_1(\Phi)$. Since $\tilde{\Phi}$ is biorthogonal to Φ and $d_\eta - d'_\eta \in \mathtt{S}_1(\tilde{\Phi})$, we must have $d_\eta = d'_\eta$. This proves the existence and uniqueness of d_η.

We now prove that $d_{\psi(\cdot-m)} = \tilde{\psi}(\cdot - m)$ for all $m \geqslant n_{\tilde{\psi}}$. Since $n_{\tilde{\psi}} \geqslant \max(-l_{\tilde{\psi}}, \frac{n_{\tilde{\phi}}-l_{\tilde{b}}}{2})$, we observe that (5.22) holds. Note that $\mathrm{fsupp}(\tilde{\psi}(\cdot - k)) = [k + l_{\tilde{\psi}}, k + h_{\tilde{\psi}}]$ for $k \in \mathbb{Z}$. Since $n_{\tilde{\psi}} \geqslant h_{\phi^L} - l_{\tilde{\psi}}$ by (5.51), for $m \geqslant n_{\tilde{\psi}}$ we have $m + l_{\tilde{\psi}} \geqslant h_{\phi^L}$ and hence $\langle \phi^L, \tilde{\psi}(\cdot - m)\rangle = 0$. Consequently, $\tilde{\psi}(\cdot - m)$ is perpendicular to all elements in Φ. By the same argument and $n_{\tilde{\psi}} \geqslant h_{\psi^L} - l_{\tilde{\psi}}$, $\tilde{\psi}(\cdot - m)$ is also perpendicular to all elements in $\{\psi^L\} \cup \{\psi(\cdot - k) \ : \ n_\psi \leqslant k < n_{\tilde{\psi}}\}$. Now we conclude that (5.52) must hold with $\eta = \psi(\cdot - m)$ and $d_\eta = \tilde{\psi}(\cdot - m)$. This proves $d_{\psi(\cdot-m)} = \tilde{\psi}(\cdot - m)$ for all $m \geqslant n_{\tilde{\psi}}$.

Let $\eta \in \{\psi^L\} \cup \{\psi(\cdot - k) \ : \ n_\psi \leqslant k < n_{\tilde{\psi}}\}$. We now prove that (5.53) must hold. By Lemma 5.8 with n_ψ being replaced with $n_{\tilde{\psi}}$, we conclude from (5.41) that

$$\phi(2\cdot - k_0) \text{ is a finite linear combination of elements in } H \text{ for all } k_0 \geqslant m_{\tilde{\phi}}, \tag{5.54}$$

where $H := \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\} \cup \{\psi(\cdot - k) \ : \ k \geqslant n_{\tilde{\psi}}\}$. By the definition of d_η in (5.52) and $\eta \in \{\psi^L\} \cup \{\psi(\cdot - k) \ : \ n_\psi \leqslant k < n_{\tilde{\psi}}\}$, we have $d_\eta \perp H$. For any integer $k \geqslant m_{\tilde{\phi}}$, by the biorthogonality of Φ and $\tilde{\Phi}$, we deduce from (5.54) and $d_\eta \perp H$ that $c_\eta(\phi(\cdot - k)) = \langle d_\eta, \sqrt{2}\phi(2\cdot - k)\rangle = 0$. This proves (5.53). Hence, $\tilde{\psi}^L$ has compact support and $\tilde{\Psi}$ satisfies item (iv) of Theorem 5.7. □

5.4 Direct approach for constructing biorthogonal wavelets on $[0,\infty)$

The general construction using the classical approach in Section 5.3 (in particular, Section 5.3.3) is often complicated and it restricts the choices of ϕ^L and $\tilde{\phi}^L$ in Sections 5.3.1 and 5.3.3. In this section, we propose a direct approach to construct all possible compactly supported biorthogonal wavelets on $[0,\infty)$ without explicitly involving the dual refinable functions $\tilde{\Phi}$ and dual wavelets $\tilde{\Psi}$.

We first address how to construct general $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\}$ with $\phi^L \subseteq L_2([0,\infty))$ satisfying item (i) of Theorem 5.7 and having compact support.

Theorem 5.13. *Let ϕ be a compactly supported refinable vector function satisfying $\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot - k)$ for some finitely supported matrix-valued filter $a \in (l_0(\mathbb{Z}))^{r\times r}$. Take $n_\phi \in \mathbb{Z}$ satisfying (5.13). Let A_L be an $N \times N$ matrix satisfying*

$$\rho(A_L) < 2^{-1/2}, \quad \textit{that is, the spectral radius of } A_L \textit{ is less than } 2^{-1/2}. \tag{5.55}$$

Define an $N \times 1$ vector function ϕ^L by

$$\phi^L := \sum_{j=1}^{\infty} 2^{j-1} A_L^{j-1} g(2^j \cdot) \quad \textit{with} \quad g := 2\sum_{k=n_\phi}^{\infty} A(k)\phi(\cdot - k), \tag{5.56}$$

where A is a finitely supported sequence of $N \times (\#\phi)$ matrices. Then ϕ^L is a well-defined compactly supported vector function in $L_2([0,\infty)) \cap H^\tau(\mathbb{R})$ for some $\tau > 0$, (5.17) holds, i.e., $\phi^L = 2A_L\phi^L(2\cdot) + 2\sum_{k=n_\phi}^{\infty} A(k)\phi(2\cdot - k)$, and $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\}$ satisfies item (i) of Theorem 5.7.

Proof. Since ϕ is a compactly supported refinable vector function with a finitely supported filter $a \in (l_0(\mathbb{Z}))^{r\times r}$, by [71, Corollary 5.8.2 or Corollary 6.3.4] (also see [67, Theorem 2.2]), the refinable vector function ϕ must belong to $H^t(\mathbb{R})$ for some $t > 0$. By our assumption in (5.55), we have $\log_2 \rho(A_L) < -1/2$ and hence, $t_{A_L} := -1/2 - \log_2 \rho(A_L) > 0$. So, $(0,t] \cap (0,t_{A_L})$ is nonempty. Let $\tau \in (0,t] \cap (0,t_{A_L})$. Then $\tau > 0$. Since $\phi \subseteq H^\tau(\mathbb{R})$ by $0 < \tau \leqslant t$ and n_ϕ satisfies (5.13), we see that g must be a compactly supported vector function in $L_2([0,\infty)) \cap H^\tau(\mathbb{R})$ and hence there exists a positive constant C independent of j (e.g., see [69, (3.7)]) such that

$$C^{-1}\|g\|_{H^\tau(\mathbb{R})} \leqslant 2^{(-1/2-\tau)j}\|2^j g(2^j\cdot)\|_{H^\tau(\mathbb{R})} \leqslant C\|g\|_{H^\tau(\mathbb{R})}$$

for all $j \in \mathbb{N} \cup \{0\}$. Applying the triangle inequality, we deduce that

$$\left\| \sum_{j=1}^{\infty} 2^{j-1} A_L^{j-1} g(2^j \cdot) \right\|_{H^\tau(\mathbb{R})} \leqslant \sum_{j=1}^{\infty} 2^{j-1} \|A_L^{j-1}\| \|g(2^j \cdot)\|_{H^\tau(\mathbb{R})} \leqslant 2^{-1} C \|g\|_{H^\tau(\mathbb{R})} \sum_{j=1}^{\infty} 2^{(1/2+\tau)j} \|A_L^{j-1}\|.$$

Since $\tau < t_{A_L}$, we have $t_{A_L} - \tau > 0$. For any $0 < \varepsilon < t_{A_L} - \tau$, since $\rho(A_L) = \lim_{j\to\infty} \|A_L^j\|^{1/j}$, there exists a positive constant C_ε such that $\|A_L^j\| \leqslant C_\varepsilon 2^{\varepsilon j} (\rho(A_L))^j$ for all $j \in \mathbb{N} \cup \{0\}$. Therefore,

$$\sum_{j=1}^{\infty} 2^{(1/2+\tau)j} \|A_L^{j-1}\| \leqslant C_\varepsilon \sum_{j=1}^{\infty} 2^{(1/2+\tau)j} 2^{\varepsilon(j-1)} (\rho(A_L))^{j-1} = 2^{(1/2+\tau)} C_\varepsilon \sum_{j=0}^{\infty} 2^{(\tau+\varepsilon-t_{A_L})j} < \infty,$$

because $\tau + \varepsilon - t_{A_L} < 0$ by $0 < \varepsilon < t_{A_L} - \tau$. Hence we proved $\phi^L \subseteq H^\tau(\mathbb{R})$ with $\tau > 0$ and thus, $\phi^L \subseteq L_2([0,\infty))$. Since g has compact support, ϕ^L in (5.56) obviously has compact support and

$$\phi^L = g(2\cdot) + 2A_L \sum_{j=1}^{\infty} 2^{j-1} A_L^{j-1} g(2^{j+1}\cdot) = g(2\cdot) + 2A_L \phi^L(2\cdot) = 2A_L \phi^L(2\cdot) + 2 \sum_{k=n_\phi}^{\infty} A(k) \phi(2\cdot - k).$$

This proves (5.17). Hence, $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\}$ satisfies item (i) of Theorem 5.7. □

To achieve polynomial reproduction, we can simply append the vector function ϕ^{p} in item (ii) of Proposition 5.4 to the vector function ϕ^L in Theorem 5.13 to create a new ϕ^L; or equivalently, we can append the associated refinement coefficients of ϕ^{p} to the coefficients A_L and A in (5.56) instead. Then remove redundant elements in the new ϕ^L by Theorem 5.5. To construct orthogonal wavelets on $[0,\infty)$, we can simply replace (S1) of Algorithm 5.1 by Theorem 5.13. As we discussed in Section 5.3.1, without loss of generality, A_L in (5.56) can be taken as a block diagonal matrix of Jordan matrices in (5.38). Define a vector function $\mathring{\phi}^L$ by appending ϕ to ϕ^L. Since ϕ^L satisfies (5.17), we see that $\mathring{\phi}^L$ is a compactly supported vector function associated with a finitely supported filter. Consequently, we can apply Theorem 5.10 to compute $\int_0^1 \phi^L(x-m)\overline{\eta(x-n)}^{\mathsf{T}} dx$ for all $m, n \in \mathbb{Z}$ for any compactly supported refinable vector function η. Generalizing results on vector subdivision schemes and refinable vector functions in [71, Chapter 5], in fact we can prove through a technical argument that the condition in (5.55) must hold if $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\}$ satisfies item (i) of Theorem 5.7 and is a Riesz sequence. We shall address this technical issue elsewhere.

The direct approach is to construct Ψ directly. The following result will be proved in

Section 5.8.

Theorem 5.14. *Let $(\{\tilde{\phi};\tilde{\psi}\},\{\phi;\psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a biorthogonal wavelet filter bank $(\{\tilde{a};\tilde{b}\},\{a;b\})$ satisfying items (1)–(4) of Theorem 5.1. Suppose that $\Phi=\{\phi^L\}\cup\{\phi(\cdot-k)\ :\ k\geqslant n_\phi\}\subseteq L_2([0,\infty))$ satisfies item (i) of Theorem 5.7, Φ is a Riesz sequence in $L_2([0,\infty))$, and ϕ^L is a compactly supported vector function in $L_2([0,\infty))\cap H^\tau(\mathbb{R})$ for some $\tau>0$. Take an integer $n_\psi\geqslant\max(-l_\psi,\frac{n_\phi-l_b}{2})$ and define $m_\phi:=\max(2n_\phi+h_{\tilde{a}},2n_\psi+h_{\tilde{b}})$ as in (5.39). Let $m,n_0\in\mathbb{N}_0$ (we often take $0\leqslant m\leqslant \mathrm{sr}(\tilde{a})$ and $n_0=0$). Construct $\Psi:=\{\psi^L\}\cup\{\psi(\cdot-k)\ :\ k\geqslant n_\psi\}$ such that*

(i) $\psi^L\subseteq \mathit{span}(\{\phi^L(2\cdot)\}\cup\{\phi(2\cdot-k)\ :\ n_\phi\leqslant k<m_\phi+n_0\})$, ψ^L has m vanishing moments with $\mathrm{vm}(\psi^L)\geqslant m$, and the set ψ^L (which is regarded as in $\mathtt{S}_1(\Phi)/\mathtt{S}_0(\Phi\cup\{\psi(\cdot-k)\ :\ k\geqslant n_\psi\})$) is a basis of the finite-dimensional quotient space $\mathtt{S}_1(\Phi)/\mathtt{S}_0(\Phi\cup\{\psi(\cdot-k)\ :\ k\geqslant n_\psi\})$.

(ii) Every element in $\phi^L(2\cdot)\cup\phi^E(2\cdot)$ is a finite linear combination of elements in $\Phi\cup\Psi$, where $\phi^E:=\{\phi(\cdot-k)\ :\ n_\phi\leqslant k<m_\phi\}$. That is, for some integers $h_C\geqslant n_\phi$ and $h_D\geqslant n_\psi$,

$$\begin{bmatrix}\phi^L(2\cdot)\\ \phi^E(2\cdot)\end{bmatrix}=A_0\phi^L+B_0\psi^L+\sum_{n_\phi\leqslant k<h_C}C(k)\phi(\cdot-k)+\sum_{n_\psi\leqslant k<h_D}D(k)\psi(\cdot-k)\qquad(5.57)$$

for some matrices A_0,B_0, $C(k)$ with $n_\phi\leqslant k<h_C$, and $D(k)$ with $n_\psi\leqslant k<h_D$.

Define $n_{\tilde{\phi}}:=\max(m_\phi,h_C,-l_{\tilde{\phi}},1-l_{\tilde{a}})$ and $n_{\tilde{\psi}}:=\max(n_\psi,h_D,-l_{\tilde{\psi}},\lceil\frac{n_{\tilde{\phi}}-l_{\tilde{b}}+1}{2}\rceil)$. Then we must have

$$\phi(2\cdot-k_0)=\sum_{k=n_\phi}^{n_{\tilde{\phi}}-1}\overline{\tilde{a}(k_0-2k)}^{\mathsf{T}}\phi(\cdot-k)+\sum_{k=n_{\tilde{\psi}}}^{n_{\tilde{\psi}}-1}\overline{\tilde{b}(k_0-2k)}^{\mathsf{T}}\psi(\cdot-k),\qquad\forall\, m_\phi\leqslant k_0<n_{\tilde{\phi}}.\quad(5.58)$$

Now we can rewrite/combine (5.57) and (5.58) together into the following equivalent form:

$$\mathring{\phi}^L(2\cdot)=\overline{\tilde{A}_L}^{\mathsf{T}}\mathring{\phi}^L+\overline{\tilde{B}_L}^{\mathsf{T}}\mathring{\psi}^L,\qquad(5.59)$$

where $\mathring{\phi}^L:=\{\phi^L\}\cup\{\phi(\cdot-k)\ :\ n_\phi\leqslant k<n_{\tilde{\phi}}\}$, $\mathring{\psi}^L:=\{\psi^L\}\cup\{\psi(\cdot-k)\ :\ n_\psi\leqslant k<n_{\tilde{\psi}}\}$, and the matrices $\tilde{A}_L,\tilde{B}_L$ are uniquely determined by $A_0,B_0,\{C(k)\}_{k=n_\phi}^{h_C-1}$, $\{D(k)\}_{k=n_\psi}^{h_D-1}$ and the filters $\tilde{a},\tilde{b}$. If

$$\rho(\tilde{A}_L)<2^{-1/2},\quad\textit{that is, the spectral radius of }\tilde{A}_L\textit{ is less than }2^{-1/2},\qquad(5.60)$$

then the following $\tilde{\phi}^L$ *and* $\tilde{\psi}^L$ *are well-defined compactly supported vector functions in* $L_2([0,\infty))$*:*

$$\tilde{\phi}^L := \sum_{j=1}^{\infty} 2^{j-1}\tilde{A}_L^{j-1}\tilde{g}(2^j\cdot) \quad \textit{with} \quad \tilde{g} := 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{A}(k)\tilde{\phi}(\cdot-k), \tag{5.61}$$

$$\tilde{\psi}^L := 2\tilde{B}_L\tilde{\phi}^L(2\cdot) + 2\sum_{k=n_{\tilde{\phi}}}^{\infty}\tilde{B}(k)\tilde{\phi}(2\cdot-k), \tag{5.62}$$

where the $(\#\mathring{\phi}^L)\times(\#\phi)$ *matrices* $\tilde{A}(k)$ *and* $(\#\mathring{\psi}^L)\times(\#\phi)$ *matrices* $\tilde{B}(k), k\in\mathbb{Z}$ *are defined by*

$$\tilde{A}(k) := \begin{bmatrix} 0_{(\#\phi^L)\times(\#\phi)} \\ \tilde{a}(k-2n_\phi) \\ \vdots \\ \tilde{a}(k-2(n_{\tilde{\phi}}-1)) \end{bmatrix} \quad \textit{and} \quad \tilde{B}(k) := \begin{bmatrix} 0_{(\#\psi^L)\times(\#\phi)} \\ \tilde{b}(k-2n_\psi) \\ \vdots \\ \tilde{b}(k-2(n_{\tilde{\psi}}-1)) \end{bmatrix}, \quad k\in\mathbb{Z}. \tag{5.63}$$

Then $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ *and* $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ *form a pair of biorthogonal Riesz bases of* $L_2([0,\infty))$ *for every* $J\in\mathbb{Z}$*, where* $\tilde{\Phi} := \{\tilde{\phi}^L\}\cup\{\tilde{\phi}(\cdot-k) \ : \ k\geqslant n_{\tilde{\phi}}\}$ *and* $\tilde{\Psi} := \{\tilde{\psi}^L\}\cup\{\tilde{\psi}(\cdot-k) \ : \ k\geqslant n_{\tilde{\psi}}\}$.

By item (3) of Theorem 5.2, items (i) and (ii) of Theorem 5.14 are necessary conditions on Ψ for $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ to be a Riesz basis of $L_2([0,\infty))$ satisfying both (5.17) and (5.18). We often take $n_\psi\in\mathbb{Z}$ to be the smallest integer such that $\psi(\cdot-k)\in\mathtt{S}_1(\Phi)$ for all $k\geqslant n_\psi$. We now discuss how to construct all possible

$$\psi^L \subseteq \operatorname{span}(\{\phi^L(2\cdot)\}\cup\{\phi(2\cdot-k) \ : \ n_\phi\leqslant k<m_\phi+n_0\})$$

satisfying both items (i) and (ii) of Theorem 5.14. Since $\Phi(2\cdot)$ is a Riesz basis of $\mathtt{S}_1(\Phi)$, we observe that

$$\eta = \sum_{h\in\Phi} c_\eta(h)h(2\cdot) \quad \text{with} \quad c_\eta := \{c_\eta(h)\}_{h\in\Phi}\in\ell_2(\Phi), \qquad \text{for } \eta\in\mathtt{S}_1(\Phi).$$

Let $S := \phi^L\cup\phi^E$ and $T := \{\phi(\cdot-k) \ : \ k\geqslant m_\phi\}$. Then $\Phi = S\cup T$ and we can write $c_\eta = c_\eta\chi_S + c_\eta\chi_T$ for all $\eta\in\mathtt{S}_1(\Phi)$. Define $H := \Phi\cup\{\psi(\cdot-k) \ : \ k\geqslant n_\psi\}$. By (5.5) and (5.17), we have $\mathtt{S}_0(H)\subseteq\mathtt{S}_1(\Phi)$. Now we can find a finite subset $H_0\subseteq H$ such that $M_0 := \{c_\eta\chi_S \ : \ \eta\in H_0\}$ is a basis of the finite-dimensional space spanned by $c_\eta\chi_S, \eta\in H$. In other words, $H_0\cup T(2\cdot)$ is another Riesz basis of $\mathtt{S}_1(\Phi)$. Next we find a generating set

(not necessarily a basis) ψ^b of the finite-dimensional space

$$W := \{g \in \operatorname{span}(\{\phi^L(2\cdot)\} \cup \{\phi(2\cdot - k) \ : \ n_\phi \leqslant k < m_\phi + n_0\}) \ : \ \operatorname{vm}(g) \geqslant m\},$$

which is not empty by taking n_0 large enough (we often set $n_0 = 0$). If $\psi(\cdot - k) \in W$ for some $k \in \mathbb{Z}$, to have as many interior wavelets as possible, then we always keep $\psi(\cdot - k)$ in ψ^b. We now find a subset $\psi^L \subseteq \psi^b$ such that $M_0 \cup M_1$ is a basis of $\mathbb{R}^{\#S}$, where $M_1 := \{c_\eta \chi_S \ : \ \eta \in \psi^L\}$. Since both $\Phi(2\cdot)$ and $H_0 \cup T(2\cdot)$ are Riesz bases of $\mathtt{S}_1(\Phi)$, item (i) must hold for the constructed ψ^L. We now prove that ψ^L satisfies item (ii) of Theorem 5.14. Let $\eta \in S$. Then $c_{\eta(2\cdot)} = \boldsymbol{\delta}_\eta \in \mathbb{R}^{\#S}$. Since $M_0 \cup M_1$ is a basis of $\mathbb{R}^{\#S}$, by $M_0 \cup M_1 = \{c_h \chi_S \ : \ h \in \psi^L \cup H_0\}$, we have $c_{\eta(2\cdot)} = \sum_{h \in \psi^L \cup H_0} d_{\eta,h} c_h \chi_S$ for some $d_{\eta,h} \in \mathbb{C}$. Observing $c_h = c_h \chi_S + c_h \chi_T$, we obtain

$$c_{\eta(2\cdot)} - \sum_{h \in \psi^L \cup H_0} d_{\eta,h} c_h = - \sum_{h \in \psi^L \cup H_0} d_{\eta,h} c_h \chi_T.$$

Using the refinable structure in (5.5) and (5.17), we conclude that every sequence c_h for $h \in \psi^L \cup H$ must have only finitely many nonzero entries. Hence, by $H_0 \subseteq H$, the sequence $\sum_{h \in \psi^L \cup H_0} d_{\eta,h} c_h \chi_T$ has only finitely many nonzero entries. Therefore, we conclude from (5.41) that $\sum_{h \in \psi^L \cup H_0} d_{\eta,h} c_h \chi_T$ must be a finite linear combination of elements in H. Consequently, $\eta(2\cdot) - \sum_{h \in \psi^L \cup H_0} d_{\eta,h} h$ must be a finite linear combination of elements in H. This proves item (ii) of Theorem 5.14.

Instead of constructing Ψ first in Theorem 5.14, we can first construct $\tilde{\Phi}$ satisfying item (ii) of Theorem 5.7 below. Then construct Ψ by Proposition 5.11 and $\tilde{\Psi}$ by Proposition 5.12. The classical approach in Section 5.3 can be improved by the following result, whose proof is given in Section 5.8.

Theorem 5.15. *Let $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a biorthogonal wavelet filter bank $(\{\tilde{a}; \tilde{b}\}, \{a; b\})$ satisfying items (1)–(4) of Theorem 5.1. Suppose that $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\} \subseteq L_2([0, \infty))$ with ϕ^L having compact support satisfies item (i) of Theorem 5.7 and Φ is a Riesz sequence in $L_2([0, \infty))$. Let $n_{\tilde{\phi}}$ be chosen as in item (S1) of Algorithm 5.2. Define $N := \#\phi^L + (n_{\tilde{\phi}} - n_\phi)(\#\phi)$ and let $\tilde{A}_L$ be an $N \times N$ matrix satisfying (5.60). For a finitely supported sequence $\tilde{A}$ of $N \times (\#\phi)$ matrices, define $\tilde{\phi}^L$ as in (5.61). By Theorem 5.13, $\tilde{\phi}^L$ is a well-defined compactly supported vector function in $L_2([0, \infty)) \cap H^\tau(\mathbb{R})$ for some $\tau > 0$. If*

$$\tilde{A}_L \overline{A_L}^{\mathsf{T}} + \sum_{k=n_{\tilde{\phi}}}^{\infty} \tilde{A}(k) \overline{A(k)}^{\mathsf{T}} = I_N, \tag{5.64}$$

where A_L *and* $\{A(k)\}_{k=n_{\tilde{\phi}}}^{\infty}$ *are augmented version in* (5.17) *with* ϕ^L *being replaced by* $\mathring{\phi}^L := \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ n_\phi \leqslant k < n_{\tilde{\phi}}\}$, *then* $\tilde{\Phi}$ *is biorthogonal to* Φ *and satisfies item (ii) of Theorem* 5.7, *where* $\tilde{\Phi} := \{\tilde{\phi}^L\} \cup \{\tilde{\phi}(\cdot - k) \ : \ k \geqslant n_{\tilde{\phi}}\}$.

If $\tilde{\Phi}$ is biorthogonal to Φ, then (5.64) must hold. Hence, (5.64) is a necessary condition for the biorthogonality between $\tilde{\Phi}$ and Φ. Theorem 5.15 generalizes Algorithm 5.2 for the classical approach.

5.5 Biorthogonal wavelets on $[0, \infty)$ satisfying homogeneous boundary conditions

In this section we study (bi)orthogonal wavelets on $[0, \infty)$ satisfying given boundary conditions.

For a polynomial $\mathsf{p}(x) = \sum_{j=0}^{\infty} c_j x^j$, it is convenient to use the notation $\mathsf{p}(\frac{d}{dx}) = \sum_{j=0}^{\infty} c_j \frac{d^j}{dx^j}$ for a differential operator. Let $\mathcal{I} := [0, \infty)$. To study wavelets on $\mathcal{I}$ with general homogeneous boundary conditions such as Robin boundary conditions, it is necessary for us to study nonstationary wavelet systems in $L_2(\mathcal{I})$. For subsets Φ_j and Ψ_j of functions in $L_2(\mathcal{I})$ with $j \in \mathbb{Z}$, we define

$$\mathsf{AS}_J(\Phi_J; \{\Psi_j\}_{j=J}^{\infty})_{\mathcal{I}} := \{2^{J/2}\varphi(2^J\cdot) \ : \ \varphi \in \Phi_J\} \cup \{2^{j/2}\eta(2^j\cdot) \ : \ \eta \in \Psi_j, j \geqslant J\}, \qquad J \in \mathbb{Z}. \tag{5.65}$$

If $\Phi_j = \Phi$ and $\Psi_j = \Psi$ for all $j \in \mathbb{Z}$, then $\mathsf{AS}_J(\Phi_J; \{\Psi_j\}_{j=J}^{\infty})_{[0,\infty)} = \mathsf{AS}_J(\Phi; \Psi)_{[0,\infty)}$ as in (5.1).

For wavelets $\Psi_j(2^j\cdot), j \geqslant J$ satisfying prescribed boundary conditions, the following result shows that all the elements in $\mathsf{AS}_J(\Phi_J; \{\Psi_j\}_{j=J}^{\infty})$ must satisfy the same prescribed boundary conditions.

Theorem 5.16. *Let* $J \in \mathbb{Z}$ *and* $\mathcal{I} = [0, \infty)$. *Let* $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ *be a compactly supported biorthogonal wavelet in* $L_2(\mathbb{R})$ *satisfying items (1)–(4) of Theorem* 5.1. *Let* $n_\phi \in \mathbb{Z}$ *such that* $\operatorname{fsupp}(\eta(\cdot - k_0)) \subseteq \mathcal{I}$ *for all* $k_0 \geqslant n_\phi$ *and* $\eta \in \phi \cup \psi \cup \tilde{\phi} \cup \tilde{\psi}$. *Let* $\{\phi_J^L, \tilde{\phi}_J^L\} \cup \{\psi_j^L, \tilde{\psi}_j^L\}_{j=J}^{\infty} \subseteq L_2(\mathcal{I})$ *have compact support and satisfy* $\lim_{j\to\infty} 2^{-j} h_{\tilde{\psi}_j^L} = 0$, *where* $[l_{\tilde{\psi}_j^L}, h_{\tilde{\psi}_j^L}] := \operatorname{fsupp}(\tilde{\psi}_j^L)$. *Define* Φ_J *and* $\Psi_j, j \geqslant J$ *by*

$$\Phi_J := \{\phi_J^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\}, \qquad \Psi_j := \{\psi_j^L\} \cup \{\psi(\cdot - k) \ : \ k \geqslant n_\phi\} \tag{5.66}$$

and

$$\tilde{\Phi}_J := \{\tilde{\phi}_J^L\} \cup \{\tilde{\phi}(\cdot - k) \ : \ k \geqslant n_\phi\}, \qquad \tilde{\Psi}_j := \{\tilde{\psi}_j^L\} \cup \{\tilde{\psi}(\cdot - k) \ : \ k \geqslant n_\phi\}.$$

Suppose that $\mathsf{AS}_J(\Phi_J;\{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ *and* $\mathsf{AS}_J(\tilde{\Phi}_J;\{\tilde{\Psi}_j\}_{j=J}^\infty)_{\mathcal{I}}$ *form a pair of biorthogonal Riesz bases in* $L_2(\mathcal{I})$. *Let* $\mathsf{p}_0,\dots,\mathsf{p}_\ell \in \mathbb{P}_{m-1}$ *be polynomials of degree less than* m. *Suppose that each function* $\eta \in \mathsf{AS}_J(\Phi_J;\{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ *has continuous derivatives of all orders less than* m *on* $[0,\varepsilon_\eta)$ *for some* $\varepsilon_\eta > 0$. *If all the wavelet functions in* $\mathsf{AS}_J(\Phi_J;\{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ *satisfy the following homogeneous boundary conditions prescribed by* $\mathsf{p}_0,\dots,\mathsf{p}_\ell$ *as follows:*

$$\mathsf{p}_0(\tfrac{d}{dx})(\eta(2^j x))|_{x=0} = \cdots = \mathsf{p}_\ell(\tfrac{d}{dx})(\eta(2^j x))|_{x=0} = 0 \quad \forall\, \eta \in \Psi_j, j \geqslant J, \tag{5.67}$$

then the refinable functions $\Phi_J(2^J\cdot)$ *in* $\mathsf{AS}_J(\Phi_J;\{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ *must satisfy the same boundary conditions*

$$\mathsf{p}_0(\tfrac{d}{dx})(\varphi(2^J x))|_{x=0} = \cdots = \mathsf{p}_\ell(\tfrac{d}{dx})(\varphi(2^J x))|_{x=0} = 0 \quad \forall\, \varphi \in \Phi_J. \tag{5.68}$$

That is, if all the elements in $\{2^{j/2}\eta(2^j\cdot) \,:\, \eta\in\Psi_j, j\geqslant J\}$ *satisfy the homogeneous boundary conditions in* (5.67), *then all elements in* $\mathsf{AS}_J(\Phi_J;\{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ *must satisfy the same boundary conditions.*

Proof. Let $\varepsilon > 0$ and consider functions $f \in L_2([2\varepsilon,\infty))$. Since

$$\mathsf{AS}_J(\Phi_J;\{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}} \quad \text{and} \quad \mathsf{AS}_J(\tilde{\Phi}_J;\{\tilde{\Psi}_j\}_{j=J}^\infty)_{\mathcal{I}}$$

form a pair of biorthogonal Riesz bases of $L_2(\mathcal{I})$, we have

$$f(x) = \sum_{\varphi\in\Phi} 2^J\langle f,\tilde{\varphi}(2^J\cdot)\rangle\varphi(2^J x) + \sum_{j=J}^\infty\sum_{\eta\in\Psi_j} 2^j\langle f,\tilde{\eta}(2^j\cdot)\rangle\eta(2^j x).$$

Because $\lim_{j\to\infty} 2^{-j}h_{\tilde{\psi}_j^L} = 0$ and $\operatorname{supp}(\tilde{\psi}_j^L(2^j\cdot)) \subseteq [0,2^{-j}h_{\tilde{\psi}_j^L}]$, there exists $\tilde{J}_\varepsilon \in \mathbb{N}$ such that

$$\operatorname{supp}(\tilde{\psi}_j^L(2^j\cdot)) \subseteq [0,2\varepsilon], \qquad \forall\, j \geqslant \tilde{J}_\varepsilon. \tag{5.69}$$

Since $\phi,\psi,\tilde{\phi}$ and $\tilde{\psi}$ have compact support, we assume that all of them are supported inside $[-N,N]$ for some $N\in\mathbb{N}$. Since $\operatorname{supp}(\eta(2^j\cdot-k)) \subseteq [2^{-j}(k-N),2^{-j}(k+N)]$ for $\eta \in \phi\cup\psi\cup\tilde{\phi}\cup\tilde{\psi}$, we observe

$$\operatorname{supp}(\tilde{\psi}(2^j\cdot-k)) \subseteq [0,2\varepsilon], \qquad \forall\, n_\phi \leqslant k \leqslant 2^{j+1}\varepsilon - N \tag{5.70}$$

and

$$\operatorname{supp}(\phi(2^j\cdot-k)) \cup \operatorname{supp}(\psi(2^j\cdot-k)) \subseteq [\varepsilon,\infty), \qquad \forall\, k \geqslant 2^j\varepsilon + N. \tag{5.71}$$

Let $J_\varepsilon \in \mathbb{N}$ such that $J_\varepsilon \geqslant \max(\tilde{J}_\varepsilon, \log_2\frac{2N}{\varepsilon})$. For $j \geqslant J_\varepsilon$ and $k\in\mathbb{Z}$, then either $k \leqslant 2^{j+1}\varepsilon - N$ or $k \geqslant 2^j\varepsilon + N$ must hold. Consequently, one of (5.70) and (5.71) must hold for all $k \geqslant n_\phi$.

Hence, by $\operatorname{supp}(f) \subseteq [2\varepsilon, \infty)$ and $J_\varepsilon \geqslant \tilde{J}_\varepsilon$, we deduce from (5.70) and (5.71) that

$$\langle f, \tilde{\eta}(2^j \cdot)\rangle \eta(2^j x) = 0 \qquad \forall\, x \in [0, \varepsilon), \eta \in \Psi_j, j \geqslant J_\varepsilon. \tag{5.72}$$

From (5.71), for $x \in [0, \varepsilon)$, we have $\langle f, \tilde{\phi}(2^J \cdot - k)\rangle \phi(2^J x - k) = 0$ for all $k \geqslant 2^J \varepsilon + N$ and $\langle f, \tilde{\psi}(2^j \cdot - k)\rangle \psi(2^j x - k) = 0$ for all $k \geqslant 2^j \varepsilon + N$. Consequently, by (5.71) and (5.72), we obtain

$$\begin{aligned} f(x) =& \langle f, (\tilde{\phi}_J^L)_{J;0}\rangle (\phi_J^L)_{J;0}(x) + \sum_{k=n_\phi}^{\lfloor 2^j \varepsilon + N\rfloor} \langle f, \tilde{\phi}_{J;k}\rangle \phi_{J;k}(x) \\ &+ \sum_{j=J}^{J_\varepsilon - 1} \Big(\langle f, (\tilde{\psi}_j^L)_{j;0}\rangle (\psi_j^L)_{j;0}(x) + \sum_{k=n_\phi}^{\lfloor 2^j \varepsilon + N\rfloor} \langle f, \tilde{\psi}_{j;k}\rangle \psi_{j;k}(x) \Big) \end{aligned} \tag{5.73}$$

for almost every $x \in [0, \varepsilon)$, where $\psi_{j;k} := 2^{j/2} \psi(2^j \cdot - k)$. By assumption, each function $\eta \in \mathsf{AS}_J(\Phi_J; \{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ is continuous on $[0, \varepsilon_\eta)$ and has continuous derivatives of all orders less than m on $[0, \varepsilon_\eta)$ for some $\varepsilon_\eta > 0$. Because there are only finitely many terms in (5.73), there must exist $0 < \varepsilon_0 < \varepsilon$ such that (5.73) holds for all $x \in [0, \varepsilon_0)$ and all terms in (5.73) have continuous derivatives of all orders less than m on $[0, \varepsilon_0)$. Applying our assumption in (5.67) and using the fact $\operatorname{supp}(f) \subset [2\varepsilon, \infty)$, we conclude from (5.73) that $\mathsf{p}_i(\frac{d}{dx}) f(x)|_{x=0} = 0$ and

$$2^J \langle f, \tilde{\phi}_J^L(2^J \cdot)\rangle \mathsf{p}_i(\tfrac{d}{dx}) \phi_J^L(2^J x)|_{x=0} + \sum_{k=n_\phi}^{\lfloor 2^j \varepsilon + N\rfloor} 2^J \langle f, \tilde{\phi}(2^J \cdot - k)\rangle \mathsf{p}_i(\tfrac{d}{dx}) \phi(2^J x - k)|_{x=0} = 0 \tag{5.74}$$

for all $i = 0, \ldots, \ell$. By the choice of n_ϕ satisfying (5.13), $\operatorname{supp}(\phi(\cdot - k)) \subseteq [1, \infty)$ for all $k \geqslant n_\phi + 1$ and hence trivially $\mathsf{p}_i(\frac{d}{dx}) \phi(2^J x - k)|_{x=0} = 0$. For simplicity, we may group $\phi(\cdot - n_\phi)$ into ϕ^L. Hence, (5.74) becomes

$$\langle f, \tilde{\phi}_J^L(2^J \cdot)\rangle \mathsf{p}_i(\tfrac{d}{dx}) \phi_J^L(2^J x)|_{x=0} = 0, \qquad \forall\, i = 0, \ldots, \ell \quad \text{and} \quad f \in L_2([2\varepsilon, \infty)). \tag{5.75}$$

In particular, (5.75) must hold with $\varepsilon = 0$. Since $\tilde{\phi}_J^L$ must be a Riesz sequence, the mapping $L_2([0, \infty)) \to \mathbb{C}^{\#\tilde{\phi}_J^L}$ with $f \mapsto \langle f, \tilde{\phi}_J^L(2^J \cdot)\rangle$ is onto. Consequently, we deduce from (5.75) that (5.68) holds for all $\varphi \in \phi_J^L$, from which we conclude that (5.68) holds for all $\varphi \in \Phi_J$. □

As a direct consequence of Theorem 5.16, we now claim that any orthogonal wavelet basis on $[0, \infty)$ satisfying boundary conditions often cannot have high vanishing moments.

Theorem 5.17. *Let $J \in \mathbb{Z}$ and $\mathcal{I} = [0, \infty)$. Let $\{\phi; \psi\}$ be a compactly supported or-*

thogonal wavelet in $L_2(\mathbb{R})$. *Let* $\{\phi_J^L\} \cup \{\psi_j^L\}_{j=J}^\infty \subseteq L_2(\mathcal{I})$ *have compact support and satisfy* $\lim_{j\to\infty} 2^{-j} h_{\psi_j^L} = 0$, *where* $[l_{\psi_j^L}, h_{\psi_j^L}] := \mathrm{fsupp}(\psi_j^L)$. *Let* $n_\phi \in \mathbb{Z}$ *such that* $\mathrm{fsupp}(\eta(\cdot - k_0)) \subseteq \mathcal{I}$ *for all* $k_0 \geqslant n_\phi$ *and* $\eta \in \phi \cup \psi$. *Define* Φ_J *and* $\Psi_j, j \geqslant J$ *as in* (5.66). *Let* $\mathsf{p}_0, \ldots, \mathsf{p}_\ell \in \mathbb{P}_{m-1}$ *and define* n_{B} *to be the largest nonnegative integer such that*

$$\mathsf{p}_i(\tfrac{d}{dx})(x^j)|_{x=0} = 0 \qquad \forall\, i = 0, \ldots, \ell \quad \text{and} \quad j = 0, \ldots, n_{\mathrm{B}} - 1. \tag{5.76}$$

Suppose that each function $\eta \in \mathsf{AS}_J(\Phi_J; \{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ *has continuous derivatives of all orders less than* m *on* $[0, \varepsilon_\eta)$ *for some* $\varepsilon_\eta > 0$. *If* $\mathsf{AS}_J(\Phi_J; \{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ *is an orthonormal basis of* $L_2(\mathcal{I})$ *such that the boundary conditions in* (5.67) *are satisfied, then for any* $j_0 \in \mathbb{N}$, *there exists at least one element* $\eta_{j_0} \in \cup_{j=j_0}^\infty \Psi_j$ *such that* η_{j_0} *has no more than* n_{B} *vanishing moments, i.e.,* $\mathrm{vm}(\eta_{j_0}) \leqslant n_B$.

Proof. Suppose not. Then there exists $j_0 \in \mathbb{N}$ such that all the elements in $\cup_{j=j_0}^\infty \Psi_j$ have $n_{\mathrm{B}} + 1$ vanishing moments. Since ϕ and ψ have compact support, we can assume that ϕ and ψ are supported inside $[-N, N]$ for some $N \in \mathbb{N}$. Define $\varepsilon := 2^{1-j_0} N > 0$ and let $h : [0, \infty) \to \mathbb{R}$ be a compactly supported continuous function such that $h(x) = 1$ for all $x \in [0, 2\varepsilon]$. Define $f(x) := x^{n_{\mathrm{B}}} h(x)$. Then $f \in L_2(\mathcal{I})$ has compact support and $f(x) = x^{n_{\mathrm{B}}}$ for all $x \in [0, 2\varepsilon]$. Noting that $\mathrm{supp}(\eta(2^j \cdot - k)) \subseteq [2^{-j}(k - N), 2^{-j}(k + N)]$ for all $\eta \in \phi \cup \psi$, we can easily verify that (5.71) holds and

$$\mathrm{supp}(\psi(2^j \cdot - k)) \subset [0, 2\varepsilon], \qquad \forall\, n_\phi \leqslant k \leqslant 2^{j+1}\varepsilon - N. \tag{5.77}$$

By $\lim_{j\to\infty} 2^{-j} h_{\psi_j^L} = 0$, there exists an integer $\tilde{J}_\varepsilon \geqslant j_0$ such that $\mathrm{supp}(\psi_j^L(2^j \cdot)) \subseteq [0, 2\varepsilon]$ for all $j \geqslant \tilde{J}_\varepsilon$. Since $f(x) = x^{n_{\mathrm{B}}}$ for $x \in [0, 2\varepsilon]$ and all elements in $\cup_{j=j_0}^\infty \Psi_j$ have $n_{\mathrm{B}} + 1$ vanishing moments, we have $\langle f, \psi_j^L(2^j \cdot)\rangle = 0$ for all $j \geqslant \tilde{J}_\varepsilon$. Let $J_\varepsilon \in \mathbb{N}$ such that $J_\varepsilon \geqslant \max(\tilde{J}_\varepsilon, \log_2 \frac{2N}{\varepsilon})$. Note that $J_\varepsilon \geqslant j_0$ by $\varepsilon = 2^{1-j_0} N$. For all $j \geqslant J_\varepsilon$, one of (5.71) and (5.77) must hold. Now it follows from the same argument as in the proof of Theorem 5.16 that

$$\begin{aligned} f(x) = &\langle f, (\phi_J^L)_{J;0}\rangle (\phi_J^L)_{J;0}(x) + \sum_{k=n_\phi}^{\lfloor 2^j\varepsilon + N\rfloor} \langle f, \phi_{J;k}\rangle \phi_{J;k}(x) \\ &+ \sum_{j=J}^{J_\varepsilon - 1} \Big(\langle f, (\psi_j^L)_{j;0}\rangle (\psi_j^L)_{j;0}(x) + \sum_{k=n_\phi}^{\lfloor 2^j\varepsilon + N\rfloor} \langle f, \psi_{j;k}\rangle \psi_{j;k}(x) \Big) \end{aligned} \tag{5.78}$$

for almost every $x \in [0, \varepsilon)$, and there exists $0 < \varepsilon_0 < \varepsilon$ such that (5.78) holds for all $x \in [0, \varepsilon_0)$ and all terms in (5.78) have continuous derivatives of all orders less than m on $[0, \varepsilon_0)$.

On the other hand, all the conditions in Theorem 5.16 are satisfied. Consequently,

all the elements in $\mathsf{AS}_J(\Phi_J;\{\Psi_j\}_{j=J}^\infty)_{\mathcal{I}}$ must satisfy the prescribed homogeneous boundary conditions. Therefore, we deduce from (5.78) that $\mathsf{p}_i(\frac{d}{dx})f(x)|_{x=0}=0$ for all $i=0,\dots,\ell$. Because $f(x)=x^{n_{\rm B}}$ for all $x\in[0,2\varepsilon]$, we conclude that (5.76) holds with $n_{\rm B}$ being replaced by $n_{\rm B}+1$, which contradicts the definition of the maximum integer $n_{\rm B}$ in (5.76). This proves the claim. □

A popular choice of homogeneous boundary conditions in the literature is

$$\mathsf{p}_0(\tfrac{d}{dx})=\frac{d^{j_0}}{dx^{j_0}},\quad\dots,\quad \mathsf{p}_\ell(\tfrac{d}{dx})=\frac{d^{j_\ell}}{dx^{j_\ell}}\quad\text{with}\quad 0\leqslant j_0<\dots<j_\ell<m. \tag{5.79}$$

Moreover, the particular choice $j_0=0,\dots,j_\ell=\ell$ in (5.79) is commonly used in the variational formulation of the boundary value problems in numerical partial differential equations, where the derivatives are in the weak/distributional sense and boundary values at 0 are interpreted in the trace sense. Spline scalar wavelets on $[0,1]$ satisfying homogeneous Dirichlet boundary conditions have been addressed in [18, 20, 35, 39, 71, 73, 88] and references therein.

Let $(\mathsf{AS}_J(\tilde\Phi;\tilde\Psi)_{[0,\infty)},\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)})$ be a (stationary) biorthogonal wavelet on $[0,\infty)$. Let $\mathsf{p}_0,\dots,\mathsf{p}_\ell\in\mathbb{P}_{m-1}$. It is easy to check that (5.67) holds with $\Psi_j:=\Psi$ for all $j\geqslant J$ if and only if $\mathsf{p}(\frac{d}{dx})\eta(x)|_{x=0}=0$ for all $\eta\in\Psi$ and $\mathsf{p}\in\mathcal{P}:=\operatorname{span}\{\mathsf{p}_0(2^j\cdot),\dots,\mathsf{p}_\ell(2^j\cdot)\ :\ j\geqslant J\}$. Note that $\mathcal{P}$ is generated by all the nonzero monomial terms in the polynomials $\mathsf{p}_0,\dots,\mathsf{p}_\ell$. Hence, if $\operatorname{span}\{\mathsf{p}_0,\dots,\mathsf{p}_\ell\}$ does not have a basis of monomials as in (5.79), then the dimension of $\mathcal{P}$ will be greater than $\ell+1$. To avoid increasing the number of boundary conditions, it is necessary to consider nonstationary wavelets in (5.65). To construct a biorthogonal wavelet $(\mathsf{AS}_J(\tilde\Phi;\tilde\Psi)_{[0,\infty)},\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)})$ on $[0,\infty)$ such that

$$\frac{d^{j_0}}{dx^{j_0}}\eta(x)|_{x=0}=\cdots=\frac{d^{j_\ell}}{dx^{j_\ell}}\eta(x)|_{x=0}=0,\qquad\forall\,\eta\in\Psi,$$

by Theorem 5.16 and the refinable structure in (5.17) and (5.18), it is necessary and sufficient that

$$\frac{d^{j_0}}{dx^{j_0}}\varphi(x)|_{x=0}=\cdots=\frac{d^{j_\ell}}{dx^{j_\ell}}\varphi(x)|_{x=0}=0,\qquad\forall\,\varphi\in\Phi. \tag{5.80}$$

Consequently, (5.80) holds if and only if all the elements in $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ satisfies the same prescribed homogeneous boundary conditions given by (5.79). For Φ satisfying the boundary conditions in (5.80), to achieve high approximation orders near the endpoint 0, it is important to have

$$\mathbb{P}_{m-1}\chi_{[0,\infty)}\subseteq\operatorname{span}\{\mathsf{p}_0\chi_{[0,\infty)},\dots,\mathsf{p}_\ell\chi_{[0,\infty)}\}+\mathsf{S}_0(\Phi)\quad\text{with}\quad\mathsf{p}_0,\dots,\mathsf{p}_\ell\text{ in (5.79)}. \tag{5.81}$$

For any Φ satisfying item (i) of Theorem 5.7, we can easily obtain a new Φ^{bc} satisfying item

(i) of Theorem 5.7 and the boundary conditions in (5.80).

Proposition 5.18. *Let* $\Phi = \{\phi^L\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi\} \subseteq L_2([0,\infty))$ *satisfy item (i) of Theorem 5.7, where* ϕ^L *and* ϕ *have compact support. Let* $\mathsf{p}(x) := (x^{j_0}, \dots, x^{j_\ell})^{\mathsf{T}}$ *with* $\{j_0, \dots, j_\ell\} \subseteq \{0, 1, \dots, m-1\}$. *Suppose that every element* $\eta \in \Phi$ *has continuous derivatives of all order less than* m *on* $[0, \varepsilon_\eta)$ *for some* $\varepsilon_\eta > 0$. *Then there exists an invertible* $(\#\phi^L + \#\phi) \times (\#\phi^L + \#\phi)$ *matrix* C_{ϕ^L} *such that*

$$M_{\mathsf{p}}(\phi^{L,I}) = \{0\} \text{ and } M_{\mathsf{p}}(\phi^{L,E}) \text{ is a basis of } \operatorname{span}(M_{\mathsf{p}}(\Phi)) \text{ with } \begin{bmatrix} \phi^{L,E} \\ \phi^{L,I} \end{bmatrix} := C_{\phi^L} \begin{bmatrix} \phi^L \\ \phi(\cdot - n_\phi) \end{bmatrix}, \tag{5.82}$$

where $M_{\mathsf{p}}(S) := \{\mathsf{p}(\frac{d}{dx})\eta(x)|_{x=0} \ : \ \eta \in S\}$ *for* $S \subseteq \mathtt{S}_0(\Phi)$. *Then*

$$\Phi^{bc} := \{\phi^{L,I}\} \cup \{\phi(\cdot - k) \ : \ k \geqslant n_\phi + 1\}$$

satisfies item (i) of Theorem 5.7, the homogeneous boundary conditions $\mathsf{p}(\frac{d}{dx})\eta(x)|_{x=0} = 0$ *for all* $\eta \in \Phi^{bc}$, *and* $\mathtt{S}_0(\Phi^{bc}) = \{\eta \in \mathtt{S}_0(\Phi) \ : \ \mathsf{p}(\frac{d}{dx})\eta(x)|_{x=0} = 0\}$.

Proof. Note that $\operatorname{fsupp}(\phi(\cdot - k)) \subseteq [1, \infty)$ for all $k \geqslant n_\phi + 1$. Trivially, $\mathsf{p}(\frac{d}{dx})\phi(x-k)|_{x=0} = 0$ for all $k \geqslant n_\phi + 1$. Thus, $M_{\mathsf{p}}(\Phi) = M_{\mathsf{p}}(\phi^L \cup \phi(\cdot - n_\phi))$, which is a finite subset of $\mathbb{R}^{\#\mathsf{p}}$. Therefore, there exists an invertible matrix C_{ϕ^L} such that (5.82) holds. Hence, all the elements in Φ^{bc} satisfy the homogeneous boundary conditions prescribed by p. Since $\mathtt{S}_0(\phi^{L,E} \cup \Phi^{bc}) = \mathtt{S}_0(\Phi)$ and C_{ϕ^L} is invertible, by (5.13) and (5.17), we have

$$\eta = \sum_{f \in \phi^{L,E}} c_\eta(f) f(2\cdot) + \sum_{f \in \phi^{L,I}} c_\eta(f) f(2\cdot) + \sum_{k=n_\phi+1}^{\infty} c_\eta(\phi(\cdot - k))\phi(2\cdot - k), \qquad \eta \in \mathtt{S}_0(\Phi). \tag{5.83}$$

Because $M_{\mathsf{p}}(g) = \{0\}$ for all $g \in \Phi^{bc}$, for $\eta \in \mathtt{S}_0(\Phi)$ with $M_{\mathsf{p}}(\eta) = \{0\}$ in (5.83), we have

$$\{0\} = M_{\mathsf{p}}(\eta) = [c_\eta(f)]_{f \in \phi^{L,E}} M_{\mathsf{p}}(\phi^{L,E}(2\cdot)) = [c_\eta(f)]_{f \in \phi^{L,E}} \operatorname{diag}(2^{j_0}, \dots, 2^{j_\ell}) M_{\mathsf{p}}(\phi^{L,E}).$$

Since $M_{\mathsf{p}}(\phi^{L,E})$ is linearly independent, the above identity forces $c_\eta(f) = 0$ for all $f \in \phi^{L,E}$. By (5.83), $\eta = \sum_{f \in \phi^{L,I}} c_\eta(f) f(2\cdot) + \sum_{k=n_\phi+1}^{\infty} c_\eta(\phi(\cdot - k))\phi(2\cdot - k)$ for all $\eta \in \Phi^{bc}$. This proves (5.17) and (5.13) for Φ^{bc} with n_ϕ being replaced by $n_\phi + 1$. Hence, Φ^{bc} satisfies item (i) of Theorem 5.7. The identity $\mathtt{S}_0(\Phi^{bc}) = \{\eta \in \mathtt{S}_0(\Phi) \ : \ \mathsf{p}(\frac{d}{dx})\eta(x)|_{x=0} = 0\}$ follows directly from (5.82) and (5.83). □

5.6 Orthogonal and biorthogonal wavelets on bounded intervals

In this section we discuss how to construct locally supported biorthogonal wavelets on a bounded interval $[0, N]$ with $N \in \mathbb{N}$ from compactly supported biorthogonal wavelets on $[0, \infty)$.

Recall that $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and $f_{j;k} := 2^{j/2} f(2^j \cdot - k)$ for $j, k \in \mathbb{Z}$. We remind the reader that a vector function is also used as an ordered set and vice versa throughout the paper. Using the classical approach in Section 5.3 or the direct approach in Section 5.4 for constructing (bi)orthogonal wavelets on $[0, \infty)$, we now discuss how to construct a locally supported (bi)orthogonal wavelet in $L_2([0, N])$ with $N \in \mathbb{N}$ from a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$. We shall provide a detailed proof in Section 5.8 for the following result, which is often employed but without a proof in the literature.

Theorem 5.19. *Let $(\{\tilde\phi; \tilde\psi\}, \{\phi; \psi\})$ be a compactly supported biorthogonal wavelet in $L_2(\mathbb{R})$ with a biorthogonal wavelet filter bank $(\{\tilde a; \tilde b\}, \{a; b\})$ satisfying items (1)–(4) of Theorem 5.1. A locally supported biorthogonal wavelet on the interval $[0, N]$ with $N \in \mathbb{N}$ can be constructed as follows:*

(S1) From the biorthogonal wavelet $(\{\tilde\phi; \tilde\psi\}, \{\phi; \psi\})$ in $L_2(\mathbb{R})$, use either the classical approach in Section 5.3 or the direct approach in Section 5.4 to construct compactly supported $\Phi, \Psi, \tilde\Phi, \tilde\Psi$ as in (5.2) and (5.15) such that $(\mathsf{AS}_J(\tilde\Phi; \tilde\Psi)_{[0,\infty)}, \mathsf{AS}_J(\Phi; \Psi)_{[0,\infty)})$ is a pair of biorthogonal Riesz bases in $L_2([0, \infty))$ for every $J \in \mathbb{N}_0$ and satisfies items (i)–(iv) of Theorem 5.7.

(S2) Similarly, perform item (S1) to the (flipped) biorthogonal wavelet $(\{\mathring{\tilde\phi}; \mathring{\tilde\psi}\}, \{\mathring\phi; \mathring\psi\})$ in $L_2(\mathbb{R})$ to construct compactly supported $\mathring\Phi, \mathring\Psi, \mathring{\tilde\Phi}, \mathring{\tilde\Psi}$, where

$$\mathring\phi := \phi(-\cdot), \quad \mathring\psi := \psi(-\cdot), \quad \mathring{\tilde\phi} := \tilde\phi(-\cdot), \quad \mathring{\tilde\psi} := \tilde\psi(-\cdot), \tag{5.84}$$

such that $(\mathsf{AS}_J(\mathring{\tilde\Phi}; \mathring{\tilde\Psi})_{[0,\infty)}, \mathsf{AS}_J(\mathring\Phi; \mathring\Psi)_{[0,\infty)})$ is a pair of biorthogonal Riesz bases in $L_2([0, \infty))$ for every $J \in \mathbb{N}_0$ and satisfies items (i)–(iv) of Theorem 5.7 similarly.

(S3) Let J_0 be the smallest nonnegative integer such that

$$\max(h_A + n_{\mathring\phi}, h_B + n_{\mathring\phi}, h_{\mathring A} + n_\phi, h_{\mathring B} + n_\phi, 2n_\phi + 2n_{\mathring\phi} - 2, 2n_\psi + 2n_{\mathring\psi} - 2) \leqslant 2^{J_0+1} N \tag{5.85}$$

and all elements in $\phi^L \cup \psi^L \cup \mathring{\phi}^L \cup \mathring{\psi}^L$ *are supported inside* $[0, 2^{J_0}N]$*, where*

$$\begin{aligned} [l_A, h_A] &:= \operatorname{fsupp}(A), \quad [l_B, h_B] := \operatorname{fsupp}(B), \\ [l_{\tilde{A}}, h_{\tilde{A}}] &:= \operatorname{fsupp}(\tilde{A}), \quad [l_{\tilde{B}}, h_{\tilde{B}}] := \operatorname{fsupp}(\tilde{B}) \end{aligned} \tag{5.86}$$

for the finitely supported filters $A, B, \tilde{A}, \tilde{B}$ *in* (5.17)*,* (5.18)*,* (5.19)*, and* (5.20)*, respectively. The integers* $h_{\mathring{A}}, h_{\mathring{B}}$ *and* $h_{\mathring{\tilde{A}}}, h_{\mathring{\tilde{B}}}$ *are defined similarly.*

(S4) Let $\tilde{J}_0$ *be the smallest nonnegative integer such that*

$$\max(h_{\tilde{A}}+n_{\mathring{\tilde{\phi}}}, h_{\tilde{B}}+n_{\mathring{\tilde{\phi}}}, h_{\tilde{A}}+n_{\tilde{\phi}}, h_{\tilde{B}}+n_{\tilde{\phi}}, 2n_{\tilde{\phi}}+2n_{\mathring{\tilde{\phi}}}-2, 2n_{\tilde{\psi}}+2n_{\mathring{\tilde{\psi}}}-2) \leqslant 2^{\tilde{J}_0+1}N, \tag{5.87}$$

all elements in $\tilde{\phi}^L \cup \tilde{\psi}^L \cup \mathring{\tilde{\phi}}^L \cup \mathring{\tilde{\psi}}^L$ *are supported inside* $[0, 2^{\tilde{J}_0}N]$*, and for all* $j \geqslant \tilde{J}_0$*,*

$$\{\phi^L_{j;0}, \psi^L_{j;0}\} \perp \{\tilde{\phi}^R_{j;2^jN-N}, \tilde{\psi}^R_{j;2^jN-N}\} \quad \text{and} \quad \{\tilde{\phi}^L_{j;0}, \tilde{\psi}^L_{j;0}\} \perp \{\phi^R_{j;2^jN-N}, \psi^R_{j;2^jN-N}\}, \tag{5.88}$$

where the right boundary refinable functions and right boundary wavelets are defined by

$$\phi^R := \mathring{\phi}^L(N-\cdot), \quad \psi^R := \mathring{\psi}^L(N-\cdot), \quad \tilde{\phi}^R := \mathring{\tilde{\phi}}^L(N-\cdot), \quad \tilde{\psi}^R := \mathring{\tilde{\psi}}^L(N-\cdot). \tag{5.89}$$

Without loss of generality, we assume $J_0 \leqslant \tilde{J}_0$*. Then the following statements hold:*

(1) for each $j \geqslant J_0$*, there exist matrices* A_j *and* B_j *such that* $\Phi_j = A_j\Phi_{j+1}$ *and* $\Psi_j = B_j\Phi_{j+1}$ *with* $\#\Psi_j = \#\Phi_{j+1} - \#\Phi_j = 2^jN(\#\phi)$ *and* $\#\Phi_j = \#\phi^L + \#\mathring{\phi}^L + (2^jN - n_{\mathring{\phi}} - n_\phi + 1)(\#\phi)$*, where*

$$\Phi_j := \{\phi^L_{j;0}\} \cup \{\phi_{j;k} \ : \ n_\phi \leqslant k \leqslant 2^jN - n_{\mathring{\phi}}\} \cup \{\phi^R_{j;2^jN-N}\}, \tag{5.90}$$

$$\Psi_j := \{\psi^L_{j;0}\} \cup \{\psi_{j;k} \ : \ n_\psi \leqslant k \leqslant 2^jN - n_{\mathring{\psi}}\} \cup \{\psi^R_{j;2^jN-N}\}. \tag{5.91}$$

(2) for every $j \geqslant \tilde{J}_0$*, there exist matrices* $\tilde{A}_j$ *and* $\tilde{B}_j$ *such that* $\tilde{\Phi}_j = \tilde{A}_j\tilde{\Phi}_{j+1}$ *and* $\tilde{\Psi}_j = \tilde{B}_j\tilde{\Phi}_{j+1}$ *hold with* $\#\tilde{\Psi}_j = \#\Psi_j = 2^jN(\#\phi)$ *and* $\#\tilde{\Phi}_j = \#\Phi_j$*, where*

$$\tilde{\Phi}_j := \{\tilde{\phi}^L_{j;0}\} \cup \{\tilde{\phi}_{j;k} \ : \ n_{\tilde{\phi}} \leqslant k \leqslant 2^jN - n_{\mathring{\tilde{\phi}}}\} \cup \{\tilde{\phi}^R_{j;2^jN-N}\}, \tag{5.92}$$

$$\tilde{\Psi}_j := \{\tilde{\psi}^L_{j;0}\} \cup \{\tilde{\psi}_{j;k} \ : \ n_{\tilde{\psi}} \leqslant k \leqslant 2^jN - n_{\mathring{\tilde{\psi}}}\} \cup \{\tilde{\psi}^R_{j;2^jN-N}\}. \tag{5.93}$$

(3) For every $J \geqslant \tilde{J}_0$*,* $(\tilde{\mathcal{B}}_J, \mathcal{B}_J)$ *forms a pair of biorthogonal Riesz bases of* $L_2([0, N])$ *and for all integers* $j \geqslant \tilde{J}_0$*, the matrix* $[\overline{A_j}^{\mathsf{T}}, \overline{B_j}^{\mathsf{T}}]$ *must be an invertible square matrix*

satisfying

$$\begin{bmatrix}\tilde{A}_j\\ \tilde{B}_j\end{bmatrix} = [\overline{A_j}^{\mathsf{T}}, \overline{B_j}^{\mathsf{T}}]^{-1}, \quad \textit{that is,} \quad \begin{bmatrix}\tilde{A}_j\\ \tilde{B}_j\end{bmatrix} [\overline{A_j}^{\mathsf{T}}, \overline{B_j}^{\mathsf{T}}] = \begin{bmatrix}A_j\\ B_j\end{bmatrix} [\overline{\tilde{A}_j}^{\mathsf{T}}, \overline{\tilde{B}_j}^{\mathsf{T}}] = I_{\#\Phi_{j+1}}, \tag{5.94}$$

where $\mathcal{B}_J := \Phi_J \cup \{\Psi_j \ : \ j \geqslant J\}$ *and* $\tilde{\mathcal{B}}_J := \tilde{\Phi}_J \cup \{\tilde{\Psi}_j \ : \ j \geqslant J\}$.

(4) $\mathbb{P}_{m-1}\chi_{[0,N]} \subseteq \mathit{span}(\Phi_j)$ *for some (or all)* $j \geqslant \tilde{J}_0$ *if and only if* $\mathrm{vm}(\tilde{\psi}^L \cup \tilde{\psi}^R \cup \tilde{\psi}) \geqslant m$. *Similarly,* $\mathbb{P}_{\tilde{m}-1}\chi_{[0,N]} \subseteq \mathit{span}(\tilde{\Phi}_j)$ *for some (or all)* $j \geqslant \tilde{J}_0$ *if and only if* $\mathrm{vm}(\psi^L \cup \psi^R \cup \psi) \geqslant \tilde{m}$.

(5) If $[\overline{A_J}^{\mathsf{T}}, \overline{B_J}^{\mathsf{T}}]$ *is invertible for every* $J_0 \leqslant J < \tilde{J}_0$, *then* $(\tilde{\mathcal{B}}_J, \mathcal{B}_J)$ *forms a pair of biorthogonal Riesz bases of* $L_2([0,N])$ *for every* $J_0 \leqslant J < \tilde{J}_0$, *where we recursively define* $\tilde{\Phi}_j := \tilde{A}_j\tilde{\Phi}_{j+1}$ *and* $\tilde{\Psi}_j := \tilde{B}_j\tilde{\Phi}_{j+1}$ *for* j *going from* $\tilde{J}_0 - 1$ *to* J_0 *with the matrices* $\tilde{A}_j$ *and* $\tilde{B}_j$ *in* (5.94).

We now make some remarks on Theorem 5.19. Note that there are no interior elements of Φ_j in (5.90) if $2^j N < n_\phi + n_{\mathring{\phi}}$. Since J_0 is often much smaller than $\tilde{J}_0$, item (5) allows us to have a locally supported Riesz basis $\mathcal{B}_{J_0}$ with simple structures and the smallest coarse scale level J_0. Suppose that $\phi = (\phi^1, \dots, \phi^r)^{\mathsf{T}}, \psi = (\psi^1, \dots, \psi^r)^{\mathsf{T}}, \tilde{\phi}, \tilde{\psi} \in (L_2(\mathbb{R}))^r$ in Theorem 5.19 have the following symmetry:

$$\phi^\ell(c_\ell^\phi - \cdot) = \epsilon_\ell^\phi \phi^\ell, \quad \tilde{\phi}^\ell(c_\ell^\phi - \cdot) = \epsilon_\ell^\phi \tilde{\phi}^\ell \quad \text{with} \quad c_\ell^\phi \in \mathbb{Z}, \epsilon_\ell^\phi \in \{-1, 1\}, \quad \ell = 1, \dots, r, \tag{5.95}$$

$$\psi^\ell(c_\ell^\psi - \cdot) = \epsilon_\ell^\psi \psi^\ell, \quad \tilde{\psi}^\ell(c_\ell^\psi - \cdot) = \epsilon_\ell^\psi \tilde{\psi}^\ell \quad \text{with} \quad c_\ell^\psi \in \mathbb{Z}, \epsilon_\ell^\psi \in \{-1, 1\}, \quad \ell = 1, \dots, r. \tag{5.96}$$

As a consequence of the above symmetry property, up to a possible sign change of some elements, $\mathsf{AS}_0(\phi(-\cdot); \psi(-\cdot))$ is the same as $\mathsf{AS}_0(\phi; \psi)$, while $\mathsf{AS}_0(\tilde{\phi}(-\cdot); \tilde{\psi}(-\cdot))$ is the same as $\mathsf{AS}_0(\tilde{\phi}; \tilde{\psi})$. Hence, using the definition in (5.84), for item (S2) in Theorem 5.19 we can simply choose

$$\mathring{\Phi} = \{\phi^L\} \cup \{\mathring{\phi}^\ell(\cdot - k) \ : \ k \geqslant n_\phi + c_\ell^\phi\}_{\ell=1}^r, \quad \mathring{\Psi} = \{\psi^L\} \cup \{\mathring{\psi}^\ell(\cdot - k) \ : \ k \geqslant n_\psi + c_\ell^\psi\}_{\ell=1}^r, \tag{5.97}$$

$$\mathring{\tilde{\Phi}} = \{\tilde{\phi}^L\} \cup \{\mathring{\tilde{\phi}}^\ell(\cdot - k) \ : \ k \geqslant n_{\tilde{\phi}} + c_\ell^\phi\}_{\ell=1}^r, \quad \mathring{\tilde{\Phi}} = \{\tilde{\psi}^L\} \cup \{\mathring{\tilde{\psi}}^\ell(\cdot - k) \ : \ k \geqslant n_{\tilde{\psi}} + c_\ell^\psi\}_{\ell=1}^r. \tag{5.98}$$

In other words, up to a possible sign change of some elements, $\mathring{\Phi}$, $\mathring{\Psi}$, $\mathring{\tilde{\Phi}}$, and $\mathring{\tilde{\Psi}}$ are the same as $\Phi, \Psi, \tilde{\Phi}, \tilde{\Psi}$, respectively. If a compactly supported biorthogonal wavelet $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ in $L_2(\mathbb{R})$ has the symmetry properties in (5.95) and (5.96), then we always take $\mathring{\Phi}$ and $\mathring{\Psi}$ in

(5.97) and $\mathring{\tilde{\Phi}}$ and $\mathring{\tilde{\Psi}}$ in (5.98) for item (S2) in Theorem 5.19 for all our examples in the next section.

5.7 Examples of orthogonal and biorthogonal wavelets on $[0,1]$

In this section we provide a few examples to illustrate our general construction methods and algorithms. Since the construction of orthogonal wavelets on $[0,1]$ is much simpler than biorthogonal wavelets on $[0,1]$, let us first provide a few examples of orthogonal multiwavelets on $[0,1]$ by Algorithm 5.1 such that the boundary wavelets have the same order of vanishing moments as the interior wavelets. We shall provide examples of wavelets on $[0,1]$ satisfying homogeneous boundary conditions as well. All our examples have the polynomial reproduction property in (5.81) for Φ satisfying (5.80), i.e., for $m := \mathrm{sr}(a)$, $\mathbb{P}_{m-1} \subseteq \mathrm{span}(\{x^n : 0 \leqslant n \leqslant \ell\}) + \mathrm{span}(\Phi_j)$ holds on $[0,1]$ and $h^{(n)}(0) = h^{(n)}(1) = 0$ for all $0 \leqslant n \leqslant \ell$, $h \in \Phi_j$ and $j \geqslant J_0$, where $\ell = -1$ (no boundary conditions) or $\ell \in \{0,1\}$. To avoid possible confusion, we shall use the notations $\phi^{L,bc}, \phi^{L,bc1}$ for ϕ^L, and $\psi^{L,bc}, \psi^{L,bc1}$ for ψ^L if they satisfy the homogeneous Dirichlet boundary conditions for $\ell = 0$ or $\ell = 1$, respectively.

Before presenting our examples, let us recall a technical quantity. For $\tau \in \mathbb{R}$, recall that $\phi \in (H^\tau(\mathbb{R}))^r$ if $\int_{\mathbb{R}} \|\widehat{\phi}(\xi)\|_{l_2}^2 (1+|\xi|^2)^\tau d\xi < \infty$. We define the smoothness exponent $\mathrm{sm}(\phi) := \sup\{\tau \in \mathbb{R} : \phi \in (H^\tau(\mathbb{R}))^r\}$. For $a, \tilde{a} \in (l_0(\mathbb{Z}))^{r\times r}$, let $\phi, \tilde{\phi}$ be compactly supported distributions satisfying $\widehat{\phi}(2\xi) = \widehat{a}(\xi)\widehat{\phi}(\xi)$ and $\widehat{\tilde{\phi}}(2\xi) = \widehat{\tilde{a}}(\xi)\widehat{\tilde{\phi}}(\xi)$ with $\overline{\widehat{\phi}(0)}^{\mathsf{T}}\widehat{\tilde{\phi}}(0) = 1$. It is known (e.g., see [71, Theorem 6.4.5] and [66]) that items (1) and (2) in Theorem 5.1 can be equivalently replaced by $\mathrm{sm}(a) > 0$ and $\mathrm{sm}(\tilde{a}) > 0$, where the technical quantity $\mathrm{sm}(a)$ is defined in [71, (5.6.44)] (also see [66, (4.3)] and [68, (3.2)]) and can be computed (see [89], [66, Theorem 7.1], and [71, Theorem 5.8.4]). The quantity $\mathrm{sm}(a)$ is closely linked to the smoothness of a refinable vector function ϕ through the inequality $\mathrm{sm}(\phi) \geqslant \mathrm{sm}(a)$. For any refinable vector function ϕ in a biorthogonal wavelet, $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ must be a Riesz sequence in $L_2(\mathbb{R})$ and hence we always have $\mathrm{sm}(\phi) = \mathrm{sm}(a)$ (e.g., see [71, Theorem 6.3.3]). See [30, 65, 66, 68, 71, 89, 93] for more details on smoothness $\mathrm{sm}(\phi)$ of refinable vector functions and the quantity $\mathrm{sm}(a)$. Recall that $\mathrm{sr}(a)$ is the highest order of sum rules satisfied by the filter a in (5.23), while $\mathrm{vm}(\psi)$ stands for the highest order of vanishing moments satisfied by ψ. We shall always take n_ψ in Theorem 5.14 to be the smallest integer such that $\psi(\cdot - k) \in \mathsf{S}_1(\Phi)$ for all $k \geqslant n_\psi$.

In what follows, we present three examples. For more examples, see [75, Section 7].

Example 5.2. Consider the compactly supported orthogonal wavelet $\{\phi;\psi\}$ in [40] satisfying $\widehat{\phi}(2\xi) = \widehat{a}(\xi)\widehat{\phi}(\xi)$ and $\widehat{\psi}(2\xi) = \widehat{b}(\xi)\widehat{\phi}(\xi)$ with $\widehat{\phi}(0) = (\sqrt{7/33}, \sqrt{3/4}, \sqrt{5/132})^{\mathsf{T}}$ and an associated finitely supported orthogonal wavelet filter bank $\{a;b\}$ given by

$$a = \left\{ \begin{bmatrix} 0 & -\frac{\sqrt{2}\sqrt{154}(3+2\sqrt{5})}{7392} & \frac{\sqrt{2}\sqrt{14}(2+5\sqrt{5})}{2464} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -\frac{3}{44} - \frac{\sqrt{5}}{22} & \frac{\sqrt{2}\sqrt{154}(67+30\sqrt{5})}{7392} & \frac{\sqrt{2}\sqrt{14}(-10+\sqrt{5})}{224} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{2}\sqrt{154}(67-30\sqrt{5})}{7392} & \frac{\sqrt{2}\sqrt{14}(10+\sqrt{5})}{224} \\ 0 & \frac{3}{8} & \frac{\sqrt{2}\sqrt{22}(-4+\sqrt{5})}{176} \\ 0 & \frac{\sqrt{2}\sqrt{22}(32+7\sqrt{5})}{528} & -\frac{5}{88} + \frac{\sqrt{5}}{22} \end{bmatrix}, \begin{bmatrix} -\frac{3}{44} + \frac{\sqrt{5}}{22} & \frac{\sqrt{2}\sqrt{154}(-3+2\sqrt{5})}{7392} & \frac{\sqrt{2}\sqrt{14}(-2+5\sqrt{5})}{2464} \\ \frac{\sqrt{2}\sqrt{154}}{44} & \frac{3}{8} & \frac{\sqrt{2}\sqrt{22}(4+\sqrt{5})}{176} \\ -\frac{\sqrt{2}\sqrt{70}}{44} & \frac{\sqrt{2}\sqrt{22}(-32+7\sqrt{5})}{528} & -\frac{5}{88} - \frac{\sqrt{5}}{22} \end{bmatrix} \right\}_{[-2,1]},$$

$$b = \left\{ \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{\sqrt{2}\sqrt{154}(3+2\sqrt{5})}{7392} & -\frac{\sqrt{2}\sqrt{14}(2+5\sqrt{5})}{2464} \\ 0 & -\frac{\sqrt{2}\sqrt{7}(1+\sqrt{5})}{672} & \frac{\sqrt{2}\sqrt{77}(-1+3\sqrt{5})}{2464} \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ \frac{3}{44} + \frac{\sqrt{5}}{22} & -\frac{\sqrt{2}\sqrt{154}(67+30\sqrt{5})}{7392} & \frac{\sqrt{2}(10-\sqrt{5})\sqrt{14}}{224} \\ -\frac{\sqrt{2}\sqrt{11}(1+\sqrt{5})}{88} & \frac{\sqrt{2}\sqrt{7}(29+13\sqrt{5})}{672} & \frac{\sqrt{2}\sqrt{77}(-75+17\sqrt{5})}{2464} \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} 0 & \frac{\sqrt{2}\sqrt{77}(-2+\sqrt{5})}{528} & \frac{\sqrt{2}\sqrt{7}(13-6\sqrt{5})}{176} \\ \frac{1}{2} & \frac{\sqrt{2}\sqrt{154}(-67+30\sqrt{5})}{7392} & -\frac{\sqrt{2}\sqrt{14}(10+\sqrt{5})}{224} \\ 0 & \frac{\sqrt{2}\sqrt{7}(-29+13\sqrt{5})}{672} & -\frac{\sqrt{2}\sqrt{77}(75+17\sqrt{5})}{2464} \end{bmatrix}, \begin{bmatrix} \frac{13\sqrt{2}}{44} & -\frac{\sqrt{2}\sqrt{77}(\sqrt{5}+2)}{528} & -\frac{\sqrt{2}\sqrt{7}(13+6\sqrt{5})}{176} \\ \frac{3}{44} - \frac{\sqrt{5}}{22} & \frac{\sqrt{2}\sqrt{154}(3-2\sqrt{5})}{7392} & \frac{\sqrt{2}(2-5\sqrt{5})\sqrt{14}}{2464} \\ \frac{\sqrt{2}\sqrt{11}(1-\sqrt{5})}{88} & \frac{\sqrt{2}\sqrt{7}(1-\sqrt{5})}{672} & -\frac{\sqrt{2}\sqrt{77}(3\sqrt{5}+1)}{2464} \end{bmatrix} \right\}_{[-2,1]}.$$

Note that $\phi = (\phi^1, \phi^2, \phi^3)^{\mathsf{T}}$ is a continuous piecewise linear vector function without symmetry and $\mathrm{fsupp}(\phi) = \mathrm{fsupp}(\psi) = [-1,1]$. Then $\mathrm{sm}(a) = 1.5$, $\mathrm{sr}(a) = 2$, and its matching filter $\upsilon \in (l_0(\mathbb{Z}))^{1\times 2}$ with $\widehat{\upsilon}(0)\widehat{\phi}(0) = 1$ is given by $\widehat{\upsilon}(0) = (\sqrt{7/33}, \sqrt{3}/2, \sqrt{5/132})$ and $\widehat{\upsilon}'(0) = \mathsf{i}(0, \sqrt{3}/4, \sqrt{165}/132 - \sqrt{1/33})$. By item (i) of Proposition 5.4 with $n_\phi = 2$, the left boundary refinable vector functions consisting of interior elements $\phi^2, \phi^3, \phi(\cdot - 1)$ and a true boundary element

$$\phi^L := \sqrt{\tfrac{7(7+\sqrt{5})}{22}}\phi^1\chi_{[0,\infty)} \quad \text{satisfying} \quad \phi^L = \phi^L(2\cdot) + 2[\sqrt{\tfrac{7(7+\sqrt{5})}{22}}, 0, 0]a(1)\phi(2\cdot - 1).$$

Hence, we can reset $n_\phi := 1$ and use only $\{\phi^L, \phi^2, \phi^3\}$ as the left boundary refinable vector function. Let $[a(k)]_{j,:}$ denote the jth row of the matrix $a(k)$. Using $n_\phi = 1$ and $n_\psi = 1$ in Algorithm 5.1, we obtain the left boundary wavelet $\{\psi^L, \psi^1\}$ with $\#\psi^L = 1$ as follows:

$$\psi^L := 2\left([\tfrac{1}{2}, 0, 0] + \lambda_1[b(0)]_{3,:}\right)\phi^L(2\cdot) + 2\lambda_1[b(1)]_{3,:}\phi(2\cdot - 1) \quad \text{with} \quad \lambda_1 := \tfrac{1}{2}\sqrt{\tfrac{7(7-\sqrt{5})}{11}}.$$

Note that $\mathring{\phi} = (\mathring{\phi}^1, \mathring{\phi}^2, \mathring{\phi}^3)^{\mathsf{T}} := \phi(-\cdot)$ has no symmetry. Using item (ii) of Proposition 5.4 with $\mathsf{p}(x) = (1, x)^{\mathsf{T}}$, we have $n_{\mathring{\phi}} = 1$ and the left boundary refinable vector function

$$\mathring{\phi}^L := \tfrac{\sqrt{14}}{\sqrt{7+\sqrt{5}}}\mathring{\phi}^1\chi_{[0,\infty)} \quad \text{satisfying} \quad \begin{aligned} \mathring{\phi}^L = \mathring{\phi}^L(2\cdot) &+ \tfrac{2\sqrt{14}}{\sqrt{7+\sqrt{5}}}[a(-1)]_{1,:}\mathring{\phi}(2\cdot - 1) \\ &+ \tfrac{2\sqrt{14}}{\sqrt{7+\sqrt{5}}}[a(-2)]_{1,:}\mathring{\phi}(2\cdot - 2). \end{aligned}$$

Using $n_{\mathring{\psi}} = 1$ in Algorithm 5.1, we obtain the left boundary wavelet $\mathring{\psi}^L$ with $\#\mathring{\psi}^L = 1$ as follows:

$$\mathring{\psi}^L := \mathring{\phi}^L(2\cdot) + 2\lambda_2[b(-1)]_{2,:}\mathring{\phi}(2\cdot - 1) + 2\lambda_2[b(-2)]_{2,:}\mathring{\phi}(2\cdot - 2) \quad \text{with} \quad \lambda_2 := \sqrt{\tfrac{7(7-\sqrt{5})}{22}}.$$

By (5.89), we have $\phi^R := \mathring{\phi}^L(1-\cdot)$ and $\psi^R := \mathring{\psi}^L(1-\cdot)$. According to Algorithm 5.1 and Theorem 5.19 with $N = 1$, we conclude that $\mathcal{B}_J = \Phi_J \cup \{\Psi_j \ : \ j \geqslant J\}$ is an orthonormal basis of $L_2([0,1])$ for every $J \in \mathbb{N}_0$, where Φ_j and Ψ_j in (5.90) and (5.91) with $n_\phi = n_\psi = n_{\mathring{\phi}} = n_{\mathring{\psi}} = 1$ are given by

$$\begin{aligned}
\Phi_j &= \{\phi^L_{j;0}, \phi^2_{j;0}, \phi^3_{j;0}\} \cup \{\phi_{j;k} : 1 \leqslant k \leqslant 2^j - 1\} \cup \{\phi^R_{j;2^j-1}\},\\
\Psi_j &= \{\psi^L_{j;0}, \psi^1_{j;0}\} \cup \{\psi_{j;k} : 1 \leqslant k \leqslant 2^j - 1\} \cup \{\psi^R_{j;2^j-1}\},
\end{aligned}$$

with $\#\phi^L = \#\phi^R = \psi^L = \psi^R = 1$, $\#\Phi_j = 3(2^j) + 1$ and $\#\Psi_j = 3(2^j)$. Note that $\operatorname{vm}(\psi^L) = \operatorname{vm}(\psi^R) = \operatorname{vm}(\psi) = 2 = \operatorname{sr}(a)$ and $\mathbb{P}_1\chi_{[0,1]} \subset \operatorname{span}(\Phi_j)$ for all $j \in \mathbb{N}_0$.

Using the classical approach in Section 5.3 and Theorem 5.19, we obtain a Riesz basis $\mathcal{B}^{bc}_J := \Phi^{bc}_J \cup \{\Psi^{bc}_j \ : \ j \geqslant J\}$ of $L_2([0,1])$ for every $J \geqslant J_0 := 1$ such that $h(0) = h(1) = 0$ for all $h \in \mathcal{B}^{bc}_J$, where

$$\begin{aligned}
\Phi^{bc}_j &= \{\phi^2_{j;0}, \phi^3_{j;0}, \phi_{j;1}\} \cup \{\phi_{j;k} \ : \ 2 \leqslant k \leqslant 2^j - 2\} \cup \{\phi_{j;2^j-1}\},\\
\Psi^{bc}_j &= \{\psi^{L,bc}_{j;0}, \psi^1_{j;0}, \psi_{j;1}\} \cup \{\psi_{j;k} \ : \ 2 \leqslant k \leqslant 2^j - 2\} \cup \{\psi_{j;2^j-1}, \psi^{R,bc}_{j;2^j-1}\}
\end{aligned}$$

with $\#\psi^{L,bc} = \#\psi^{R,bc} = 1$, $\#\Phi^{bc}_j = 3(2^j) - 1$ and $\#\Psi^{bc}_j = 3(2^j)$, where $\psi^{R,bc} := \mathring{\psi}^{L,bc}(1-\cdot)$ and

$$\begin{aligned}
\psi^{L,bc} &:= -\tfrac{\sqrt{11}(4+\sqrt{5})}{33}\phi^2(2\cdot) + \phi^3(2\cdot) + \left[0, \tfrac{\sqrt{11}(4-\sqrt{5})}{33}, 1\right]\phi(2\cdot - 1),\\
\mathring{\psi}^{L,bc} &:= \left[\tfrac{2}{\sqrt{7}}, \tfrac{\sqrt{11}(\sqrt{5}-4)}{33}, -1\right]\mathring{\phi}(2\cdot - 1).
\end{aligned}$$

Note that $\operatorname{vm}(\psi^{L,bc}) = \operatorname{vm}(\psi^{R,bc}) = \operatorname{vm}(\psi) = 2$ and $\Phi^{bc}_j = \Phi_j \backslash \{\phi^L_{j;0}, \phi^R_{j;2^j-1}\}$ as in Proposition 5.18. Moreover, the dual Riesz basis $\tilde{\mathcal{B}}^{bc}_J$ of $\mathcal{B}^{bc}_J$ is given by $\tilde{\mathcal{B}}^{bc}_J = \tilde{\Phi}^{bc}_j \cup \{\tilde{\Psi}^{bc}_j \ : \ j \geqslant J\}$ with $J \geqslant \tilde{J}_0 = 2$ and

$$\begin{aligned}
\tilde{\Phi}^{bc}_j &= \{\tilde{\phi}^{L,bc}_{j;0}\} \cup \{\phi_{j;k} \ : \ 2 \leqslant k \leqslant 2^j - 2\} \cup \{\tilde{\phi}^{R,bc}_{j;2^j-1}\}, \quad \text{with} \quad \tilde{\phi}^{R,bc} := \mathring{\tilde{\phi}}^{L,bc}(1-\cdot),\\
\tilde{\Psi}^{bc}_j &= \{\tilde{\psi}^{L,bc}_{j;0}\} \cup \{\psi_{j;k} \ : \ 2 \leqslant k \leqslant 2^j - 2\} \cup \{\tilde{\psi}^{R,bc}_{j;2^j-1}\}, \quad \text{with} \quad \tilde{\psi}^{R,bc} := \mathring{\tilde{\psi}}^{L,bc}(1-\cdot),
\end{aligned}$$

$\#\tilde{\phi}^{L,bc} = \#\tilde{\psi}^{L,bc} = 5$, $\#\tilde{\phi}^{R,bc} = 3$, and $\#\tilde{\psi}^{R,bc} = 4$, where

$$\tilde{\phi}^{L,bc} := \sqrt{\tfrac{22}{7(7+\sqrt{5})}}\left[0, \tfrac{4\sqrt{7}(4-\sqrt{5})}{11}, \tfrac{21-8\sqrt{5}}{11}, 0, 0\right]^{\mathsf{T}} \phi^L + [\phi^2(\cdot-1), \phi^3(\cdot-1), \phi(\cdot-2)]^{\mathsf{T}},$$

$$\mathring{\tilde{\phi}}^{L,bc} := \lambda_2^{-1}\left[\tfrac{21+8\sqrt{5}}{11}, 0, -\tfrac{4\sqrt{7}(4+\sqrt{5})}{11}\right]^{\mathsf{T}} \mathring{\phi}^L + \mathring{\phi}(\cdot-1),$$

$$\tilde{\psi}^{L,bc} := \begin{bmatrix} \frac{\sqrt{11}(114-51\sqrt{5})}{104} & \frac{315-54\sqrt{5}}{104} & \frac{9\sqrt{7}(-2+\sqrt{5})}{26} & \frac{3\sqrt{11}(-2+\sqrt{5})}{104} & \frac{18\sqrt{5}-27}{104} \\ \frac{3\sqrt{154}(31\sqrt{5}-69)}{1144} & \frac{\sqrt{14}(167\sqrt{5}-41)}{1144} & \frac{\sqrt{2}(207-49\sqrt{5})}{143} & -\frac{3\sqrt{154}(3\sqrt{5}-1)}{1144} & -\frac{9\sqrt{14}(1+\sqrt{5})}{104} \\ 0 & 0 & 0 & 0 & 0 \\ \frac{3\sqrt{77}(6\sqrt{5}-13)}{308} & \frac{\sqrt{7}(27\sqrt{5}-86)}{308} & \frac{8-2\sqrt{5}}{11} & -\frac{3\sqrt{77}(3+2\sqrt{5})}{308} & -\frac{\sqrt{7}(-6+\sqrt{5})}{28} \\ -\frac{3\sqrt{14}(5\sqrt{5}-11)}{56} & -\frac{\sqrt{154}(25\sqrt{5}-67)}{616} & \frac{\sqrt{22}(\sqrt{5}-3)}{11} & \frac{3\sqrt{14}(1+\sqrt{5})}{56} & \frac{\sqrt{154}(13\sqrt{5}-47)}{616} \end{bmatrix} \tilde{\phi}^{L,bc}(2\cdot)$$
$$+ \begin{bmatrix} 0_{2\times 3} \\ 2b(0) \end{bmatrix} \phi(2\cdot-2) + \begin{bmatrix} 0_{2\times 3} \\ 2b(1) \end{bmatrix} \phi(2\cdot-3),$$

$$\mathring{\tilde{\psi}}^{L,bc} := \begin{bmatrix} -\frac{\sqrt{2}(837+377\sqrt{5})}{11} & \frac{3\sqrt{154}(1503\sqrt{5}+3361)}{616} & -\frac{\sqrt{14}(8683\sqrt{5}+19321)}{616} \\ \frac{8+2\sqrt{5}}{11} & -\frac{3\sqrt{77}(13+6\sqrt{5})}{308} & \frac{\sqrt{7}(27\sqrt{5}+86)}{308} \\ \frac{\sqrt{22}(\sqrt{5}+3)}{11} & -\frac{3\sqrt{14}(11+5\sqrt{5})}{56} & \frac{\sqrt{154}(67+25\sqrt{5})}{616} \\ \frac{\sqrt{7}(504+225\sqrt{5})}{11} & -\frac{\sqrt{11}(1347\sqrt{5}+3012)}{44} & \frac{2592\sqrt{5}+5715}{44} \end{bmatrix} \mathring{\tilde{\phi}}^{L,bc}(2\cdot)$$
$$+ \begin{bmatrix} 0 & -\frac{3\sqrt{154}(189+83\sqrt{5})}{616} & \frac{9\sqrt{14}(13\sqrt{5}+31)}{56} \\ 1 & \frac{3\sqrt{77}(-3+2\sqrt{5})}{308} & -\frac{\sqrt{7}(6+\sqrt{5})}{28} \\ 0 & \frac{3\sqrt{14}(\sqrt{5}-1)}{56} & -\frac{\sqrt{154}(13\sqrt{5}+47)}{616} \\ 0 & \frac{\sqrt{11}(75\sqrt{5}+168)}{44} & -\frac{81+36\sqrt{5}}{4} \end{bmatrix} \mathring{\phi}(2\cdot-2) + \begin{bmatrix} 2b(-1) \\ 0_{1\times 3} \end{bmatrix} \mathring{\phi}(2\cdot-3)$$
$$+ \begin{bmatrix} 2b(-2) \\ 0_{1\times 3} \end{bmatrix} \mathring{\phi}(2\cdot-4).$$

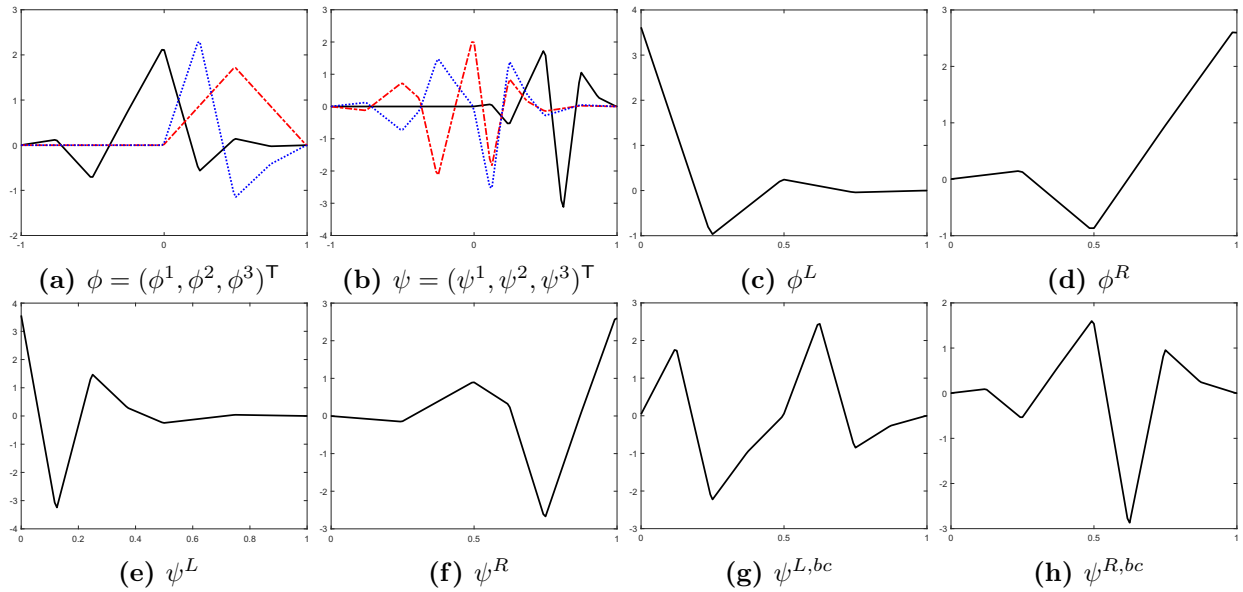
Note that $\tilde{\phi}^{L,bc}$ and $\mathring{\tilde{\phi}}^{L,bc}$ satisfy the refinement equation in (5.19) as follows:

$$\tilde{\phi}^{L,bc} = \begin{bmatrix} \frac{3}{4} & \frac{\sqrt{11}(-4+\sqrt{5})}{44} & \frac{\sqrt{77}}{11} & \frac{3}{4} & \frac{\sqrt{11}(4+\sqrt{5})}{44} \\ -\frac{9\sqrt{11}(-4+\sqrt{5})}{44} & \frac{65-8\sqrt{5}}{44} & \frac{\sqrt{7}(-4+\sqrt{5})}{11} & \frac{3\sqrt{11}(-4+\sqrt{5})}{44} & -\frac{1}{4} \\ \frac{\sqrt{77}(39-18\sqrt{5})}{308} & \frac{\sqrt{7}(86-27\sqrt{5})}{308} & \frac{2\sqrt{5}-8}{11} & \frac{\sqrt{77}(9+6\sqrt{5})}{308} & \frac{\sqrt{7}(\sqrt{5}-6)}{28} \\ & & 0_{2\times 5} & & \end{bmatrix} \tilde{\phi}^{L,bc}(2\cdot)$$
$$+ \begin{bmatrix} 0_{2\times 3} \\ 2a(0) \end{bmatrix} \phi(2\cdot-2) + \begin{bmatrix} 0_{2\times 3} \\ 2a(1) \end{bmatrix} \phi(2\cdot-3),$$

$$\mathring{\tilde{\phi}}^{L,bc} = \begin{bmatrix} -\frac{8+2\sqrt{5}}{11} & \frac{\sqrt{77}(18\sqrt{5}+39)}{308} & -\frac{\sqrt{7}(27\sqrt{5}+86)}{308} \\ \frac{\sqrt{77}}{11} & \frac{3}{4} & \frac{\sqrt{11}(4+\sqrt{5})}{44} \\ \frac{\sqrt{7}(4+\sqrt{5})}{11} & -\frac{9\sqrt{11}(4+\sqrt{5})}{44} & \frac{65+8\sqrt{5}}{44} \end{bmatrix} \mathring{\tilde{\phi}}^{L,bc}(2\cdot)$$

$$+\begin{bmatrix} 1 & -\frac{\sqrt{77}(6\sqrt{5}-9)}{308} & \frac{\sqrt{7}(6+\sqrt{5})}{28} \\ 0 & \frac{3}{4} & \frac{\sqrt{11}(-4+\sqrt{5})}{44} \\ 0 & \frac{3\sqrt{11}(4+\sqrt{5})}{44} & -\frac{1}{4} \end{bmatrix} \phi(2\cdot-2) + 2a(-1)\phi(2\cdot-3) + 2a(-2)\phi(2\cdot-4).$$

We can also directly check that all the conditions in Theorem 5.14 are satisfied for the Riesz basis $\mathcal{B}_J^{bc}$ with $J \geqslant 2$. To avoid complicated presentation, we only mention that the condition in (5.60) is satisfied with $\rho(\tilde{A}_L^{bc}) = 1/2$ for both left and right dual boundary elements. Note that the dual Riesz basis $\tilde{\mathcal{B}}_J^{bc}$ for $J = 1$ has to be computed via item (5) of Theorem 5.19. See Fig. 5.1 for the graphs of ϕ, ψ and all boundary elements.



(a) $\phi = (\phi^1, \phi^2, \phi^3)^\mathsf{T}$ **(b)** $\psi = (\psi^1, \psi^2, \psi^3)^\mathsf{T}$ **(c)** ϕ^L **(d)** ϕ^R

(e) ψ^L **(f)** ψ^R **(g)** $\psi^{L,bc}$ **(h)** $\psi^{R,bc}$

Figure 5.1: The generators of the orthonormal basis $\mathcal{B}_J$ and the Riesz basis $\mathcal{B}_J^{bc}$ of $L_2([0,1])$ in Example 5.2 with $J \geqslant 1$ such that $h(0) = h(1) = 0$ for all $h \in \mathcal{B}_J^{bc}$. The black, red, and blue lines correspond to the first, second, and third components of a vector function. Note that $\mathrm{vm}(\psi^L) = \mathrm{vm}(\psi^R) = \mathrm{vm}(\psi^{L,bc}) = \mathrm{vm}(\psi^{R,bc}) = \mathrm{vm}(\psi) = 2$.

Example 5.3. Using the CBC algorithm in [65, 71], we construct a biorthogonal wavelet $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ with $\widehat{\phi}(0) = \widehat{\tilde{\phi}}(0) = (1,0)^\mathsf{T}$ and a biorthogonal wavelet filter bank $(\{\tilde{a}; \tilde{b}\}, \{a; b\})$ given by

$$a = \left\{ \begin{bmatrix} \frac{1}{4} & \frac{3}{8} \\ -\frac{1}{16} & -\frac{1}{16} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{bmatrix}, \begin{bmatrix} \frac{1}{4} & -\frac{3}{8} \\ \frac{1}{16} & -\frac{1}{16} \end{bmatrix} \right\}_{[-1,1]},$$

$$b = \left\{ \begin{bmatrix} 0 & 0 \\ \frac{2}{97} & \frac{24}{679} \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} & -\frac{15}{4} \\ \frac{77}{1164} & \frac{2921}{2761} \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} & \frac{15}{4} \\ -\frac{77}{1164} & \frac{2921}{2761} \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ -\frac{2}{97} & \frac{24}{679} \end{bmatrix} \right\}_{[-2,2]},$$

$$\tilde{a}=\left\{\begin{bmatrix}-\frac{13}{2432} & -\frac{91}{29184}\\ \frac{3}{152} & \frac{7}{608}\end{bmatrix},\begin{bmatrix}\frac{39}{2432} & \frac{13}{3648}\\ -\frac{9}{152} & -\frac{1}{76}\end{bmatrix},\begin{bmatrix}-\frac{1}{12} & -\frac{1699}{43776}\\ \frac{679}{1216} & \frac{4225}{14592}\end{bmatrix},\begin{bmatrix}\frac{569}{2432} & \frac{647}{10944}\\ -\frac{1965}{1216} & -\frac{37}{96}\end{bmatrix},\begin{bmatrix}\frac{2471}{3648} & 0\\ 0 & \frac{7291}{7296}\end{bmatrix},\right.$$
$$\left.\begin{bmatrix}\frac{569}{2432} & -\frac{647}{10944}\\ \frac{1965}{1216} & -\frac{37}{96}\end{bmatrix},\begin{bmatrix}-\frac{1}{12} & \frac{1699}{43776}\\ -\frac{679}{1216} & \frac{4225}{14592}\end{bmatrix},\begin{bmatrix}\frac{39}{2432} & -\frac{13}{3648}\\ \frac{9}{152} & -\frac{1}{76}\end{bmatrix},\begin{bmatrix}-\frac{13}{2432} & \frac{91}{29184}\\ -\frac{3}{152} & \frac{7}{608}\end{bmatrix}\right\}_{[-4,4]},$$
$$\tilde{b}=\left\{\begin{bmatrix}-\frac{1}{4864} & -\frac{7}{58368}\\ 0 & 0\end{bmatrix},\begin{bmatrix}\frac{3}{4864} & \frac{1}{7296}\\ 0 & 0\end{bmatrix},\begin{bmatrix}\frac{1}{24} & \frac{2161}{87552}\\ -\frac{679}{4864} & -\frac{4753}{58368}\end{bmatrix},\begin{bmatrix}-\frac{611}{4864} & -\frac{605}{21888}\\ \frac{2037}{4864} & \frac{679}{7296}\end{bmatrix},\begin{bmatrix}\frac{1219}{7296} & 0\\ 0 & \frac{7469}{29814}\end{bmatrix},\right.$$
$$\left.\begin{bmatrix}-\frac{611}{4864} & \frac{605}{21888}\\ -\frac{2037}{4864} & \frac{679}{7296}\end{bmatrix},\begin{bmatrix}\frac{1}{24} & -\frac{2161}{87552}\\ \frac{679}{4864} & -\frac{4753}{58368}\end{bmatrix},\begin{bmatrix}\frac{3}{4864} & -\frac{1}{7296}\\ 0 & 0\end{bmatrix},\begin{bmatrix}-\frac{1}{4864} & \frac{7}{58368}\\ 0 & 0\end{bmatrix}\right\}_{[-4,4]}.$$

Note that ϕ is the well-known Hermite cubic splines with $\mathrm{fsupp}(\phi)=[-1,1]$. Note that $\mathrm{fsupp}(\psi)=[-2,2]$ and $\mathrm{fsupp}(\tilde{\phi})=\mathrm{fsupp}(\tilde{\psi})=[-4,4]$. Then $\mathrm{sm}(a)=2.5$, $\mathrm{sm}(\tilde{a})=0.281008$, $\mathrm{sr}(a)=\mathrm{sr}(\tilde{a})=4$, and the matching filters $v,\tilde{v}\in(l_0(\mathbb{Z}))^{1\times 2}$ with $\widehat{v}(0)\widehat{\phi}(0)=\widehat{\tilde{v}}(0)\widehat{\tilde{\phi}}(0)=1$ are given (see [65, 66]) by $\widehat{v}(0,0)=(1,0)$, $\widehat{v}'(0)=(0,\mathsf{i})$, $\widehat{v}''(0)=\widehat{v}'''(0)=(0,0)$ and

$$\widehat{\tilde{v}}(0)=(1,0),\quad \widehat{v}'(0)=\mathsf{i}(0,\tfrac{1}{15}),\quad \widehat{\tilde{v}}''(0)=(-\tfrac{2}{15},0),\quad \widehat{v}'''(0)=\mathsf{i}(0,-\tfrac{2}{105}).$$

We use the direct approach as discussed in Section 5.4. By item (i) of Proposition 5.4 with $n_\phi=1$, the left boundary refinable vector function is $\phi^L:=\phi\chi_{[0,\infty)}$ with $\#\phi^L=2$ and satisfies

$$\phi^L=(\phi_1^L,\phi_2^L)^{\mathsf{T}}:=(\phi_1^L(2\cdot),\tfrac{1}{2}\phi_2^L(2\cdot))^{\mathsf{T}}+2a(1)\phi(2\cdot-1). \tag{5.99}$$

Taking $n_\psi=2$ and $m_\phi=7$ in Theorem 5.14, we have

$$\psi^L=\begin{bmatrix}\psi_1^L\\ \psi_2^L\\ \psi_3^L\end{bmatrix}:=\begin{bmatrix}\phi_1^L(2\cdot)\\ \phi_2^L(2\cdot)\\ 0\end{bmatrix}+\begin{bmatrix}\frac{31}{54} & \frac{533}{36}\\ \frac{1}{36} & \frac{7}{4}\\ 1 & \frac{390}{61}\end{bmatrix}\phi(2\cdot-1)+\begin{bmatrix}-\frac{29}{27} & \frac{59}{9}\\ -\frac{1}{9} & \frac{2}{3}\\ -\frac{52}{61} & \frac{660}{61}\end{bmatrix}\phi(2\cdot-2)+\begin{bmatrix}0 & 0\\ 0 & 0\\ -\frac{9}{61} & 0\end{bmatrix}\phi(2\cdot-3),$$

which satisfies items (i) and (ii) of Theorem 5.14 with $\mathrm{fsupp}(C)=[1,5]$, $\mathrm{fsupp}(D)=[2,5]$ and

$$A_0=\begin{bmatrix}\frac{79}{64} & \frac{49043}{394624} & -\frac{54639}{98656} & -\frac{1399}{24664} & \frac{48111}{49328} & \frac{3653}{24664} & -\frac{97227}{98656} & \frac{1341}{6166} & \frac{64593}{197312} & -\frac{152367}{789248} & \frac{675}{197312} & -\frac{75}{98656} & -\frac{225}{197312} & \frac{525}{789248}\\ -\frac{279}{64} & -\frac{58639}{394624} & \frac{823347}{98656} & -\frac{6135}{24664} & -\frac{561771}{49328} & -\frac{33789}{24664} & \frac{1097199}{98656} & -\frac{30279}{12332} & -\frac{729081}{197312} & \frac{1718679}{789248} & -\frac{7155}{197312} & \frac{795}{98656} & \frac{2385}{197312} & -\frac{5565}{789248}\end{bmatrix}^{\mathsf{T}},$$

$$B_0=\begin{bmatrix}-\frac{15}{64} & -\frac{49043}{394624} & \frac{54639}{98656} & \frac{1399}{24664} & -\frac{48111}{49328} & -\frac{3653}{24664} & \frac{97227}{98656} & -\frac{1341}{6166} & -\frac{64593}{197312} & \frac{152367}{789248} & -\frac{675}{197312} & \frac{75}{98656} & \frac{225}{197312} & -\frac{525}{789248}\\ \frac{279}{128} & \frac{847887}{789248} & -\frac{823347}{197312} & \frac{6135}{49328} & \frac{561771}{98656} & \frac{33789}{49328} & -\frac{1097199}{197312} & \frac{30279}{24664} & \frac{729081}{394624} & -\frac{1718679}{1578496} & \frac{7155}{394624} & -\frac{795}{197312} & -\frac{2385}{394624} & \frac{5565}{1578496}\\ \frac{61}{9216} & \frac{1222501}{170477568} & -\frac{282125}{4735488} & -\frac{82289}{2663712} & \frac{2913787}{7103232} & \frac{1504321}{10654848} & -\frac{847027}{1578496} & \frac{46787}{394624} & \frac{563091}{3156992} & -\frac{1325957}{12627968} & \frac{4941}{3156992} & -\frac{549}{1578496} & -\frac{1647}{3156992} & \frac{3843}{12627968}\end{bmatrix}^{\mathsf{T}},$$

$$C(1)=\begin{bmatrix}-\frac{1}{256} & -\frac{113627}{14206464} & \frac{95129}{1183872} & \frac{10735}{221976} & \frac{552251}{591936} & \frac{33025}{887904} & -\frac{8905}{394624} & -\frac{247}{98656} & \frac{4459}{2367744} & -\frac{324935}{28412928} & \frac{13351}{789248} & -\frac{13351}{3551616} & -\frac{13351}{2367744} & \frac{93457}{28412928}\\ \frac{3}{128} & \frac{99131}{2367744} & -\frac{80633}{197312} & -\frac{8923}{36996} & -\frac{363803}{98656} & -\frac{4033}{147984} & \frac{1200891}{197312} & -\frac{67723}{49328} & -\frac{806635}{394624} & \frac{5513543}{4735488} & \frac{18123}{394624} & -\frac{6041}{591936} & -\frac{6041}{394624} & \frac{422887}{4735488}\end{bmatrix}^{\mathsf{T}},$$

$$C(2)=\begin{bmatrix}0 & -\frac{19}{221976} & \frac{19}{18498} & \frac{19}{27747} & \frac{247}{9249} & \frac{76}{27747} & \frac{513}{6166} & \frac{285}{3083} & \frac{16363565}{22493568} & -\frac{7958533}{269922816} & \frac{1757267}{7497856} & -\frac{1997741}{33740352} & -\frac{772}{9249} & \frac{5248657}{134961408}\\ 0 & \frac{679}{1183872} & -\frac{679}{98656} & -\frac{679}{147984} & -\frac{8827}{49328} & -\frac{679}{36996} & -\frac{54999}{98656} & -\frac{30555}{49328} & -\frac{329829}{937232} & \frac{565273}{468616} & \frac{3023955}{1874464} & -\frac{360659}{937232} & -\frac{54999}{98656} & \frac{2166985}{7497856}\end{bmatrix}^{\mathsf{T}},$$

$$C(3)=\begin{bmatrix}0 & -\frac{13}{2367744} & \frac{13}{197312} & \frac{13}{295968} & \frac{169}{98656} & \frac{13}{73992} & \frac{1053}{197312} & \frac{585}{98656} & -\frac{23611}{295968} & -\frac{72635}{1775808} & \frac{46169}{197312} & \frac{52487}{887904} & \frac{15235991}{22493568} & \frac{455}{89974272}\\ 0 & \frac{1}{49328} & -\frac{3}{12332} & -\frac{1}{6166} & -\frac{39}{6166} & -\frac{2}{3083} & -\frac{243}{12332} & -\frac{135}{6166} & \frac{20441109}{3748928} & \frac{13373051}{44987136} & -\frac{60584455}{3748928} & -\frac{2167229}{5623392} & \frac{15}{468616} & \frac{22477733}{22493568}\end{bmatrix}^{\mathsf{T}},$$

$$C(4) = \begin{bmatrix} 0_{1\times 8} & -\frac{13}{2432} & -\frac{91}{29184} & \frac{39}{2432} & \frac{13}{3648} & -\frac{1}{12} & -\frac{1699}{43776} \\ 0_{1\times 8} & \frac{3}{152} & \frac{7}{608} & -\frac{9}{152} & -\frac{1}{76} & \frac{679}{1216} & \frac{4225}{14592} \end{bmatrix}^{\mathsf{T}}, \qquad C(5) = \begin{bmatrix} 0_{1\times 12} & -\frac{13}{2432} & -\frac{91}{29184} \\ 0_{1\times 12} & \frac{3}{152} & \frac{7}{608} \end{bmatrix}^{\mathsf{T}},$$

$$D(2) = \begin{bmatrix} 0 & \frac{19}{443952} & -\frac{19}{36996} & -\frac{19}{55494} & -\frac{247}{18498} & -\frac{38}{27747} & -\frac{513}{12332} & -\frac{285}{6166} & \frac{6259489}{44987136} & \frac{8864935}{539845632} & -\frac{1886753}{14995712} & \frac{1868255}{67480704} & \frac{386}{9249} & -\frac{6673003}{269922816} \\ 0 & -\frac{679}{4735488} & \frac{679}{394624} & \frac{679}{591936} & \frac{8827}{197312} & \frac{679}{147984} & \frac{54999}{394624} & \frac{30555}{197312} & \frac{18333}{197312} & \frac{79443}{394624} & -\frac{164997}{394624} & \frac{18333}{197312} & \frac{54999}{394624} & -\frac{128331}{1578496} \end{bmatrix}^{\mathsf{T}},$$

$$D(3) = \begin{bmatrix} 0 & -\frac{1}{4735488} & \frac{1}{394624} & \frac{1}{591936} & \frac{13}{197312} & \frac{1}{147984} & \frac{81}{394624} & \frac{45}{197312} & \frac{24745}{591936} & \frac{87377}{3551616} & -\frac{49571}{394624} & -\frac{49085}{1775808} & \frac{7516339}{44987136} & \frac{35}{179948544} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{679}{4864} & -\frac{4753}{58368} & \frac{2037}{4864} & \frac{679}{7296} & 0 & \frac{7469}{29184} \end{bmatrix}^{\mathsf{T}},$$

$$D(4) = \begin{bmatrix} 0_{1\times 8} & -\frac{1}{4864} & -\frac{7}{58368} & \frac{3}{4864} & \frac{1}{7296} & \frac{1}{24} & \frac{2161}{87552} \\ 0_{1\times 8} & 0 & 0 & 0 & 0 & -\frac{679}{4864} & -\frac{4753}{58368} \end{bmatrix}^{\mathsf{T}}, \qquad D(5) = \begin{bmatrix} 0_{1\times 12} & -\frac{1}{4864} & -\frac{7}{58368} \\ 0_{1\times 12} & 0 & 0 \end{bmatrix}^{\mathsf{T}}.$$

Since (5.60) is satisfied with $\rho(\tilde{A}_L) = 1/2$, we conclude from Theorems 5.14 and 5.19 with $N = 1$ that $\mathcal{B}_J = \Phi_J \cup \{\Psi_j \ : \ j \geqslant J\}$ is a Riesz basis of $L_2([0,1])$ for all $J \geqslant J_0 := 2$, where Φ_j and Ψ_j in (5.90) and (5.91) with $n_\phi = n_{\mathring{\phi}} = 1$ and $n_\psi = n_{\mathring{\psi}} = 2$ are given by $\phi^R := \phi^L(1-\cdot)$, $\psi^R := \psi^L(1-\cdot)$ and

$$\Phi_j := \{\phi^L_{j;0}, \phi_{j;1}, \phi_{j;2}, \phi_{j;3}\} \cup \{\phi_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\phi^R_{j;2^j-1}, \phi_{j;2^j-1}, \phi_{j;2^j-2}, \phi_{j;2^j-3}\},$$
$$\Psi_j := \{\psi^L_{j;0}, \psi_{j;2}, \psi_{j;3}\} \cup \{\psi_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\psi^R_{j;2^j-1}, \psi_{j;2^j-2}, \psi_{j;2^j-3}\},$$

with $\#\phi^L = \#\phi^R = 2$, $\#\psi^L = \#\psi^R = 3$, $\#\Phi_j = 2^{j+1} + 2$ and $\#\Psi_j = 2^{j+1}$. Note that $\mathrm{vm}(\psi^L) = \mathrm{vm}(\psi^R) = \mathrm{vm}(\psi) = 4 = \mathrm{sr}(\tilde{a})$. The dual Riesz basis $\tilde{\mathcal{B}}_j$ of $\mathcal{B}_j$ with $j \geqslant \tilde{J}_0 := 3$ is given by Theorem 5.14 through (5.61) and (5.62). We rewrite $\tilde{\phi}^L$ in (5.61) as $\{\tilde{\phi}^L, \tilde{\phi}(\cdot-4), \tilde{\phi}(\cdot-5), \tilde{\phi}(\cdot-6)\}$ with true boundary elements $\tilde{\phi}^L$ and $\#\tilde{\phi}^L = 8$, and rewrite $\tilde{\psi}^L$ in (5.62) as $\{\tilde{\psi}^L, \tilde{\psi}(\cdot-4), \tilde{\psi}(\cdot-5)\}$ with true boundary elements $\tilde{\psi}^L$ and $\#\tilde{\psi}^L = 7$. Hence, $\tilde{\mathcal{B}}_J = \tilde{\Phi}_J \cup \{\tilde{\Psi}_j \ : \ j \geqslant J\}$ for $J \geqslant 3$ is given by

$$\tilde{\Phi}_j := \{\tilde{\phi}^L_{j;0}\} \cup \{\tilde{\phi}_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\tilde{\phi}^R_{j;2^j-1}\}, \quad \text{with} \quad \tilde{\phi}^R := \tilde{\phi}^L(1-\cdot),$$
$$\tilde{\Psi}_j := \{\tilde{\psi}^L_{j;0}\} \cup \{\tilde{\psi}_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\tilde{\psi}^R_{j;2^j-1}\}, \quad \text{with} \quad \tilde{\psi}^R := \tilde{\psi}^L(1-\cdot).$$

Note that $\mathrm{vm}(\tilde{\psi}^L) = \mathrm{vm}(\tilde{\psi}^R) = \mathrm{vm}(\tilde{\psi}) = 4 = \mathrm{sr}(a)$ and $\mathbb{P}_3\chi_{[0,1]} \subset \mathrm{span}(\Phi_j)$ for all $j \geqslant 2$. According to Theorem 5.19 with $N = 1$, $(\tilde{\mathcal{B}}_J, \mathcal{B}_J)$ forms a biorthogonal Riesz basis of $L_2([0,1])$ for every $J \geqslant 3$.

By item (ii) of Proposition 5.4 with $n_\phi = 1$ and $\mathsf{p}(x) = (x, x^2, x^3)^{\mathsf{T}}$, the left boundary refinable vector function is $\phi^{L,bc} := \phi^L_2$ (the second entry of ϕ^L) and satisfies $\phi^{L,bc} = \frac{1}{2}\phi^{L,bc}(2\cdot) + [\frac{1}{8}, -\frac{1}{8}]\phi(2\cdot-1)$ by (5.99). Taking $n_\psi = 2$ and $m_\phi = 7$ in Theorem 5.14, we have $\psi^{L,bc} := (\psi^{L,bc}_1, \psi^L_2, \psi^L_3)^{\mathsf{T}}$ with

$$\psi^{L,bc}_1 := \psi^L_1 - \phi^L_1(2\cdot) - [\tfrac{1121}{2376}, \tfrac{533}{36}]\phi(2\cdot-1) + [\tfrac{989}{594}, \tfrac{17}{18}]\phi(2\cdot-2) + [-\tfrac{61}{88}, \tfrac{195}{44}]\phi(2\cdot-3)$$

with $\#\psi^{L,bc} = 3$ satisfying both items (i) and (ii) of Theorem 5.14, where A^{bc}_0, B^{bc}_0, C^{bc}, and D^{bc} can be easily derived from A_0, B_0, C, and D. More precisely, A^{bc}_0 is obtained from $U^{-1}A_0$ by taking out its first row and first column, and B^{bc}_0, C^{bc}, D^{bc} are obtained from

$U^{-1}B_0, U^{-1}C, U^{-1}D$, respectively by removing their first rows, where the invertible matrix U is given by

$$U := I_{14} + B_0 \begin{bmatrix} -1 & 0 & -\frac{1121}{2376} & -\frac{533}{36} & \frac{989}{594} & \frac{17}{18} & -\frac{61}{88} & \frac{195}{44} & 0_{1\times 6} \\ & & & & 0_{2\times 14} & & & & \end{bmatrix}.$$

Since (5.60) is satisfied with $\rho(\tilde{A}_L^{bc}) = 1/2$, we conclude from Theorems 5.14 and 5.19 that $\mathcal{B}_J^{bc} = \Phi_J^{bc} \cup \{\Psi_j^{bc} \,:\, j \geqslant J\}$ is a Riesz basis of $L_2([0,1])$ for every $J \geqslant J_0 := 2$ such that $h(0) = h(1) = 0$ for all $h \in \mathcal{B}_J^{bc}$, where Φ_j^{bc} and Ψ_j^{bc} in (5.90) and (5.91) with $n_\phi = n_{\mathring{\phi}} = 1$ and $n_\psi = n_{\mathring{\psi}} = 2$ are given by

$$\begin{aligned}
\Phi_j^{bc} &:= \{\phi_{j;0}^{L,bc}, \phi_{j;1}, \phi_{j;2}, \phi_{j;3}\} \cup \{\phi_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\phi_{j;2^j-1}^{R,bc}, \phi_{j;2^j-1}, \phi_{j;2^j-2}, \phi_{j;2^j-3}\}, \\
\Psi_j^{bc} &:= \{\psi_{j;0}^{L,bc}, \psi_{j;2}, \psi_{j;3}\} \cup \{\psi_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\psi_{j;2^j-1}^{R,bc}, \psi_{j;2^j-2}, \psi_{j;2^j-3}\},
\end{aligned}$$

where $\phi^{R,bc} := \phi^{L,bc}(1-\cdot)$ and $\psi^{R,bc} := \psi^{L,bc}(1-\cdot)$ with $\#\phi^{L,bc} = 1$, $\#\psi^{L,bc} = 3$, and $\#\Phi_j^{bc} = \#\Psi_j^{bc} = 2^{j+1}$. For the case $j = 2$, $\Phi_2^{bc} = \{\phi_{2;0}^{L,bc}, \phi_{2;1}, \phi_{2;2}, \phi_{2;3}, \phi_{2;3}^{R,bc}\}$ and $\Psi_2^{bc} = \{\psi_{2;0}^{L,bc}, \psi_{2;2}, \psi_{2;3}^{R,bc}\}$ after removing repeated elements. Note $\mathrm{vm}(\psi^{L,bc}) = \mathrm{vm}(\psi^{R,bc}) = \mathrm{vm}(\psi) = 4$ and $\Phi_j^{bc} = \Phi_j \backslash \{(\phi_1^L)_{j;0}, (\phi_1^R)_{j;2^j-1}\}$ as in Proposition 5.18. The dual Riesz basis $\tilde{\mathcal{B}}_j^{bc}$ of $\mathcal{B}_j^{bc}$ with $j \geqslant \tilde{J}_0 := 3$ is given by Theorem 5.14 through (5.61) and (5.62). We rewrite $\tilde{\phi}^{L,bc}$ in (5.61) as $\{\tilde{\phi}^{L,bc}, \tilde{\phi}(\cdot-4), \tilde{\phi}(\cdot-5), \tilde{\phi}(\cdot-6)\}$ with true boundary elements $\tilde{\phi}^{L,bc}$ and $\#\tilde{\phi}^{L,bc} = 7$, and $\tilde{\psi}^{L,bc}$ in (5.62) as $\{\tilde{\psi}^{L,bc}, \tilde{\psi}(\cdot-4), \tilde{\psi}(\cdot-5)\}$ with true boundary elements $\tilde{\psi}^{L,bc}$ and $\#\tilde{\psi}^{L,bc} = 7$. Hence, $\tilde{\mathcal{B}}_J^{bc} = \tilde{\Phi}_J^{bc} \cup \{\tilde{\Psi}_j^{bc} \,:\, j \geqslant J\}$ is given by

$$\begin{aligned}
\tilde{\Phi}_j^{bc} &:= \{\tilde{\phi}_{j;0}^{L,bc}\} \cup \{\tilde{\phi}_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\tilde{\phi}_{j;2^j-1}^{R,bc}\}, \quad \text{with} \quad \tilde{\phi}^{R,bc} := \tilde{\phi}^{L,bc}(1-\cdot), \\
\tilde{\Psi}_j^{bc} &:= \{\tilde{\psi}_{j;0}^{L,bc}\} \cup \{\tilde{\psi}_{j;k} : 4 \leqslant k \leqslant 2^j - 4\} \cup \{\tilde{\psi}_{j;2^j-1}^{R,bc}\}, \quad \text{with} \quad \tilde{\psi}^{R,bc} := \tilde{\psi}^{L,bc}(1-\cdot).
\end{aligned}$$

Note that $\mathrm{vm}(\tilde{\psi}^{L,bc}) = \mathrm{vm}(\tilde{\psi}^{R,bc}) = 0$ and $\{x\chi_{[0,1]}, x^2\chi_{[0,1]}, x^3\chi_{[0,1]}\} \subset \mathrm{span}(\Phi_j^{bc})$ for all $j \geqslant 2$. By Theorem 5.19 with $N = 1$, $(\tilde{\mathcal{B}}_J^{bc}, \mathcal{B}_J^{bc})$ forms a biorthogonal Riesz basis of $L_2([0,1])$ for every $J \geqslant 3$.

By item (ii) of Proposition 5.4 with $n_\phi = 1$ and $\mathsf{p}(x) = (x^2, x^3)^\mathsf{T}$, we have $\phi^{L,bc1} := \emptyset$. Taking $n_\psi = 2$ and $m_\phi = 7$ in Theorem 5.14, we have $\psi^{L,bc1} := (\psi_1^{L,bc}, \psi_2^{L,bc1}, \psi_3^L)$ with

$$\psi_2^{L,bc1} := \psi_2^L - \phi_2^L(2\cdot) - [\tfrac{1}{36}, \tfrac{7}{4}]\phi(2\cdot-1) + [\tfrac{10}{9}, \tfrac{1048}{183}]\phi(2\cdot-2) - [\tfrac{52}{61}, -\tfrac{660}{61}]\phi(2\cdot-3) - [\tfrac{9}{61}, 0]\phi(2\cdot-4)$$

with $\#\psi^{L,bc1} = 3$ satisfying both items (i) and (ii) of Theorem 5.14, where A_0^{bc1} is obtained from $V^{-1}A_0$ by taking out its first two rows and the first two columns, and $B_0^{bc1}, C^{bc1}, D^{bc1}$ are obtained from $V^{-1}B_0$, $V^{-1}C$, $V^{-1}D$, respectively by removing their first two rows, where

the invertible matrix V is given by

$$V := I_{14} + B_0 \begin{bmatrix} -1 & 0 & -\frac{1121}{2376} & -\frac{533}{36} & \frac{989}{594} & \frac{17}{18} & -\frac{61}{88} & \frac{195}{44} & 0 & 0_{1\times 5} \\ 0 & -1 & -\frac{1}{36} & -\frac{7}{4} & \frac{10}{9} & \frac{1048}{183} & -\frac{52}{61} & \frac{660}{61} & -\frac{9}{61} & 0_{1\times 5} \\ & & & & 0_{1\times 14} & & & & & \end{bmatrix}.$$

Since (5.60) is satisfied with $\rho(\tilde{A}_L^{bc1}) = 1/2$, we conclude from Theorems 5.14 and 5.19 with $N = 1$ that $\mathcal{B}_J^{bc1} = \Phi_J^{bc1} \cup \{\Psi_j^{bc1} : j \geqslant J\}$ is a Riesz basis of $L_2([0,1])$ for every $J \geqslant J_0 := 2$ such that $h(0) = h'(0) = h(1) = h'(1) = 0$ for all $h \in \mathcal{B}_J^{bc1}$, where Φ_j^{bc1} and Ψ_j^{bc1} in (5.90) and (5.91) with $n_\phi = n_{\mathring{\phi}} = 1$ and $n_\psi = n_{\mathring{\psi}} = 2$ are given by: for $j = 2, 3$,

$$\Phi_j^{bc1} := \{\phi_{j;k} : 1 \leqslant k \leqslant 2^j - 1\}, \qquad \Psi_j^{bc1} := \{\psi_{j;0}^{L,bc1}\} \cup \{\psi_{j;k} : 2 \leqslant k \leqslant 2^j - 2\} \cup \{\psi_{j;2^j-1}^{R,bc1}\},$$

where $\psi^{R,bc1} := \psi^{L,bc1}(1 - \cdot)$ with $\#\psi^{L,bc1} = \#\psi^{R,bc1} = 3$, and for $j \geqslant 4$,

$$\Phi_j^{bc1} := \{\phi_{j;k} : 1 \leqslant k \leqslant 4\} \cup \{\phi_{j;k} : 5 \leqslant k \leqslant 2^j - 5\} \cup \{\phi_{j;2^j-k} : 1 \leqslant k \leqslant 4\},$$
$$\Psi_j^{bc1} := \{\psi_{j;0}^{L,bc1}, \psi_{j;2}, \psi_{j;3}, \psi_{j;4}\} \cup \{\psi_{j;k} : 5 \leqslant k \leqslant 2^j - 5\} \cup \{\psi_{j;2^j-1}^{R,bc1}, \psi_{j;2^j-2}, \psi_{j;2^j-3}, \psi_{j;2^j-4}\},$$

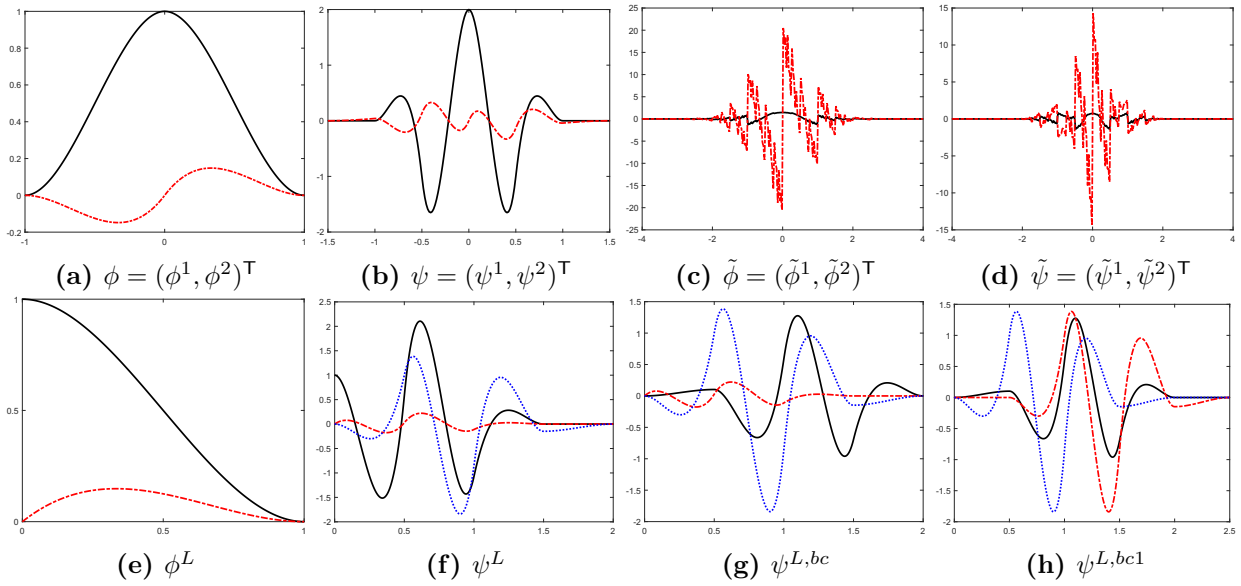
with $\#\Phi_j^{bc1} = 2^{j+1} - 2$ and $\#\Psi_j^{bc1} = 2^{j+1}$. Note that $\mathrm{vm}(\psi^{L,bc1}) = \mathrm{vm}(\psi^{R,bc1}) = \mathrm{vm}(\psi) = 4$ and $\Phi_j^{bc1} = \Phi_j^{bc} \backslash \{\phi_{j;0}^{L,bc}, \phi_{j,2^j-1}^{R,bc}\} = \Phi_j \backslash \{\phi_{j;0}^L, \phi_{j;2^j-1}^R\}$ as in Proposition 5.18. We rewrite $\tilde{\phi}^{L,bc1}$ in (5.61) as $\{\tilde{\phi}^{L,bc1}, \tilde{\phi}(\cdot - 5), \tilde{\phi}(\cdot - 6)\}$ with true boundary elements $\tilde{\phi}^{L,bc1}$ and $\#\tilde{\phi}^{L,bc1} = 8$. Note that $\#\tilde{\psi}^{L,bc1} = 9$. The dual Riesz basis $\tilde{\mathcal{B}}_J^{bc1} := \tilde{\Phi}_J^{bc1} \cup \{\tilde{\Psi}_j^{bc1} : j \geqslant J\}$ of $\mathcal{B}_j^{bc1}$ with $j \geqslant \tilde{J}_0 := 4$ is given by

$$\tilde{\Phi}_j^{bc1} := \{\tilde{\phi}_{j;0}^{L,bc1}\} \cup \{\tilde{\phi}_{j;k} : 5 \leqslant k \leqslant 2^j - 5\} \cup \{\tilde{\phi}_{j;2^j-1}^{R,bc1}\}, \quad \text{with} \quad \tilde{\phi}^{R,bc1} := \tilde{\phi}^{L,bc1}(1 - \cdot),$$
$$\tilde{\Psi}_j^{bc1} := \{\tilde{\psi}_{j;0}^{L,bc1}\} \cup \{\tilde{\psi}_{j;k} : 5 \leqslant k \leqslant 2^j - 5\} \cup \{\tilde{\psi}_{j;2^j-1}^{R,bc1}\}, \quad \text{with} \quad \tilde{\psi}^{R,bc1} := \tilde{\psi}^{L,bc1}(1 - \cdot).$$

Note that $\mathrm{vm}(\tilde{\psi}^{L,bc1}) = \mathrm{vm}(\tilde{\psi}^{R,bc1}) = 0$ and $\{x^2\chi_{[0,1]}, x^3\chi_{[0,1]}\} \subset \mathrm{span}(\Phi_j^{bc1})$ for all $j \geqslant 2$. According to Theorem 5.19 with $N = 1$, $(\tilde{\mathcal{B}}_J^{bc1}, \mathcal{B}_J^{bc1})$ forms a biorthogonal Riesz basis of $L_2([0,1])$ for every $J \geqslant 4$. See Fig. 5.2 for the graphs of ϕ, ψ and their associated boundary elements.

Example 5.4. Consider the scalar biorthogonal wavelet $(\{\tilde{\phi}; \tilde{\psi}\}, \{\phi; \psi\})$ in [29] with $\widehat{\phi}(0) = \widehat{\tilde{\phi}}(0) = 1$ and a biorthogonal wavelet filter bank $(\{\tilde{a}; \tilde{b}\}, \{a; b\})$ given by

$$a = \left\{\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\right\}_{[-1,1]}, \quad b = \left\{-\tfrac{1}{8}, -\tfrac{1}{4}, \tfrac{3}{4}, -\tfrac{1}{4}, -\tfrac{1}{8}\right\}_{[-1,3]},$$
$$\tilde{a} = \left\{-\tfrac{1}{8}, \tfrac{1}{4}, \tfrac{3}{4}, \tfrac{1}{4}, -\tfrac{1}{8}\right\}_{[-2,2]}, \quad \tilde{b} = \left\{-\tfrac{1}{4}, \tfrac{1}{2}, -\tfrac{1}{4}\right\}_{[0,2]}.$$



(a) $\phi = (\phi^1, \phi^2)^\mathsf{T}$ **(b)** $\psi = (\psi^1, \psi^2)^\mathsf{T}$ **(c)** $\tilde{\phi} = (\tilde{\phi}^1, \tilde{\phi}^2)^\mathsf{T}$ **(d)** $\tilde{\psi} = (\tilde{\psi}^1, \tilde{\psi}^2)^\mathsf{T}$

(e) ϕ^L **(f)** ψ^L **(g)** $\psi^{L,bc}$ **(h)** $\psi^{L,bc1}$

Figure 5.2: The generators of the Riesz bases $\mathcal{B}_J$, $\mathcal{B}_J^{bc}$, $\mathcal{B}_J^{bc1}$ for $L_2([0,1])$ with $J \geqslant 2$ in Example 5.3 such that $\eta(0) = \eta(1)$ for all $\eta \in \mathcal{B}_J^{bc}$ and $h(0) = h'(0) = h(1) = h'(1) = 0$ for all $h \in \mathcal{B}_J^{bc1}$. The black, red, and blue lines correspond to the first, second, and third components of a vector function. Note that $\phi^{L,bc}$ in $\mathcal{B}_J^{bc}$ is the second entry of ϕ^L, $\phi^{L,bc1} = \emptyset$ in $\mathcal{B}_J^{bc1}$, and $\mathrm{vm}(\psi^L) = \mathrm{vm}(\psi^{L,bc}) = \mathrm{vm}(\psi^{L,bc1}) = \mathrm{vm}(\psi) = 4$.

Then, $\mathrm{sm}(a) = 1.5$, $\mathrm{sm}(\tilde{a}) \approx 0.440765$, and $\mathrm{sr}(a) = \mathrm{sr}(\tilde{a}) = 2$. Note that ϕ is a piecewise linear function. By item (i) of Proposition 5.4 with $n_\phi = 1$, we have the left boundary refinable function $\phi^L := \phi\chi_{[0,\infty)} = \phi^L(2\cdot) + \frac{1}{2}\phi(2\cdot - 1)$. We use the direct approach in Section 5.4. Taking $n_\psi = 1$ and $m_\phi = 4$ in Theorem 5.14, we have $\psi^L = \phi^L(2\cdot) - \frac{5}{6}\phi(2\cdot - 1) + \frac{1}{3}\phi(2\cdot - 2)$, which satisfies items (i) and (ii) of Theorem 5.14 with $\mathrm{fsupp}(C) = [1,2]$, $\mathrm{fsupp}(D) = [1,1]$, and

$$A_0 = [\tfrac{2}{3}, \tfrac{2}{3}, -\tfrac{1}{3}, 0]^\mathsf{T}, \quad B_0 = [\tfrac{1}{3}, -\tfrac{2}{3}, \tfrac{1}{3}, 0]^\mathsf{T}, \quad C(1) = [-\tfrac{7}{72}, \tfrac{7}{36}, \tfrac{7}{9}, \tfrac{1}{4}]^\mathsf{T},$$
$$C(2) = [\tfrac{1}{72}, -\tfrac{1}{36}, -\tfrac{1}{9}, \tfrac{1}{4}]^\mathsf{T}, \quad D(1) = [\tfrac{1}{36}, -\tfrac{1}{18}, -\tfrac{2}{9}, \tfrac{1}{2}]^\mathsf{T}.$$

Since (5.60) is satisfied with $\rho(\tilde{A}_L) = 1/2$, we conclude from Theorems 5.14 and 5.19 with $N = 1$ that $\mathcal{B}_J = \Phi_J \cup \{\Psi_j : j \geqslant J\}$ is a Riesz basis of $L_2([0,1])$ for all $J \geqslant J_0 := 1$, where Φ_j and Ψ_j in (5.90) and (5.91) with $n_\phi = n_{\mathring{\phi}} = n_\psi = 1$ and $n_{\mathring{\psi}} = 2$ are given by

$$\Phi_j := \{\phi_{j;0}^L, \phi_{j;1}, \phi_{j;2}\} \cup \{\phi_{j;k} : 3 \leqslant k \leqslant 2^j - 3\} \cup \{\phi_{j;2^j-1}^R, \phi_{j;2^j-1}, \phi_{j;2^j-2}\},$$
$$\Psi_j := \{\psi_{j;0}^L, \psi_{j;1}\} \cup \{\psi_{j;k} : 2 \leqslant k \leqslant 2^j - 3\} \cup \{\psi_{j;2^j-1}^R, \psi_{j;2^j-2}\},$$

where $\phi^R := \phi^L(1-\cdot)$ and $\psi^R := \psi^L(1-\cdot)$ with $\#\phi^L = \#\phi^R = \#\psi^L = \#\psi^R = 1$, $\#\Phi_j = 2^j + 1$ and $\#\Psi_j = 2^j$. For the cases $j = 1$ and $j = 2$, $\Phi_1 = \{\phi_{1;0}^L, \phi_{1;1}, \phi_{1;1}^R\}$, $\Psi_1 = \{\psi_{1;0}^L, \psi_{1;1}^R\}$,

and $\Phi_2 = \{\phi^L_{2,0}, \phi_{2;1}, \phi_{2;2}, \phi_{2;3}, \phi^R_{2;3}\}$ after removing repeated elements. Note that $\mathrm{vm}(\psi^L) = \mathrm{vm}(\psi^R) = \mathrm{vm}(\psi) = 2 = \mathrm{sr}(\tilde{a})$. The dual Riesz basis $\tilde{\mathcal{B}}_j$ of $\mathcal{B}_j$ with $j \geqslant \tilde{J}_0 := 3$ is given by Theorem 5.14 through (5.61) and (5.62). We rewrite $\tilde{\phi}^L$ in (5.61) as $\{\tilde{\phi}^L, \tilde{\phi}(\cdot - 3)\}$ with true boundary elements $\tilde{\phi}^L$ and $\#\tilde{\phi}^L = 3$, and rewrite $\tilde{\psi}^L$ in (5.62) as $\{\tilde{\psi}^L, \tilde{\psi}(\cdot - 2), \tilde{\psi}(\cdot - 3)\}$ with true boundary elements $\tilde{\psi}^L$ and $\#\tilde{\psi}^L = 2$. Hence, $\tilde{\mathcal{B}}_J = \tilde{\Phi}_J \cup \{\tilde{\Psi}_j \,:\, j \geqslant J\}$ for $J \geqslant 3$ is given by

$$\begin{aligned}
\tilde{\Phi}_j &:= \{\tilde{\phi}^L_{j;0}\} \cup \{\tilde{\phi}_{j;k} : 3 \leqslant k \leqslant 2^j - 3\} \cup \{\tilde{\phi}^R_{j;2^j-1}\}, \quad \text{with} \quad \tilde{\phi}^R := \tilde{\phi}^L(1 - \cdot), \\
\tilde{\Psi}_j &:= \{\tilde{\psi}^L_{j;0}\} \cup \{\tilde{\psi}_{j;k} : 2 \leqslant k \leqslant 2^j - 3\} \cup \{\tilde{\psi}^R_{j;2^j-1}\}, \quad \text{with} \quad \tilde{\psi}^R := \tilde{\psi}^L(1 - \cdot).
\end{aligned}$$

Note that $\mathrm{vm}(\tilde{\psi}^L) = \mathrm{vm}(\tilde{\psi}^R) = \mathrm{vm}(\tilde{\psi}) = 2 = \mathrm{sr}(a)$ and $\mathbb{P}_1\chi_{[0,1]} \subset \mathrm{span}(\Phi_j)$ for all $j \geqslant 1$. According to Theorem 5.19 with $N = 1$, $(\tilde{\mathcal{B}}_J, \mathcal{B}_J)$ forms a biorthogonal Riesz basis of $L_2([0,1])$ for every $J \geqslant 3$.

By item (ii) of Proposition 5.4 with $n_\phi = 1$ and $\mathsf{p}(x) = x$, the left boundary refinable vector function is $\phi^{L,bc} := \emptyset$. Taking $n_\psi = 1$ and $m_\phi = 4$ in Theorem 5.14, we have $\#\psi^{L,bc} = 1$ and

$$\psi^{L,bc} := \psi^L - \phi^L(2\cdot) + \tfrac{4}{3}\phi(2\cdot - 1) - \tfrac{4}{3}\phi(2\cdot - 2) + \tfrac{1}{2}\phi(2\cdot - 3)$$

satisfying both items (i) and (ii) of Theorem 5.14, where A^{bc}_0, B^{bc}_0, C^{bc}, and D^{bc} can be easily derived from A_0, B_0, C, and D. More precisely, A^{bc}_0 is obtained from $U^{-1}A_0$ by taking out its first row and first column, and B^{bc}_0, C^{bc}, D^{bc} are obtained from $U^{-1}B_0, U^{-1}C, U^{-1}D$, respectively by removing their first rows, where the invertible matrix U is given by

$$U := I_4 + B_0[-1, \tfrac{4}{3}, -\tfrac{4}{3}, \tfrac{1}{2}].$$

Since (5.60) is satisfied with $\rho(\tilde{A}^{bc}_L) = 1/2$, we conclude from Theorems 5.14 and 5.19 with $N = 1$ that $\mathcal{B}^{bc}_J = \Phi^{bc}_J \cup \{\Psi^{bc}_j \,:\, j \geqslant J\}$ is a Riesz basis of $L_2([0,1])$ for every $J \geqslant J_0 := 2$ such that $h(0) = h(1) = 0$ for all $h \in \mathcal{B}^{bc}_J$, where Φ^{bc}_j and Ψ^{bc}_j in (5.90) and (5.91) with $n_\phi = n_{\mathring{\phi}} = n_\psi = 1$ and $n_{\mathring{\psi}} = 2$ are given by

$$\begin{aligned}
\Phi^{bc}_j &:= \{\phi_{j;1}, \phi_{j;2}\} \cup \{\phi_{j;k} : 3 \leqslant k \leqslant 2^j - 3\} \cup \{\phi_{j;2^j-1}, \phi_{j;2^j-2}\}, \\
\Psi^{bc}_j &:= \{\psi^{L,bc}_{j;0}, \psi_{j;1}\} \cup \{\psi_{j;k} : 2 \leqslant k \leqslant 2^j - 3\} \cup \{\psi^{R,bc}_{j;2^j-1}, \psi_{j;2^j-2}\},
\end{aligned}$$

where $\phi^{L,bc} = \phi^{R,bc} = \emptyset$ and $\psi^{R,bc} := \psi^{L,bc}(1 - \cdot)$ with $\#\psi^{L,bc} = \#\psi^{R,bc} = 1$, $\#\Phi^{bc}_j = 2^j - 1$, and $\#\Psi^{bc}_j = 2^j$. For the case $j = 2$, $\Phi^{bc}_2 = \{\phi_{2;1}, \phi_{2;2}, \phi_{2;3}\}$ after removing repeated elements. Note that $\mathrm{vm}(\psi^{L,bc}) = \mathrm{vm}(\psi^{R,bc}) = \mathrm{vm}(\psi) = 2$ and $\Phi^{bc}_j = \Phi_j \backslash \{\phi^L_{j;0}, \phi^R_{j;2^j-1}\}$ as in Proposition 5.18. We rewrite $\tilde{\phi}^{L,bc}$ in (5.61) as $\{\tilde{\phi}^{L,bc}, \tilde{\phi}(\cdot - 3)\}$ with true boundary elements

$\tilde{\phi}^{L,bc}$ and $\#\tilde{\phi}^{L,bc} = 2$, and $\tilde{\psi}^{L,bc}$ in (5.62) as $\{\tilde{\psi}^{L,bc}, \tilde{\psi}(\cdot - 2), \tilde{\psi}(\cdot - 3)\}$ with true boundary elements $\tilde{\psi}^{L,bc}$ and $\#\tilde{\psi}^{L,bc} = 2$. The dual Riesz basis $\tilde{\mathcal{B}}_J^{bc} := \tilde{\Phi}_J^{bc} \cup \{\tilde{\Psi}_j^{bc} : j \geqslant J\}$ of $\mathcal{B}_J^{bc}$ with $J \geqslant \tilde{J}_0 := 3$ is given by

$$\tilde{\Phi}_j^{bc} := \{\tilde{\phi}_{j;0}^{L,bc}\} \cup \{\tilde{\phi}_{j;k} : 3 \leqslant k \leqslant 2^j - 3\} \cup \{\tilde{\phi}_{j;2^j-1}^{R,bc}\}, \quad \text{with} \quad \tilde{\phi}^{R,bc} := \tilde{\phi}^{L,bc}(1 - \cdot),$$
$$\tilde{\Psi}_j^{bc} := \{\tilde{\psi}_{j;0}^{L,bc}\} \cup \{\tilde{\psi}_{j;k} : 2 \leqslant k \leqslant 2^j - 3\} \cup \{\tilde{\psi}_{j;2^j-1}^{R,bc}\}, \quad \text{with} \quad \tilde{\psi}^{R,bc} := \tilde{\psi}^{L,bc}(1 - \cdot).$$

Note $x\chi_{[0,1]} \subset \operatorname{span}(\Phi_j^{bc})$ for all $j \geqslant 2$. By Theorem 5.19 with $N = 1$, $(\tilde{\mathcal{B}}_J^{bc}, \mathcal{B}_J^{bc})$ forms a biorthogonal Riesz basis of $L_2([0,1])$ for $J \geqslant 3$. See Fig. 5.3 for the graphs of $\phi, \psi, \tilde{\phi}, \tilde{\psi}$ and all boundary elements.

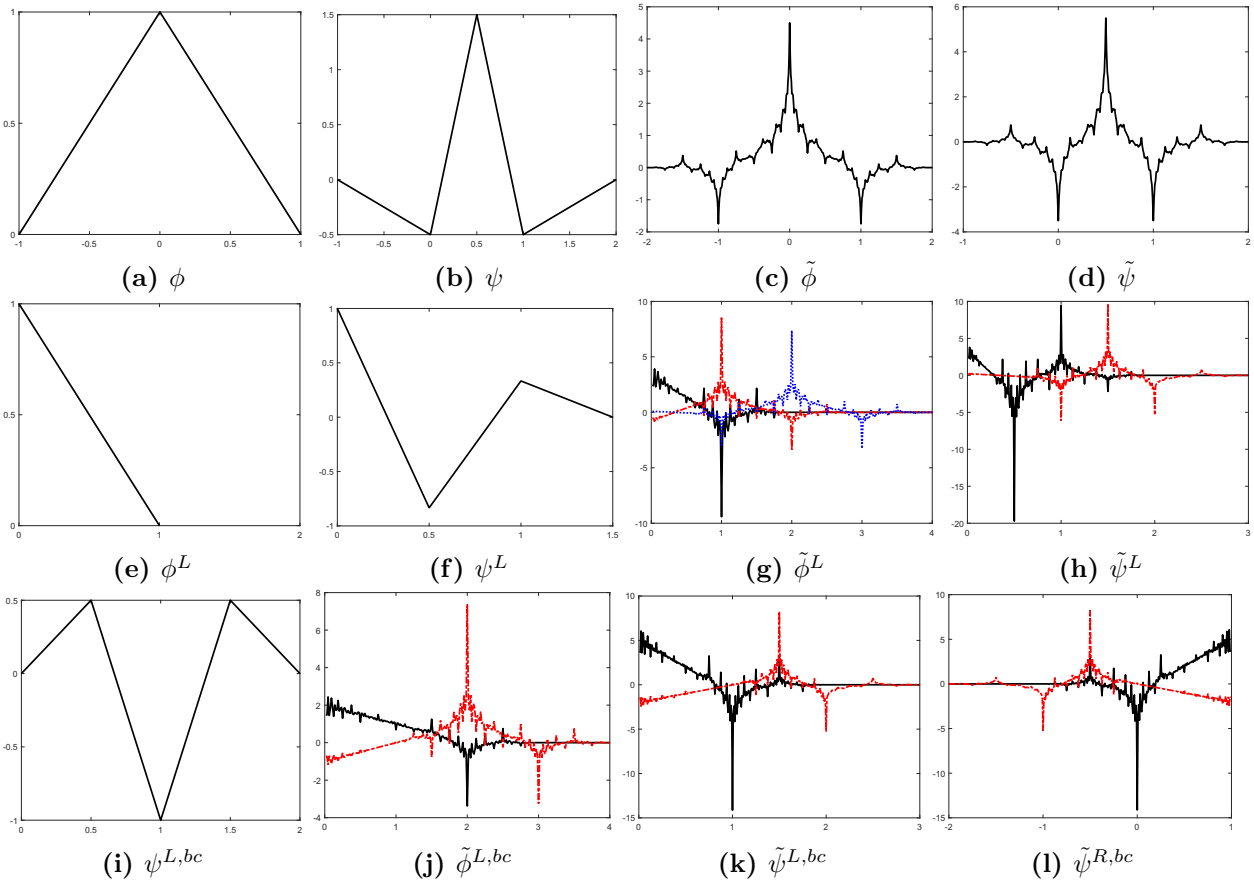


Figure 5.3: The generators of the biorthogonal wavelet bases $(\tilde{\mathcal{B}}_J, \mathcal{B}_J)$ and $(\tilde{\mathcal{B}}_J^{bc}, \mathcal{B}_J^{bc})$ of $L_2([0,1])$ for $J \geqslant 3$ in Example 5.4 with $h(0) = h(1) = 0$ for all $h \in \mathcal{B}_J^{bc}$. The black, red, and blue lines correspond to the first, second, and third components of a vector function. Note that $\operatorname{vm}(\psi^L) = \operatorname{vm}(\psi^{L,bc}) = \operatorname{vm}(\psi) = 2$.

5.8 Proofs of Theorems 5.2, 5.7, 5.10, 5.14, 5.15 and 5.19

In this section, we provide the detailed proofs for Theorems 5.2, 5.7, 5.10, 5.14, 5.15 and 5.19.

Proof of Theorem 5.2. Since $n_\phi \geqslant \max(-l_\phi, -l_a)$ and $n_\psi \geqslant \max(-l_\psi, \frac{n_\phi - l_b}{2})$, the relations in (5.13) and (5.14) must hold, see Section 5.2.2. Hence, $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)} \subseteq L_2([0,\infty))$. By our assumption that $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is a Riesz basis of $L_2([0,\infty))$, it must have a unique dual Riesz basis, denoted by $\tilde{\mathcal{B}}$ here, in $L_2([0,\infty))$. We first prove that $\tilde{\mathcal{B}} = \mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ for some $\tilde{\Phi}, \tilde{\Psi}$ in (5.15). Since both ϕ^L and ψ^L have compact support, we have $\mathrm{fsupp}(\phi^L) \cup \mathrm{fsupp}(\psi^L) \subseteq [0, N]$ for some $N \in \mathbb{N}$. Consequently, we have

$$\mathrm{supp}(\phi^L(2^j\cdot)) \cup \mathrm{supp}(\psi^L(2^j\cdot)) \subseteq [0, 2^{-j}N] \subseteq [0, N] \quad \text{for all} \quad j \in \mathbb{N}_0.$$

We take $n_{\tilde{\phi}} \geqslant \max(-l_{\tilde{\phi}}, -l_{\tilde{a}}, n_\phi)$ such that the supports of all $\tilde{\phi}(\cdot - k), k \geqslant n_{\tilde{\phi}}$ do not essentially overlap with $[0, N]$. Similarly, we take $n_{\tilde{\psi}} \geqslant \max(-l_{\tilde{\psi}}, \frac{n_{\tilde{\phi}} - l_{\tilde{b}}}{2}, n_\psi)$ such that the supports of all $\tilde{\psi}(\cdot - k), k \geqslant n_{\tilde{\psi}}$ do not essentially overlap with $[0, N]$. Consequently, (5.21) and (5.22) must hold and we trivially have

$$\langle \tilde{\phi}(\cdot - k_0), \phi^L(2^j\cdot)\rangle = 0, \quad \langle \tilde{\phi}(\cdot - k_0), \psi^L(2^j\cdot)\rangle = 0, \qquad \forall\, k_0 \geqslant n_{\tilde{\phi}}, j \in \mathbb{N}_0, \tag{5.100}$$

and

$$\langle \tilde{\psi}(\cdot - k_0), \phi^L(2^j\cdot)\rangle = 0, \quad \langle \tilde{\psi}(\cdot - k_0), \psi^L(2^j\cdot)\rangle = 0, \qquad \forall\, k_0 \geqslant n_{\tilde{\psi}}, j \in \mathbb{N}_0. \tag{5.101}$$

Let $k_0 \geqslant n_{\tilde{\phi}}$ be arbitrarily fixed. Since $(\{\tilde{\phi};\tilde{\psi}\}, \{\phi;\psi\})$ is a biorthogonal wavelet in $L_2(\mathbb{R})$, it is trivial that

$$\langle \tilde{\phi}(\cdot - k_0), \phi(\cdot - k_0)\rangle = I_r \quad \text{and} \quad \langle \tilde{\phi}(\cdot - k_0), h\rangle = 0 \qquad \forall h \in \mathsf{AS}_0(\phi;\psi)\backslash\{\phi(\cdot - k_0)\}.$$

In particular, we have $\langle \tilde{\phi}(\cdot - k_0), \phi(\cdot - k)\rangle = 0$ for all $k \geqslant n_\phi$ but $k \neq k_0$ and $\langle \tilde{\phi}(\cdot - k_0), \psi_{j;k}\rangle = 0$ for all $j \in \mathbb{N}_0$ and $k \in \mathbb{Z}$. Now it follows from (5.100) that $\tilde{\phi}(\cdot - k_0)$ must be the unique biorthogonal element/vector in $L_2([0,\infty))$ corresponding to the element $\phi(\cdot - k_0) \in \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$; more precisely, $\langle \tilde{\phi}(\cdot - k_0), \phi(\cdot - k_0)\rangle = I_r$ and $\langle \tilde{\phi}(\cdot - k_0), h\rangle = 0$ for all $h \in \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}\backslash\{\phi(\cdot - k_0)\}$.

Let $j_0 \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and $k_0 \geqslant n_{\tilde{\psi}}$ be arbitrarily fixed. We now show that $\tilde{\psi}_{j_0;k_0}$ is the unique biorthogonal element in $L_2([0,\infty))$ corresponding to the element $\psi_{j_0;k_0} \in \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$. Indeed, it follows from the biorthogonality relation between $\mathsf{AS}_0(\tilde{\phi};\tilde{\psi})$ and

$\mathsf{AS}_0(\phi;\psi)$ that

$$\langle\tilde{\psi}_{j_0;k_0},\phi_{j_0;k}\rangle=0\quad\text{and}\quad\langle\tilde{\psi}_{j_0;k_0},\psi_{j;k}\rangle=\boldsymbol{\delta}(j-j_0)\boldsymbol{\delta}(k-k_0),\qquad\forall\, j\in\mathbb{N}_0,k\in\mathbb{Z}.\tag{5.102}$$

By (5.101), we have

$$\langle\tilde{\psi}_{j_0;k_0},\phi^L(2^{j_0}\cdot)\rangle=0,\quad\langle\tilde{\psi}_{j_0;k_0},\psi^L(2^j\cdot)\rangle=2^{-j_0/2}\langle\tilde{\psi}(\cdot-k_0),\psi^L(2^{j-j_0}\cdot)\rangle=0,\qquad\forall\, j\geqslant j_0.\tag{5.103}$$

By the identities in (5.17), (5.18), (5.21) and (5.22), we see that every element in

$$S_{j_0}:=\Phi\cup\{2^{j/2}\eta(2^j\cdot)\ :\ j=0,\dots,j_0-1,\eta\in\Psi\}\tag{5.104}$$

is a finite linear combination of $\{2^{j_0/2}\phi^L(2^{j_0}\cdot)\}\cup\{\phi_{j_0;k}\ :\ k\geqslant n_\phi\}$. Consequently, it follows from (5.102) and (5.103) that $\langle\tilde{\psi}_{j_0;k_0},h\rangle=0$ for all $h\in S_{j_0}$. Hence, by (5.102), we proved that $\tilde{\psi}_{j_0;k_0}$ is the unique biorthogonal element in $L_2([0,\infty))$ corresponding to the element $\psi_{j_0;k_0}\in\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$.

Since $n_{\tilde{\phi}}\geqslant n_\phi$, we add the elements $\phi(\cdot-k),n_\phi\leqslant k<n_{\tilde{\phi}}$ to the vector function ϕ^L to form a new vector function $\mathring{\phi}^L$. We define $\mathring{\psi}^L$ similarly by adding $\psi(\cdot-k),n_\psi\leqslant k<n_{\tilde{\psi}}$ to ψ^L. Let $\tilde{\phi}^L$ be the unique biorthogonal element/vector in $L_2([0,\infty))$ corresponding to $\mathring{\phi}^L$, and $\tilde{\psi}^L$ be the unique biorthogonal element/vector in $L_2([0,\infty))$ corresponding to $\mathring{\psi}^L$ for the Riesz basis $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ of $L_2([0,\infty))$. Let $j_0\in\mathbb{N}_0$ be arbitrarily fixed. We now prove that $2^{j_0/2}\tilde{\psi}^L(2^{j_0}\cdot)$ is the unique biorthogonal element in $L_2([0,\infty))$ corresponding to $2^{j_0/2}\mathring{\psi}^L(2^{j_0}\cdot)\in\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$. By the definition of $\tilde{\psi}^L$, we have

$$\langle 2^{j_0/2}\tilde{\psi}^L(2^{j_0}\cdot),\psi_{j;k}\rangle=\langle\tilde{\psi}^L,\psi_{j-j_0;k}\rangle=0,\qquad\forall\, j\geqslant j_0,k\geqslant n_{\tilde{\phi}}$$

and

$$\langle 2^{j_0/2}\tilde{\psi}^L(2^{j_0}\cdot),2^{j/2}\mathring{\psi}^L(2^j\cdot)\rangle=\langle\tilde{\psi}^L,2^{(j-j_0)/2}\mathring{\psi}^L(2^{j-j_0}\cdot)\rangle=\boldsymbol{\delta}(j-j_0)I_{\#\tilde{\psi}^L},\qquad\forall\, j\geqslant j_0.$$

Define the set S_{j_0} as in (5.104). As we proved before, every element in S_{j_0} must be a finite linear combination of $\{2^{j_0/2}\mathring{\phi}^L(2^{j_0}\cdot)\}\cup\{\phi_{j_0;k}\ :\ k\geqslant n_{\tilde{\phi}}\}$. By the definition of $\tilde{\psi}^L$, we must have

$$\langle 2^{j_0/2}\tilde{\psi}^L(2^{j_0}\cdot),2^{j_0/2}\mathring{\phi}^L(2^{j_0}\cdot)\rangle=\langle\tilde{\psi}^L,\mathring{\phi}^L\rangle=0$$

and

$$\langle 2^{j_0/2}\tilde{\psi}^L(2^{j_0}\cdot),\phi_{j_0;k}\rangle=\langle\tilde{\psi}^L,\phi(\cdot-k)\rangle=0,\qquad\forall\, k\geqslant n_{\tilde{\phi}}.$$

Consequently, we proved that $\langle 2^{j_0/2}\tilde{\psi}^L(2^{j_0}\cdot),h\rangle=0$ for all $h\in S_{j_0}$. This shows that

$2^{j_0/2}\tilde{\psi}^L(2^{j_0}\cdot)$ must be the unique biorthogonal element in $L_2([0,\infty))$ corresponding to $2^{j_0/2}\mathring{\psi}^L(2^{j_0}\cdot) \in \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$. In summary, we proved $\tilde{\mathcal{B}} = \mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$. Note that $\#\tilde{\phi}^L = \#\mathring{\phi}^L = \#\phi^L + (n_{\tilde{\phi}} - n_\phi)(\#\phi)$ and $\#\tilde{\psi}^L = \#\psi^L = \#\psi^L + (n_{\tilde{\psi}} - n_\psi)(\#\psi)$. Therefore, (5.16) holds. To complete the proof of item (1) in Theorem 5.2, next we prove that both $\tilde{\phi}^L$ and $\tilde{\psi}^L$ must have compact support. Since $(\{\tilde{\phi};\tilde{\psi}\},\{\phi;\psi\})$ is a biorthogonal wavelet in $L_2(\mathbb{R})$, we can expand $\tilde{\phi}^L$ in (1.2) with $J=0$ as follows:

$$\tilde{\phi}^L = \sum_{k\in\mathbb{Z}} \langle \tilde{\phi}^L, \phi(\cdot-k)\rangle \tilde{\phi}(\cdot-k) + \sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}} \langle \tilde{\phi}^L, \psi_{j;k}\rangle \tilde{\psi}_{j;k}.$$

Since $\tilde{\phi}^L$ is perpendicular to all elements in $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}\backslash\{\mathring{\phi}^L\}$, we see from the above identity that

$$\tilde{\phi}^L = \sum_{k=-\infty}^{n_{\tilde{\phi}}-1} \langle \tilde{\phi}^L, \phi(\cdot-k)\rangle \tilde{\phi}(\cdot-k) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{n_{\tilde{\psi}}-1} \langle \tilde{\phi}^L, \psi_{j;k}\rangle \tilde{\psi}_{j;k}, \tag{5.105}$$

from which we deduce that $\tilde{\phi}^L$ must be supported inside $(-\infty, M]$ with $M := \max(n_{\tilde{\phi}} + h_{\tilde{\phi}}, n_{\tilde{\psi}} + h_{\tilde{\psi}})$. Because $\tilde{\phi}^L$ lies in $L_2([0,\infty))$ and hence is supported inside $[0,\infty)$, we conclude that $\tilde{\phi}^L$ must have compact support with $\mathrm{fsupp}(\tilde{\phi}^L) \subseteq [0,M]$. By the same argument, we can prove that (5.105) holds with $\tilde{\phi}^L$ being replaced by $\tilde{\psi}^L$ and hence $\tilde{\psi}^L$ also has compact support. This proves item (1).

We now prove item (2). Since $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ and $\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ form a pair of biorthogonal Riesz bases in $L_2([0,\infty))$, by a simple scaling argument (e.g., see [70, Proposition 4 and (2.6)] and [71, Theorem 4.3.3]), it is straightforward to verify that $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ and $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ form a pair of biorthogonal Riesz bases in $L_2([0,\infty))$ for every $J\in\mathbb{Z}$. Expanding $\tilde{\phi}^L$ under the biorthogonal basis formed by $\mathsf{AS}_1(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ and $\mathsf{AS}_1(\Phi;\Psi)_{[0,\infty)}$, we have

$$\tilde{\phi}^L = \sum_{h\in\mathsf{AS}_1(\Phi;\Psi)_{[0,\infty)}} \langle \tilde{\phi}^L, h\rangle \tilde{h} = 2\langle \tilde{\phi}^L, \mathring{\phi}^L(2\cdot)\rangle \tilde{\phi}^L(2\cdot) + 2\sum_{k=n_{\tilde{\phi}}}^{\infty} \langle \tilde{\phi}^L, \phi(2\cdot-k)\rangle \tilde{\phi}(2\cdot-k),$$

since $\langle \tilde{\phi}^L, h\rangle = 0$ for all $h\in\Psi(2^j\cdot)$ with $j\geqslant 0$. Hence, (5.19) holds with $\tilde{A}_L := \langle \tilde{\phi}^L, \mathring{\phi}^L(2\cdot)\rangle$ and $\tilde{A}(k) := \langle \tilde{\phi}^L, \phi(2\cdot-k)\rangle$ for $k\geqslant n_{\tilde{\phi}}$. Since both $\tilde{\phi}^L$ and ϕ have compact support, the sequence A must be finitely supported. The identity in (5.20) can be proved similarly by expanding $\tilde{\psi}^L$ instead of $\tilde{\phi}^L$, under the biorthogonal basis formed by $\mathsf{AS}_1(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ and $\mathsf{AS}_1(\Phi;\Psi)_{[0,\infty)}$. Using the same argument as in the proof of (5.13), we see that the identities in (5.21) and (5.22) follow directly from (5.6) and the assumption that $n_{\tilde{\phi}} \geqslant \max(-l_{\tilde{\phi}}, -l_a)$

and $n_{\tilde{\psi}} \geqslant \max(-l_{\tilde{\psi}}, \frac{n_{\tilde{\phi}}-l_{\tilde{b}}}{2})$. This proves item (2).

To prove item (3), since $(\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}, \mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ is a pair of biorthogonal Riesz bases in $L_2([0,\infty))$, noting that $\Phi(2\cdot) \perp \tilde{\Psi}(2^j\cdot)$ for all $j \geqslant 1$, for $\eta \in \Phi(2\cdot)$ we have

$$\eta = \langle \eta, \tilde{\phi}^L\rangle \mathring{\phi}^L + \sum_{k=n_{\tilde{\phi}}}^{\infty} \langle \eta, \tilde{\phi}(\cdot-k)\rangle \phi(\cdot-k) + \langle \eta, \tilde{\psi}^L\rangle \mathring{\psi}^L + \sum_{k=n_{\tilde{\psi}}}^{\infty} \langle \eta, \tilde{\psi}(\cdot-k)\rangle \psi(\cdot-k).$$

Since all functions in $\Phi \cup \Psi \cup \tilde{\Phi} \cup \tilde{\Psi}$ have compact support and $\Phi \cup \Psi$ is a Riesz sequence, we conclude from the above identity that item (3) holds. □

Proof of Theorem 5.7. By assumption $\phi^L \cup \psi^L \subseteq H^\tau(\mathbb{R})$ for some $\tau > 0$, since $\mathsf{AS}_0(\phi;\psi)$ is a Bessel sequence in $L_2(\mathbb{R})$, ψ must have at least one vanishing moment and we conclude from Theorem 5.6 that $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ is a Bessel sequence in $L_2([0,\infty))$. Similarly, by $\tilde{\phi}^L \cup \tilde{\psi}^L \subseteq H^\tau(\mathbb{R})$ for some $\tau > 0$, we conclude from Theorem 5.6 that $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ is a Bessel sequence in $L_2([0,\infty))$. Now the rest of the argument is quite standard for proving that $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ and $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ form a pair of biorthogonal Riesz bases in $L_2([0,\infty))$. By scaling, it suffices to prove the claim for $J = 0$. Define S_{j_0} as in (5.104) for $j_0 \in \mathbb{N}$. Using items (i)–(iv) and the same argument as in the proof of Theorem 5.2, we see that every element in S_{j_0} is a finite linear combination of $\Phi(2^{j_0}\cdot) := \{\phi^L(2^{j_0}\cdot)\} \cup \{\phi(2^{j_0}\cdot - k) \ : \ k \geqslant n_\phi\}$. Now by the biorthogonality between $\Phi \cup \Psi$ and $\tilde{\Phi} \cup \tilde{\Psi}$, it follows from the same argument as in the proof of Theorem 5.2 that $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ and $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ must be biorthogonal to each other in $L_2([0,\infty))$. Consequently, by the standard argument (e.g., see the proof of (4)$\Longrightarrow$(1) in Theorem 5.5), we conclude that both $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ and $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ are Riesz sequences in $L_2([0,\infty))$. By item (iii), we have $\mathsf{S}_0(\Psi) = \mathsf{S}_1(\Phi) \cap (\mathsf{S}_0(\tilde{\Phi}))^\perp$. For any $f \in \mathsf{S}_1(\Phi)$, define $g := \sum_{\eta\in\Phi}\langle f, \tilde{\eta}\rangle \eta$. Then $f - g \perp \tilde{\Phi}$ and we conclude that $f = g + (f-g)$ such that $g \in \mathsf{S}_0(\Phi)$ and $f - g \in \mathsf{S}_1(\Phi) \cap (\mathsf{S}_0(\tilde{\Phi}))^\perp = \mathsf{S}_0(\Psi)$. This proves $\mathsf{S}_1(\Phi) \subseteq \mathsf{S}_0(\Phi \cup \Psi)$. Because $\mathsf{S}_0(\Phi \cup \Psi) \subseteq \mathsf{S}_1(\Phi)$ is trivial, we conclude that $\mathsf{S}_0(\Phi \cup \Psi) = \mathsf{S}_1(\Phi)$. By the scaling argument, we must have

$$\mathsf{S}_0(\Phi \cup \Psi \cup \Psi(2\cdot) \cup \cdots \cup \Psi(2^{j-1}\cdot)) = \mathsf{S}_1(\Phi \cup \Psi \cup \cdots \cup \Psi(2^{j-2}\cdot)) = \cdots = \mathsf{S}_j(\Phi).$$

Hence, $\mathsf{S}_0(\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)})$ contains $\cup_{j=1}^\infty \mathsf{S}_j(\Phi)$, which includes the subset $\cup_{j=1}^\infty \{\phi(2^j\cdot - k) \ : \ k \geqslant n_\phi\}$, whose linear span is dense in $L_2([0,\infty))$ due to $\lim_{j_0\to\infty} 2^{-j_0} n_\phi = 0$. Therefore, the linear span of $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is dense in $L_2([0,\infty))$. This proves that $\mathsf{AS}_0(\Phi;\Psi)_{[0,\infty)}$ is a Riesz basis of $L_2([0,\infty))$. Similarly, $\mathsf{AS}_0(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ is also a Riesz basis of $L_2([0,\infty))$. This completes the proof of Theorem 5.7. □

Proof of Theorem 5.10. Define $\phi_j := \phi(\cdot - j)\chi_{[0,1]}$ for $j = 1-h_\phi, \ldots, -l_\phi$ (for other $j \in \mathbb{Z}$, ϕ_j

is identically zero). By $\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot -k)$, for $j \in \mathbb{Z}$, we have $\phi(\cdot - j) = 2\sum_{k\in\mathbb{Z}} a(k-2j)\phi(2\cdot -k)$. Multiplying $\chi_{[0,1]}$ on both sides of this identity, we particularly have

$$\phi_j(x) = \phi(x-j)\chi_{[0,1]}(x) = 2\sum_{k\in\mathbb{Z}} a(k-2j)\phi(2x-k)\chi_{[0,1]}(x).$$

Note that

$$\phi(2x-k)\chi_{[0,1/2]}(x) = \phi(2x-k)\chi_{[0,1]}(2x) = \Big[\phi(\cdot-k)\chi_{[0,1]}(\cdot)\Big](2x) = \phi_k(2x),$$
$$\phi(2x-k)\chi_{[1/2,1]}(x) = \phi(2x-k)\chi_{[0,1]}(2x-1) = \Big[\phi(\cdot+1-k)\chi_{[0,1]}(\cdot)\Big](2x-1) = \phi_{k-1}(2x-1).$$

Hence, we have

$$\begin{aligned}\phi_j(x) &= 2\sum_{k\in\mathbb{Z}} a(k-2j)\phi(2x-k)\chi_{[0,1]}(x) = 2\sum_{k\in\mathbb{Z}} a(k-2j)[\phi_k(2x)+\phi_{k-1}(2x-1)]\\ &= 2\sum_{k=1-h_\phi}^{-l_\phi} a(k-2j)\phi_k(2x) + 2\sum_{k=1-h_\phi}^{-l_\phi} a(k+1-2j)\phi_k(2x-1),\end{aligned}$$

since all ϕ_k for $k \in \mathbb{Z}\backslash[1-h_\phi, -l_\phi]$ are identically zero. By $\vec{\phi} = (\phi_{1-h_\phi}, \ldots, \phi_{-l_\phi})^{\mathsf{T}}$, this proves the first identity in (5.44). The proof of the second identity in (5.44) is similar.

We now prove (5.45). Noting that $\vec{\phi}(2x)$ is supported inside $[0, 1/2]$ and $\vec{\phi}(2x-1)$ is supported inside $[1/2, 1]$, we deduce from (5.44) that

$$\begin{aligned}M &= \int_0^1 \vec{\phi}(x)\overline{\vec{\phi}(x)}^{\mathsf{T}}\,dx\\ &= 4\int_0^1 \Big(\tilde{A}_0\vec{\phi}(2x) + \tilde{A}_1\vec{\phi}(2x-1)\Big)\Big(\overline{\vec{\phi}(2x)}^{\mathsf{T}}\overline{A_0}^{\mathsf{T}} + \overline{\vec{\phi}(2x-1)}^{\mathsf{T}}\overline{A_1}^{\mathsf{T}}\Big)dx\\ &= 4\tilde{A}_0\int_0^1 \vec{\phi}(2x)\overline{\vec{\phi}(2x)}^{\mathsf{T}}\,dx\overline{A_0}^{\mathsf{T}} + 4\tilde{A}_1\int_0^1 \vec{\phi}(2x-1)\overline{\vec{\phi}(2x-1)}^{\mathsf{T}}\,dx\overline{A_1}^{\mathsf{T}}\\ &= 2\tilde{A}_0\int_0^1 \vec{\phi}(x)\overline{\vec{\phi}(x)}^{\mathsf{T}}\,dx\overline{A_0}^{\mathsf{T}} + 2\tilde{A}_1\int_0^1 \vec{\phi}(x)\overline{\vec{\phi}(x)}^{\mathsf{T}}\,dx\overline{A_1}^{\mathsf{T}}\\ &= 2\tilde{A}_0M\overline{A_0}^{\mathsf{T}} + 2\tilde{A}_1M\overline{A_1}^{\mathsf{T}}.\end{aligned}$$

This proves (5.45). We now prove that up to a multiplicative constant (5.45) has a unique solution. By $\mathrm{vec}(M)$ we denote the column vector by arranging the columns of M one by one. Then (5.45) is equivalent to

$$T\mathrm{vec}(M) = \mathrm{vec}(M) \quad \text{and} \quad T := 2(\overline{A_0}\otimes\tilde{A}_0 + \overline{A_1}\otimes\tilde{A}_1), \tag{5.106}$$

where $\otimes$ stands for the right Kronecker product of matrices. To prove (S3), it suffices to prove that 1 is a simple eigenvalue of the matrix T in (5.106). Define $s := \#\vec{\phi}$ and $\tilde{s} := \#\vec{\tilde{\phi}}$. Note that $\vec{\phi}$ is a compactly supported refinable vector function in $L_2(\mathbb{R})$ and the integer shifts of $\vec{\phi}$ are linearly independent. By [71, Corollary 5.6.12 and Proposition 5.6.2], we conclude that 1 must be a simple eigenvalue of $A_0 + A_1$ and the mask/filter associated with $\vec{\phi}$ must have at least order one sum rule, that is, by (5.23), there must exist a nontrivial row vector $\vec{v} \in \mathbb{C}^{1\times s}$ such that $\vec{v}A_0 = \vec{v}A_1 = \frac{1}{2}\vec{v}$ and $\vec{v}\widehat{\vec{\phi}}(0) = 1$. Since 1 is a simple eigenvalue of $A_0 + A_1$, such row vector $\vec{v}$ must be unique by $\vec{v}(A_0 + A_1) = \vec{v}$. Similarly, there exists a unique row vector $\vec{\tilde{v}} \in \mathbb{C}^{1\times \tilde{s}}$ such that $\vec{\tilde{v}}\tilde{A}_0 = \vec{\tilde{v}}\tilde{A}_1 = \frac{1}{2}\vec{\tilde{v}}$ and $\vec{\tilde{v}}\widehat{\vec{\tilde{\phi}}}(0) = 1$. By $\vec{v}A_0 = \vec{v}A_1 = \frac{1}{2}\vec{v}$ and $\vec{\tilde{v}}\tilde{A}_0 = \vec{\tilde{v}}\tilde{A}_1 = \frac{1}{2}\vec{\tilde{v}}$, we trivially have

$$\begin{aligned}(\overline{\vec{v}} \otimes \overline{\vec{\tilde{v}}})T &= (\overline{\vec{v}} \otimes \overline{\vec{\tilde{v}}})2(\overline{A_0} \otimes \tilde{A}_0 + \overline{A_1} \otimes \tilde{A}_1) = 2(\overline{\vec{v}A_0}) \otimes (\vec{\tilde{v}}\tilde{A}_0) + 2(\overline{\vec{v}A_1}) \otimes (\vec{\tilde{v}}\tilde{A}_1)\\ &= \frac{1}{2}(\overline{\vec{v}} \otimes \overline{\vec{\tilde{v}}}) + \frac{1}{2}(\overline{\vec{v}} \otimes \overline{\vec{\tilde{v}}}) = \overline{\vec{v}} \otimes \overline{\vec{\tilde{v}}}.\end{aligned}$$

Hence 1 must be an eigenvalue of T. Next we prove that the eigenvalue 1 of T has multiplicity one by employing the joint spectral radius technique in [64]. Let U be the space of all column vectors $u \in \mathbb{C}^s$ such that $\vec{v}u = 0$. By $\vec{v}A_0 = \vec{v}A_1 = \frac{1}{2}\vec{v}$, it is trivial to observe that $A_0U \subseteq U$ and $A_1U \subseteq U$. Since all the entries in $\vec{\phi}$ are compactly supported functions in $L_2(\mathbb{R})$ and the integer shifts of $\vec{\phi}$ are linearly independent, we must have (e.g., see [71, Theorems 5.6.11 and 5.7.4]. Also c.f. [64, Theorem 3.3]) that

$$\lim_{n\to\infty} 2^{n/2}\|\{A_0, A_1\}^n u\|_{l_2} = 0, \qquad \forall\, u \in U \tag{5.107}$$

and for every $w \in \mathbb{C}^s$, there exists a positive constant C_w such that

$$2^{n/2}\|\{A_0, A_1\}^n w\|_{l_2} \leqslant C_w, \qquad \forall\, n \in \mathbb{N}, \tag{5.108}$$

where as in [64, Section 2] we define

$$\|\{A_0, A_1\}^n u\|_{l_2}^2 := \sum_{\gamma_1=0}^{1} \cdots \sum_{\gamma_n=0}^{1} \|A_{\gamma_1} \cdots A_{\gamma_n} u\|^2.$$

Similar conclusions hold for $\tilde{A}_0$ and $\tilde{A}_1$. Take particular vectors $w := \widehat{\vec{\phi}}(0)$ and $\tilde{w} := \widehat{\vec{\tilde{\phi}}}(0)$. Hence, we must have $\vec{v}w = 1$ and $\vec{\tilde{v}}\tilde{w} = 1$. Now considering $T^n(\overline{u} \otimes \tilde{w})$ with $u \in U$ and using

the Cauchy-Schwarz inequality, we conclude that

$$\|T^n(\overline{u}\otimes\tilde{w})\| \leqslant 2^n\|\{A_0,A_1\}u\|_{l_2}\|\{\tilde{A}_0,\tilde{A}_1\}\tilde{w}\|_{l_2} \leqslant C_{\tilde{w}}2^{n/2}\|\{A_0,A_1\}u\|_{l_2}\to 0$$

as $n\to\infty$. Similarly, for $\tilde{u}\in\tilde{U}$, as $n\to\infty$, we have

$$\|T^n(\overline{w}\otimes\tilde{u})\| \leqslant 2^n\|\{A_0,A_1\}w\|_{l_2}\|\{\tilde{A}_0,\tilde{A}_1\}\tilde{u}\|_{l_2} \leqslant C_{w}2^{n/2}\|\{\tilde{A}_0,\tilde{A}_1\}\tilde{u}\|_{l_2}\to 0.$$

Also, for all $u\in U$ and $\tilde{u}\in\tilde{U}$, we similarly have $\lim_{n\to\infty}\|T^n(\overline{u}\otimes\tilde{u})\|=0$. Note that w and U span the whole space $\mathbb{C}^s$ while $\tilde{w}$ and $\tilde{U}$ span $\mathbb{C}^{\tilde{s}}$, where $s:=\#\vec{\phi}$ and $\tilde{s}:=\#\vec{\tilde{\phi}}$. The above three identities prove that all the other eigenvalues of T must be less than one in modulus. Hence, 1 is a simple eigenvalue of T. Hence, up to a multiplicative constant, M is the unique solution to (5.45).

Note that $\sum_{k\in\mathbb{Z}}\vec{v}\vec{\phi}(\cdot-k)=1$ and $\sum_{k\in\mathbb{Z}}\vec{\tilde{v}}\vec{\tilde{\phi}}(\cdot-k)=1$. Since both $\vec{\phi}$ and $\vec{\tilde{\phi}}$ are supported inside $[0,1]$, we must have $\vec{v}\vec{\phi}(x)=1$ and $\vec{\tilde{v}}\vec{\tilde{\phi}}(x)=1$ for almost every $x\in[0,1]$ and hence

$$\vec{\tilde{v}}M\overline{\vec{v}}^{\mathsf{T}}=\int_0^1\vec{\tilde{v}}\vec{\tilde{\phi}}(x)\overline{\vec{v}\vec{\phi}(x)}^{\mathsf{T}}dx=\int_0^1 1dx=1.$$

This proves (5.47) and completes the proof. □

Proof of Theorem 5.14. By $m_\phi=\max(2n_\phi+h_{\tilde{a}},2n_\psi+h_{\tilde{b}})$, we have $m_\phi\geqslant 2n_\phi+h_{\tilde{a}}\geqslant n_\phi$ since $n_\phi\geqslant -l_a\geqslant -h_{\tilde{a}}$. By the definition of $n_{\tilde{\phi}}$ and $n_{\tilde{\psi}}$, we trivially have $n_{\tilde{\phi}}\geqslant\max(-l_{\tilde{\phi}},-l_{\tilde{a}})$ and $n_{\tilde{\psi}}\geqslant\max(-l_{\tilde{\psi}},\frac{n_{\tilde{\phi}}-l_{\tilde{b}}}{2})$. Therefore, (5.21) and (5.22) must hold. We now prove (5.58). By the definition of $n_{\tilde{\phi}}$ and $n_{\tilde{\psi}}$, we also have $n_{\tilde{\phi}}\geqslant 1-l_{\tilde{a}}$ and $n_{\tilde{\psi}}\geqslant\lceil\frac{n_{\tilde{\phi}}-l_{\tilde{b}}+1}{2}\rceil$. Hence, we have $\frac{k_0-l_{\tilde{a}}}{2}\leqslant n_{\tilde{\phi}}-1$ and $\frac{k_0-l_{\tilde{b}}}{2}\leqslant n_{\tilde{\psi}}-1$ for all $k_0<n_{\tilde{\phi}}$. Therefore, for all $k_0\in\mathbb{Z}$ satisfying $m_\phi\leqslant k_0<n_{\tilde{\phi}}$, it follows from (5.42) and Lemma 5.8 that (5.58) must hold.

Define (infinite) column vector functions by

$$\vec{\phi}:=\{\phi(\cdot-k)\ :\ k\geqslant n_{\tilde{\phi}}\}\quad\text{and}\quad\vec{\psi}:=\{\psi(\cdot-k)\ :\ k\geqslant n_{\tilde{\psi}}\}.$$

Abusing notations a little bit by using the same notations for augmented A_L,B_L and A,B with ϕ^L,ψ^L being replaced by $\mathring{\phi}^L,\mathring{\psi}^L$, respectively, we can equivalently rewrite (5.13), (5.14), (5.17) and (5.18) as

$$\begin{bmatrix}\mathring{\phi}^L\\ \vec{\phi}\\ \mathring{\psi}^L\\ \vec{\psi}\end{bmatrix}=2\mathcal{M}\begin{bmatrix}\mathring{\phi}^L(2\cdot)\\ \vec{\phi}(2\cdot)\end{bmatrix}\quad\text{with}\quad\mathcal{M}:=\begin{bmatrix}A_L & M_A\\ 0 & M_a\\ B_L & M_B\\ 0 & M_b\end{bmatrix},\tag{5.109}$$

where M_A, M_B, M_a, M_b are matrices associated with filters A, B, a, b, respectively. More precisely, using (5.17), (5.18), (5.13) and (5.14), we have $M_A := [A(k)]_{n_{\tilde{\phi}} \leqslant k < \infty}$ (which is equivalent to $M_A \vec{\phi} = \sum_{k=n_{\tilde{\phi}}}^{\infty} A(k)\phi(\cdot - k)$), $M_B := [B(k)]_{n_{\tilde{\phi}} \leqslant k < \infty}$, and

$$M_a := [a(k-2k_0)]_{n_{\tilde{\phi}} \leqslant k_0 < \infty, n_{\tilde{\phi}} \leqslant k < \infty} \quad \text{and} \quad M_b := [b(k-2k_0)]_{n_{\tilde{\psi}} \leqslant k_0 < \infty, n_{\tilde{\phi}} \leqslant k < \infty}, \tag{5.110}$$

where k_0 is row index and k is column index. By $n_{\tilde{\phi}} \geqslant m_\phi$, we see from Lemma 5.8 and (5.42) that

$$\vec{\phi}(2\cdot) = \overline{M_{\tilde{A}}}^{\mathsf{T}} \mathring{\phi}^L + \overline{M_{\tilde{a}}}^{\mathsf{T}} \vec{\phi} + \overline{M_{\tilde{B}}}^{\mathsf{T}} \mathring{\psi}^L + \overline{M_{\tilde{b}}}^{\mathsf{T}} \vec{\psi}, \tag{5.111}$$

where $M_{\tilde{A}}, M_{\tilde{B}}, M_{\tilde{a}}, M_{\tilde{b}}$ are matrices uniquely determined by the filters $\tilde{a}$ and $\tilde{b}$. More precisely,

$$M_{\tilde{A}} := \begin{bmatrix} 0_{(\#\phi^L)\times\infty} \\ (\tilde{a}(k-2k_0))_{n_\phi \leqslant k_0 < n_{\tilde{\phi}}, n_{\tilde{\phi}} \leqslant k < \infty} \end{bmatrix}, \qquad M_{\tilde{B}} := \begin{bmatrix} 0_{(\#\psi^L)\times\infty} \\ (\tilde{b}(k-2k_0))_{n_\psi \leqslant k_0 < n_{\tilde{\psi}}, n_{\tilde{\phi}} \leqslant k < \infty} \end{bmatrix}, \tag{5.112}$$

and $M_{\tilde{a}}$, $M_{\tilde{b}}$ are defined similarly as in (5.110) using $\tilde{a}$ and $\tilde{b}$ instead of a and b, where k_0 is row index and k is column index. Therefore, we deduce from (5.59) and the above identity in (5.111) that

$$\begin{bmatrix} \mathring{\phi}^L(2\cdot) \\ \vec{\phi}(2\cdot) \end{bmatrix} = \overline{\tilde{\mathcal{M}}}^{\mathsf{T}} \begin{bmatrix} \mathring{\phi}^L \\ \vec{\phi} \\ \mathring{\psi}^L \\ \vec{\psi} \end{bmatrix} \quad \text{with} \quad \tilde{\mathcal{M}} := \begin{bmatrix} \tilde{A}_L & M_{\tilde{A}} \\ 0 & M_{\tilde{a}} \\ \tilde{B}_L & M_{\tilde{B}} \\ 0 & M_{\tilde{b}} \end{bmatrix}. \tag{5.113}$$

By assumption, Φ is a Riesz sequence in $L_2([0,\infty))$ and hence linearly independent. By item (i), the elements in $\Phi \cup \Psi$ must be linearly independent. Consequently, we deduce from (5.109) and (5.113) that $\tilde{\mathcal{M}}\overline{\mathcal{M}}^{\mathsf{T}} = 2^{-1}I$ and $\overline{\mathcal{M}}^{\mathsf{T}}\tilde{\mathcal{M}} = 2^{-1}I$, where I here stands for the infinite identity matrix.

We now prove that $\tilde{A}$ and $\tilde{B}$ in (5.63) are finitely supported. For all $k \geqslant 2n_{\tilde{\phi}} + h_{\tilde{a}}$, we have $k - 2k_0 > h_{\tilde{a}}$ for all $k_0 = n_\phi, \ldots, n_{\tilde{\phi}} - 1$ and hence $\tilde{a}(k-2k_0) = 0$. So, $\{\tilde{A}(k)\}_{k=n_{\tilde{\phi}}}^{\infty}$ in (5.63) is finitely supported. Similarly, for all $k \geqslant 2n_{\tilde{\psi}} + h_{\tilde{b}}$, we have $k - 2k_0 > h_{\tilde{b}}$ for all $k_0 = n_\psi, \ldots, n_{\tilde{\psi}} - 1$ and hence $\tilde{b}(k-2k_0) = 0$. So, $\{\tilde{B}(k)\}_{k=n_{\tilde{\phi}}}^{\infty}$ in (5.63) is finitely supported. Since $\rho(\tilde{A}_L) < 2^{-1/2}$ in (5.60), we conclude from Theorem 5.13 that $\tilde{\phi}^L$ in (5.61) is a well-defined compactly supported vector function in $L_2([0,\infty)) \cap H^\tau(\mathbb{R})$ for some $\tau > 0$ and satisfies (5.19). Since $\tilde{B}$ is finitely supported, $\tilde{\psi}^L$ in (5.62) is a well-defined compactly supported vector function in $L_2([0,\infty)) \cap H^\tau(\mathbb{R})$ and satisfies (5.20).

We now prove that $\tilde{\Phi}$ must be biorthogonal to Φ. Define $\vec{\tilde{\phi}} := \{\tilde{\phi}(\cdot - k) \ : \ k \geqslant n_{\tilde{\phi}}\}$. By (5.112), we have $M_{\tilde{A}}\vec{\tilde{\phi}} = \sum_{k=n_{\tilde{\phi}}}^{\infty} \tilde{A}(k)\tilde{\phi}(\cdot - k)$. Since we assumed that Φ is a Riesz sequence,

we conclude from Theorem 5.5 that item (4) of Theorem 5.5 holds. If necessary, enlarging $n_{\tilde{\phi}}$, then we can assume that $\tilde{H} = \tilde{\eta}^L \cup \vec{\tilde{\phi}}$ in item (4) of Theorem 5.5 is biorthogonal to Φ with $\#\tilde{\eta}^L = \#\mathring{\phi}^L$. Define $f_0 := \mathring{\phi}^L$ and $\tilde{f}_0 := \tilde{\eta}^L$. For $n \in \mathbb{N}$, we can recursively define

$$f_n := 2A_L f_{n-1}(2\cdot) + g(2\cdot) \quad \text{and} \quad \tilde{f}_n := 2\tilde{A}_L \tilde{f}_{n-1}(2\cdot) + \tilde{g}(2\cdot), \qquad n \in \mathbb{N}, \tag{5.114}$$

where $g := 2\sum_{k=n_{\tilde{\phi}}}^{\infty} A(k)\phi(\cdot - k)$, $\tilde{g}$ is given in (5.61), and A_L and A are augmented version in (5.17). Let $F_n := f_n \cup \vec{\phi}$ and $\tilde{F}_n := \tilde{f}_n \cup \vec{\tilde{\phi}}$. By the choice of $\tilde{f}_0 = \tilde{\eta}^L$ and $f_0 = \mathring{\phi}^L$, we have $\tilde{F}_0 = \tilde{H}$ and $F_0 = \Phi$. Therefore, $\tilde{F}_0$ is biorthogonal to F_0 by Theorem 5.5. Suppose that $\tilde{F}_{n-1}$ is biorthogonal to F_{n-1} (induction hypothesis), i.e., $\langle \tilde{F}_{n-1}, F_{n-1}\rangle = I$. We now prove the claim for n. Note that

$$F_n = 2\mathcal{N}F_{n-1}(2\cdot), \quad \tilde{F}_n = 2\tilde{\mathcal{N}}\tilde{F}_{n-1}(2\cdot) \quad \text{with} \quad \mathcal{N} := \begin{bmatrix} A_L & M_A \\ 0 & M_a \end{bmatrix}, \quad \tilde{\mathcal{N}} := \begin{bmatrix} \tilde{A}_L & M_{\tilde{A}} \\ 0 & M_{\tilde{a}} \end{bmatrix}. \tag{5.115}$$

It follows trivially from the identity $\tilde{\mathcal{M}}\overline{\mathcal{M}}^{\mathsf{T}} = 2^{-1}I$ that $\tilde{\mathcal{N}}\overline{\mathcal{N}}^{\mathsf{T}} = 2^{-1}I$. Therefore, by induction hypothesis $\langle \tilde{F}_{n-1}, F_{n-1}\rangle = I$, we have

$$\langle \tilde{F}_n, F_n\rangle = 4\tilde{\mathcal{N}}\langle \tilde{F}_{n-1}(2\cdot), F_{n-1}(2\cdot)\rangle\overline{\mathcal{N}}^{\mathsf{T}} = 2\tilde{\mathcal{N}}\overline{\mathcal{N}}^{\mathsf{T}} = I.$$

This proves the claim for n. By mathematical induction, we proved that $\tilde{F}_n$ is biorthogonal to F_n for all $n \in \mathbb{N}$. By (5.19) and (5.114) with $f_0 = \mathring{\phi}^L$, we trivially have $f_n = \mathring{\phi}^L$ and hence, $F_n = \Phi$ for all $n \in \mathbb{N}$. So, $\tilde{F}_n$ is biorthogonal to Φ for all $n \in \mathbb{N}$. We deduce from the definition of $\tilde{f}_n$ in (5.114) that

$$\tilde{f}_n = 2^n \tilde{A}_L^n \tilde{f}_0(2^n\cdot) + \sum_{j=1}^{n} 2^{j-1}\tilde{A}_L^{j-1} g(2^j\cdot).$$

Since $\rho(\tilde{A}_L) < 2^{-1/2}$ and $\|2^n\tilde{A}_L^n\tilde{f}_0(2^n\cdot)\|_{L_2(\mathbb{R})} \leqslant \|\tilde{f}_0\|_{L_2(\mathbb{R})}2^{n/2}\|\tilde{A}_L^n\|$, we conclude that

$$\lim_{n\to\infty} \|2^n\tilde{A}_L^n\tilde{f}_0(2^n\cdot)\|_{L_2(\mathbb{R})} = 0.$$

This proves $\lim_{n\to\infty}\|\tilde{f}_n - \tilde{\phi}^L\|_{L_2(\mathbb{R})} = 0$. Since $\tilde{\Phi} = \lim_{n\to\infty}\tilde{F}_n$ in $L_2(\mathbb{R})$ and $\tilde{F}_n$ is biorthogonal to Φ, we conclude that $\tilde{\Phi}$ must be biorthogonal to Φ.

By (5.112), we have $\tilde{\psi}^L = 2\tilde{B}_L\tilde{\phi}^L(2\cdot) + 2M_{\tilde{B}}\vec{\tilde{\phi}}(2\cdot)$. Also note that $\tilde{\phi}^L = 2\tilde{A}_L\tilde{\phi}^L(2\cdot) + 2M_{\tilde{A}}\vec{\tilde{\phi}}(2\cdot)$, $\vec{\tilde{\phi}} = 2M_{\tilde{a}}\vec{\tilde{\phi}}(2\cdot)$, and $\vec{\tilde{\psi}} = 2M_{\tilde{b}}\vec{\tilde{\phi}}(2\cdot)$ with $\vec{\tilde{\psi}} := \{\tilde{\psi}(\cdot - k) \ : \ k \geqslant n_{\tilde{\psi}}\}$. Using the identities $\tilde{\mathcal{M}}\overline{\mathcal{M}}^{\mathsf{T}} = 2^{-1}I$ and $\overline{\mathcal{M}}^{\mathsf{T}}\tilde{\mathcal{M}} = 2^{-1}I$, we can now check that all items (i)–(iv) of Theorem 5.7 are satisfied. Because we assumed $\phi^L \subseteq H^\tau(\mathbb{R})$ for some $\tau > 0$, by the choice

of ψ^L in item (i), we must have $\psi^L \subseteq H^\tau(\mathbb{R})$. Note that we already proved $\tilde{\phi}^L \cup \tilde{\psi}^L \subseteq H^\tau(\mathbb{R})$ for some $\tau > 0$. Since all the conditions in Theorem 5.7 are satisfied, we conclude from Theorem 5.7 that $\mathsf{AS}_J(\tilde{\Phi};\tilde{\Psi})_{[0,\infty)}$ and $\mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)}$ form a pair of biorthogonal Riesz bases in $L_2([0,\infty))$ for every $J \in \mathbb{Z}$. □

Proof of Theorem 5.15. By the choice of $n_{\tilde{\phi}}$ in item (S1) of Algorithm 5.2, for all $k \geqslant n_{\tilde{\phi}}$, we have $\langle \tilde{\phi}(\cdot - k), \phi(\cdot - k)\rangle = I_r$ and $\langle \tilde{\phi}(\cdot - k), \eta\rangle = 0$ for all $\eta \in \Phi \backslash \{\phi(\cdot - k)\}$. Define $\vec{\phi} := \{\phi(\cdot - k) \,:\, k \geqslant n_{\tilde{\phi}}\}$ and $\vec{\tilde{\phi}} := \{\tilde{\phi}(\cdot - k) \,:\, k \geqslant n_{\tilde{\phi}}\}$. Then (5.17) and (5.19) become

$$\mathring{\phi}^L = 2A_L \mathring{\phi}^L(2\cdot) + 2M_A \vec{\phi}(2\cdot), \qquad \tilde{\phi}^L = 2\tilde{A}_L \tilde{\phi}^L(2\cdot) + 2M_{\tilde{A}} \vec{\tilde{\phi}}(2\cdot),$$

where $M_A := [A(k)]_{n_{\tilde{\phi}} \leqslant k < \infty}$ and $M_{\tilde{A}} := [\tilde{A}(k)]_{n_{\tilde{\phi}} \leqslant k < \infty}$ with k being the column index. Since $n_{\tilde{\phi}} \geqslant \max(-l_{\tilde{\phi}}, -l_{\tilde{a}})$ and $n_{\tilde{\phi}} \geqslant n_\phi \geqslant \max(-l_\phi, -l_a)$, by (5.13) and (5.21), we must have $\vec{\phi} = 2M_a \vec{\phi}(2\cdot)$ and $\vec{\tilde{\phi}} = 2M_{\tilde{a}} \vec{\tilde{\phi}}(2\cdot)$, where M_a is defined in (5.110) and $M_{\tilde{a}}$ is defined similarly. Define $\mathcal{N}$ and $\tilde{\mathcal{N}}$ as in (5.115). Using (5.7) and (5.64), we must have $\tilde{\mathcal{N}}\overline{\mathcal{N}}^{\mathsf{T}} = 2^{-1} I$. Since Φ is a Riesz sequence, by the same argument as in the proof of Theorem 5.14, we conclude that $\tilde{\Phi}$ is biorthogonal to Φ and $\tilde{\Phi}$ satisfies item (ii) of Theorem 5.7. □

Proof of Theorem 5.19. Let $j \geqslant J_0$. By assumption in (S3), all the boundary elements in $\Phi_j \cup \Psi_j$ belong to $L_2([0,N])$. Since $\operatorname{fsupp}(\phi(\cdot - k)) \subseteq [0,\infty)$ for all $k \geqslant n_\phi$, to show $\Phi_j \subseteq L_2([0,N])$, it suffices to prove that $\operatorname{fsupp}(\phi(\cdot - k)) \subseteq (-\infty, 2^j N]$ for all $k \leqslant 2^j N - n_{\mathring{\phi}}$, which is equivalent to $\phi(2^j N - \cdot - k) \in L_2([0,\infty))$. For all $k \leqslant 2^j N - n_{\mathring{\phi}}$, we note that $\phi(2^j N - \cdot - k) = \mathring{\phi}(\cdot - (2^j N - k))$ and $2^j N - k \geqslant n_{\mathring{\phi}}$. By the definition of $n_{\mathring{\phi}}$, we must have

$$\phi(2^j N - \cdot - k) = \mathring{\phi}(\cdot - (2^j N - k)) \subseteq L_2([0,\infty)).$$

This proves $\phi_{j;k} \in L_2([0,N])$ for all $n_\phi \leqslant k \leqslant 2^j N - n_{\mathring{\phi}}$. Consequently, we proved $\Phi_j \subseteq L_2([0,N])$ for all $j \geqslant J_0$. Similarly, we have $\Psi_j \subseteq L_2([0,N])$ for all $j \geqslant J_0$.

We now prove item (1). By (5.18) and $h_B + n_{\mathring{\phi}} \leqslant 2^{j+1}N$ in (5.85), we have $h_B \leqslant 2^{j+1}N - n_{\mathring{\phi}}$ and

$$\psi^L_{j;0} = \sqrt{2} B_L \phi^L_{j+1;0} + \sqrt{2} \sum_{k=n_\phi}^{h_B} B(k) \phi_{j+1;k} = \sqrt{2} B_L \phi^L_{j+1;0} + \sqrt{2} \sum_{k=n_\phi}^{2^{j+1}N - n_{\mathring{\phi}}} B(k) \phi_{j+1;k}$$

with $B(h_B + 1) = \cdots = B(2^{j+1}N - n_{\mathring{\phi}}) = 0$ due to $[l_B, h_B] = \operatorname{fsupp}(B)$ and $2^{j+1}N \geqslant h_B + n_{\mathring{\phi}}$ in (5.85) for all $j \geqslant J_0$. By $n_{\mathring{\psi}} \geqslant \max(-l_{\mathring{\psi}}, \frac{n_{\mathring{\phi}} - l_{\mathring{b}}}{2})$, we have $l_{\mathring{b}} + 2n_{\mathring{\psi}} \geqslant n_{\mathring{\phi}}$. Since $\mathring{b} = b(-\cdot)$, we have $l_{\mathring{b}} = -h_b$ and hence, we proved $h_b - 2n_{\mathring{\psi}} \leqslant -n_{\mathring{\phi}}$, from which we get

$h_b + 2(2^j N - n_{\mathring{\psi}}) \leqslant 2^{j+1}N - n_{\mathring{\phi}}$. By $n_\psi \geqslant \max(-l_\psi, \frac{n_\phi - l_b}{2})$, we have $l_b + 2n_\psi \geqslant n_\phi$. So, for every $k = n_\psi, \dots, 2^j N - n_{\mathring{\psi}}$, we have $n_\phi \leqslant l_b + 2k \leqslant h_b + 2k \leqslant 2^{j+1}N - n_{\mathring{\phi}}$ and thus we deduce from $\psi = 2\sum_{n=l_b}^{h_b} b(n)\phi(2\cdot - n)$ that

$$\psi_{j;k} = \sqrt{2} \sum_{n=l_b+2k}^{h_b+2k} b(n-2k)\phi_{j+1;n} = \sqrt{2} \sum_{n=n_\phi}^{2^{j+1}N - n_{\mathring{\phi}}} b(n-2k)\phi_{j+1;n}, \qquad k = n_\psi, \dots, 2^j N - n_{\mathring{\psi}}.$$

By (5.18) for $\mathring{\psi}$ and $h_{\mathring{B}} + n_\phi \leqslant 2^{j+1}N$ in (5.85), noting that $\psi^R = \psi^L(N - \cdot)$ and $\mathring{\phi} = \phi(-\cdot)$, we have $2^{j+1}N - h_{\mathring{B}} \geqslant n_\phi$ and

$$\begin{aligned}
\psi^R_{j;2^j N - N} &= \mathring{\psi}^L_{j;0}(N - \cdot) \\
&= \sqrt{2}\mathring{B}_L \mathring{\phi}^L_{j+1;0}(N - \cdot) + \sqrt{2} \sum_{k=n_{\mathring{\phi}}}^{h_{\mathring{B}}} \mathring{B}(k) \mathring{\phi}_{j+1;k}(N - \cdot) \\
&= \sqrt{2}\mathring{B}_L \phi^R_{j+1;2^{j+1}N - N} + \sqrt{2} \sum_{k=n_{\mathring{\phi}}}^{h_{\mathring{B}}} \mathring{B}(k) \phi_{j+1;2^{j+1}N-k} \\
&= \sqrt{2}\mathring{B}_L \phi^R_{j+1;2^{j+1}N - N} + \sqrt{2} \sum_{k=n_\phi}^{2^{j+1}N - n_{\mathring{\phi}}} \mathring{B}(2^{j+1}N - k) \phi_{j+1;k},
\end{aligned}$$

where we used $2^{j+1}N - h_{\mathring{B}} \geqslant n_\phi$ due to our assumption $h_{\mathring{B}} + n_\phi \leqslant 2^{j+1}N$ in (5.85) for $j \geqslant J_0$. Hence, we proved the existence of a matrix B_j such that $\Psi_j = B_j \Phi_{j+1}$. The existence of a matrix A_j can be proved similarly by the same argument. Similarly we can prove the first part of item (2).

Due to item (S1), by item (ii) of Theorem 5.7 or (5.16) in Theorem 5.2, we must have $\#\tilde{\phi}^L - \#\phi^L = (n_{\tilde{\phi}} - n_\phi)(\#\phi)$ and $\#\tilde{\mathring{\phi}}^L - \#\mathring{\phi}^L = (n_{\tilde{\mathring{\phi}}} - n_{\mathring{\phi}})(\#\phi)$, from which we have

$$\#\tilde{\phi}^L + \#\tilde{\mathring{\phi}}^L - (n_{\tilde{\mathring{\phi}}} + n_{\tilde{\phi}})(\#\phi) = (\#\phi^L + \#\mathring{\phi}^L) - (n_{\mathring{\phi}} + n_\phi)(\#\phi).$$

By (5.89), we have $\#\tilde{\phi}^R = \#\tilde{\mathring{\phi}}^L$ and $\#\phi^R = \#\mathring{\phi}^L$. Note that $2^j N \geqslant n_\phi + n_{\mathring{\phi}} - 1$ for all $j \geqslant J_0$ by (5.85) and $2^j N \geqslant n_{\tilde{\phi}} + n_{\tilde{\mathring{\phi}}} - 1$ for all $j \geqslant \tilde{J}_0$ by (5.87). Consequently,

$$\#\Phi_j = \#\phi^L + \#\mathring{\phi}^L + (2^j N - n_{\mathring{\phi}} - n_\phi + 1)(\#\phi), \qquad j \geqslant J_0. \tag{5.116}$$

Using the above two identities and $\tilde{J}_0 \geqslant J_0$, for $j \geqslant \tilde{J}_0$, we deduce that

$$\#\tilde{\Phi}_j = \#\tilde{\phi}^L + \#\tilde{\mathring{\phi}}^L + (2^j N - n_{\tilde{\mathring{\phi}}} - n_{\tilde{\phi}} + 1)(\#\phi) = \#\phi^L + \#\mathring{\phi}^L + (2^j N - n_{\mathring{\phi}} - n_\phi + 1)(\#\phi) = \#\Phi_j.$$

By Theorem 5.7, we must have $\#\tilde{\psi}^L - \#\psi^L = (n_{\tilde{\psi}} - n_\psi)(\#\psi)$ and $\#\mathring{\tilde{\psi}}^L - \#\mathring{\psi}^L = (n_{\mathring{\tilde{\psi}}} - n_{\mathring{\psi}})(\#\psi)$. By the same argument, we must have $\#\tilde{\Psi}_j = \#\Psi_j$. By the proved identities $\#\tilde{\Phi}_j = \#\Phi_j$ and $\#\tilde{\Psi}_j = \#\Psi_j$ for $j \geqslant \tilde{J}_0$, we observe from (5.88) that $\tilde{\Phi}_j \cup \tilde{\Psi}_j$ is biorthogonal to $\Phi_j \cup \Psi_j$ for all $j \geqslant \tilde{J}_0$.

Since $\tilde{\Phi}_j \cup \tilde{\Psi}_j$ is biorthogonal to $\Phi_j \cup \Psi_j$, the proved identities $\Phi_j = A_j \Phi_{j+1}$ and $\Psi_j = B_j \Phi_{j+1}$ imply $\#\Phi_j + \#\Psi_j \leqslant \#\Phi_{j+1}$ and $\Phi_j \cup \Psi_j$ is a Riesz sequence. To prove the other direction, by (5.42),

$$\phi_{j+1;m} = \sqrt{2} \sum_{k=\lceil \frac{m-h_{\tilde{a}}}{2} \rceil}^{\lfloor \frac{m-l_{\tilde{a}}}{2} \rfloor} \overline{\tilde{a}(m-2k)}^{\mathsf{T}} \phi_{j;k} + \sqrt{2} \sum_{k=\lceil \frac{m-h_{\tilde{b}}}{2} \rceil}^{\lfloor \frac{m-l_{\tilde{b}}}{2} \rfloor} \overline{\tilde{b}(m-2k)}^{\mathsf{T}} \psi_{j;k}.$$

Define $m_1 := \max(2n_\phi + h_{\tilde{a}}, 2n_\psi + h_{\tilde{b}})$ and $m_2 := \max(2n_{\mathring{\phi}} - l_{\tilde{a}}, 2n_{\mathring{\psi}} - l_{\tilde{b}})$. Using (5.8), we conclude from the above identity that $\phi_{j+1;m} \in \operatorname{span}(\Phi_j \cup \Psi_j)$ for all $m_1 \leqslant m \leqslant 2^{j+1}N - m_2$. On the other hand, for sufficiently large j, it follows directly from item (3) of Theorem 5.2 that

$$\phi^L_{j+1;0} \cup \{\phi_{j+1;k}\}_{k=n_\phi}^{m_1-1} \subseteq \operatorname{span}(\Phi_j \cup \Psi_j)$$

and

$$\phi^R_{j+1;2^jN-N} \cup \{\phi_{j+1;k}\}_{k=2^{j+1}N-m_2+1}^{2^{j+1}N-n_{\mathring{\phi}}} \subseteq \operatorname{span}(\Phi_j \cup \Psi_j).$$

This proves that $\Phi_{j+1} \subseteq \operatorname{span}(\Phi_j \cup \Psi_j)$ for sufficiently large j. Since $\Phi_j \cup \Psi_j$ is a Riesz sequence, we conclude from $\Phi_{j+1} \subseteq \operatorname{span}(\Phi_j \cup \Psi_j)$ that $\#\Phi_{j+1} \leqslant \#\Phi_j + \#\Psi_j$. Hence, we proved $\#\Phi_{j+1} = \#\Phi_j + \#\Psi_j$ and we deduce from (5.116) that $\#\Psi_j = \#\Phi_{j+1} - \#\Phi_j = 2^jN(\#\phi)$ for sufficiently large j. Note that $2^jN \geqslant n_\psi + n_{\mathring{\psi}} - 1$ for all $j \geqslant J_0$ by (5.85) and $2^jN \geqslant n_{\tilde{\psi}} + n_{\mathring{\tilde{\psi}}} - 1$ for all $j \geqslant \tilde{J}_0$ by (5.87). From the definition of Ψ_j, noting that $\#\psi = \#\phi$ (see item (3) of Theorem 5.1) and $\#\Psi_j = 2^jN(\#\phi)$, we have

$$\#\psi^L + \#\psi^R = \#\Psi_j - (2^jN - n_\psi - n_{\mathring{\psi}} + 1)(\#\psi) = (n_\psi + n_{\mathring{\psi}} - 1)(\#\phi). \tag{5.117}$$

Now for any arbitrary $j \geqslant J_0$, by definition of Ψ_j in (5.91) and the above identity, we have

$$\begin{aligned}
\#\Psi_j &= \#\psi^L + \#\psi^R + (2^jN - n_{\mathring{\psi}} - n_\psi + 1)(\#\phi) \\
&= (n_\psi + n_{\mathring{\psi}} - 1)(\#\phi) + (2^jN - n_{\mathring{\psi}} - n_\psi + 1)(\#\phi) \\
&= 2^jN(\#\phi).
\end{aligned}$$

Consequently, $\#\Psi_j = 2^jN(\#\phi) = \#\Phi_{j+1} - \#\Phi_j$ and hence $\#\Phi_{j+1} = \#\Phi_j + \#\Psi_j$ for all $j \geqslant J_0$. This proves both items (1) and (2).

Next we prove item (3). By proved items (1) and (2), $[\overline{A_j}^{\mathsf{T}}, \overline{B_j}^{\mathsf{T}}]$ must be a square matrix for all $j \geqslant J_0$ and $[\tilde{A}_j^{\mathsf{T}}, \tilde{B}_j^{\mathsf{T}}]$ must be a square matrix for all $j \geqslant \tilde{J}_0$. Since $\tilde{\Phi}_j \cup \tilde{\Psi}_j$ is biorthogonal to $\Phi_j \cup \Psi_j$, we must have (5.94). Now by items (1) and (2), we can directly check that $\mathcal{B}_J$ and $\tilde{\mathcal{B}}_J$ are biorthogonal to each other for all $J \geqslant \tilde{J}_0$. Note that $\mathcal{B}_J$ is a Bessel sequence, since $\mathcal{B}_J \subseteq \mathsf{AS}_J(\Phi;\Psi)_{[0,\infty)} \cup \{\eta(N-\cdot) : \eta \in \mathsf{AS}_J(\mathring{\Phi};\mathring{\Psi})_{[0,\infty)}\}$. By a similar reasoning, $\tilde{\mathcal{B}}_J$ is also a Bessel sequence. By the same standard argument as in the proof of (4)$\Longrightarrow$(1) in Theorem 5.5, both $\mathcal{B}_J$ and $\tilde{\mathcal{B}}_J$ are Riesz sequences in $L_2([0,N])$ for all $J \geqslant \tilde{J}_0$. By the proved item (1), we have $\mathrm{span}(\mathcal{B}_J) \supset \cup_{j=J}^{\infty}\{\phi_{j;k} \ : \ n_\phi \leqslant k \leqslant 2^j N - n_{\mathring{\phi}}\}$, which spans a dense subset of $L_2([0,N])$. That is, $\mathrm{span}(\mathcal{B}_J)$ is dense in $L_2([0,N])$ and hence, $\mathcal{B}_J$ must be a Riesz basis of $L_2([0,N])$. Similarly, we can prove that $\tilde{\mathcal{B}}_J$ is a Riesz basis of $L_2([0,N])$ for all $J \geqslant \tilde{J}_0$. Since $\mathcal{B}_J$ and $\tilde{\mathcal{B}}_J$ are biorthogonal to each other, this proves that $\tilde{\mathcal{B}}_J$ and $\mathcal{B}_J$ form a pair of biorthogonal Riesz bases for $L_2([0,N])$ for all $J \geqslant \tilde{J}_0$. This proves item (3).

Using item (3), item (4) can be easily proved by the same argument as in Lemma 5.3.

By item (3), for $J_0 \leqslant J < \tilde{J}_0$ such that J decreases from $\tilde{J}_0 - 1$ to J_0, using (5.94) and the biorthogonality relation between $\mathcal{B}_J$ and $\tilde{\mathcal{B}}_J$, we can recursively prove that $\tilde{\mathcal{B}}_J$ and $\mathcal{B}_J$ form a pair of biorthogonal Riesz bases for $L_2([0,N])$. This proves item (5). □

Chapter 6

A Wavelet Galerkin Method for an Electromagnetic Scattering Problem

Now that we have discussed in detail the construction of wavelets on a bounded interval in Chapter 5, we are ready to present our wavelet Galerkin method for solving an electromagnetic scattering from a large cavity problem. As we shall soon see, this method falls in the category of high order schemes, which is equipped with a natural preconditioner stemming from the wavelet basis. The model problem of this chapter is as follows

$$\begin{aligned} &\Delta u + \kappa^2 u = f \quad \text{in} \quad \Omega, \\ &u = 0 \quad \text{on} \quad \partial\Omega\backslash\Gamma, \\ &\frac{\partial u}{\partial \boldsymbol{\nu}} = \mathcal{T}(u) + g \quad \text{on} \quad \Gamma, \end{aligned} \tag{6.1}$$

where $\kappa > 0$ is a constant wavenumber, $\Omega := (0,1)^2$, $\Gamma := (0,1) \times \{1\}$, $f \in L_2(\Omega)$, $g \in H^{1/2}(\Gamma)$, $\boldsymbol{\nu}$ is the unit outward normal,

$$\mathcal{T}(u) := \frac{\mathrm{i}\kappa}{2} \fint_0^1 \frac{1}{|x - x'|} H_1^{(1)}(\kappa|x - x'|) u(x', 1) dx', \tag{6.2}$$

$\fint$ denotes the Hadamard finite part integral, and $H_1^{(1)}$ is the Hankel function of the first kind of degree 1. We briefly discuss the derivation of the model problem in Section 6.1. The implementation of our wavelet Galerkin scheme is discussed in Section 6.2. Finally, we present some numerical experiments in Section 6.3.

6.1 Derivation of the model problem

In practice, such a scattering problem is often encountered in stealth/tracking technology. The model derivation in this section closely follows the discussion in [42]. The Radar Cross Section (RCS) measures the detectability of an object by a radar system. The RCS of cavities in an object (e.g., a jet engine's inlet ducts, exhaust nozzles) contributes the most to the overall RCS of an object. Therefore, accurate measurements of the RCS of these cavities are important. This is where numerical methods for the scattering problem come into play.

In order to derive the model problem, we introduce several simplifying physical assumptions. We assume that the cavity is embedded in an infinite ground plane. The ground plane and cavity walls are perfect electric conductors (PECs). The medium is non-magnetic with a constant permeability, μ, and a constant permittivity, ε. Furthermore, we assume that no currents are present and the fields are source free. Let E and H respectively denote the total electric and magnetic fields. So far, our current setup can be modelled by the following Maxwell's equation with time dependence $e^{-\mathrm{i}\omega t}$, where ω stands for the angular frequency

$$\begin{aligned} \nabla \times E - \mathrm{i}\omega\mu H &= 0, \\ \nabla \times H + \mathrm{i}\omega\varepsilon E &= 0. \end{aligned} \tag{6.3}$$

Since we assume that the ground plane and cavity walls are PECs, we equip the above problem with the boundary condition $\boldsymbol{\nu} \times E = 0$ on the surface of PECs, where $\boldsymbol{\nu}$ is again the unit outward normal. We further assume that the medium and the cavity are invariant with respect to the z-axis. The cross-section of the cavity, denoted by Ω, is rectangular. More specifically, $\Omega = (0,1)^2$. Meanwhile, Γ corresponds to the top of the cavity or the aperture. As stated before, $\Gamma = (0,1) \times \{1\}$. We restrict our attention to the transverse magnetic (TM) polarization. This means that the magnetic field is transverse/perpendicular to the z-axis; moreover, the total electric and magnetic fields take the form $E = (0, 0, u(x,y))$ and $H = (H_x, H_y, 0)$ for some functions $u(x,y)$, H_x, and H_y. Plugging these particular E, H into (6.3) and recalling the boundary condition, we obtain the 2D homogeneous Helmholtz equation defined on the cavity and the upper half space with the homogeneous Dirichlet boundary condition at the surface of PECs, and the scattered field satisfying the Sommerfeld's radiation boundary condition at infinity. By using the half-space Green's function with homogeneous Dirichlet boundary condition (e.g., [11]) or the Fourier transform (e.g., [3, 10]), we can introduce a non-local boundary condition on Γ such that the previous unbounded problem is converted to a bounded problem.

For the standard scattering problem, we want to determine the scattered field u^s in the

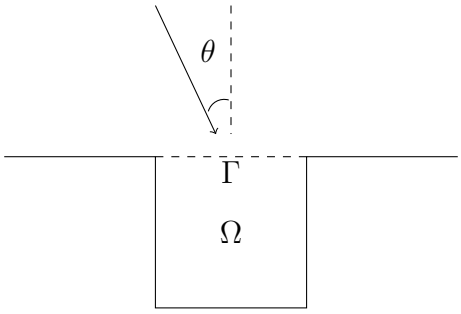


Figure 6.1: Geometry of the scattering from a cavity problem, where $\Omega := (0,1)^2$.

half space and the cavity given an incident plane wave $u^{inc} = e^{\mathrm{i}\alpha x - \mathrm{i}\beta(y-1)}$, where $\alpha = \kappa\sin(\theta)$, $\beta = \kappa\cos(\theta)$, and the incident angle $\theta \in (-\pi/2, \pi/2)$. In particular, $u^s = u - u^{inc} + e^{\mathrm{i}\alpha x + \mathrm{i}\beta(y-1)}$, where u is found by solving the following problem

$$\begin{aligned} &\Delta u + \kappa^2 \varepsilon_r u = 0 \quad \text{in} \quad \Omega, \\ &u = 0 \quad \text{on} \quad \partial\Omega\backslash\Gamma, \\ &\frac{\partial u}{\partial \boldsymbol{\nu}} = \mathcal{T}(u) - 2\mathrm{i}\beta e^{\mathrm{i}\alpha x} \quad \text{on} \quad \Gamma, \end{aligned}$$

where ε_r is the medium's relative permittivity and the non-local boundary operator $\mathcal{T}$ is defined in (6.2). Note that in the model problem (6.1), we assume that $\varepsilon_r = 1$, and allow both f and g to vary. See Fig. 6.1 for an illustration.

6.2 Implementation

Define $\mathcal{H} := \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega\backslash\Gamma\}$. The weak formulation of the problem (6.1) is to find $u \in \mathcal{H}$ such that

$$a(u,v) := \langle \nabla u, \nabla v\rangle_\Omega - \kappa^2\langle u, v\rangle_\Omega - \langle \mathcal{T}(u), v\rangle_\Gamma = \langle g, v\rangle_\Gamma - \langle f, v\rangle_\Omega, \quad \forall v \in \mathcal{H}. \tag{6.4}$$

The existence and uniqueness of the solution to (6.4) have been studied in [3, Theorem 4.1].

We now turn to the wavelet aspects of the numerical scheme. Following the notations and definitions of Chapter 5, we consider a biorthogonal wavelet $(\{\tilde\phi;\tilde\psi\},\{\phi;\psi\})$ with $\widehat{\phi}(0) = (\frac{1}{3},\frac{2}{3})^\mathsf{T}$, $\widehat{\tilde\phi}(0) = (1,1)^\mathsf{T}$, and a biorthogonal wavelet filter bank $(\{\tilde a,\tilde b\},\{a,b\})$ given by

$$a = \left\{ \begin{bmatrix} 0 & -\frac{1}{16} \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \frac{3}{16} \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} & \frac{3}{16} \\ 0 & \frac{3}{8} \end{bmatrix}, \begin{bmatrix} 0 & -\frac{1}{16} \\ \frac{1}{2} & \frac{3}{8} \end{bmatrix} \right\}_{[-2,1]},$$

$$b = \left\{ \begin{bmatrix} 0 & -\frac{1}{32} \\ 0 & -\frac{1}{8} \end{bmatrix}, \begin{bmatrix} \frac{3}{8} & -\frac{9}{32} \\ -\frac{3}{2} & \frac{15}{8} \end{bmatrix}, \begin{bmatrix} \frac{1}{2} & -\frac{9}{32} \\ 0 & -\frac{15}{8} \end{bmatrix}, \begin{bmatrix} \frac{3}{8} & -\frac{1}{32} \\ \frac{3}{2} & \frac{1}{8} \end{bmatrix} \right\}_{[-2,1]},$$
$$\tilde{a} = \left\{ \begin{bmatrix} \frac{3}{32} & -\frac{1}{8} \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} -\frac{3}{16} & \frac{3}{8} \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \frac{11}{16} & \frac{3}{8} \\ -\frac{3}{32} & \frac{3}{8} \end{bmatrix}, \begin{bmatrix} -\frac{3}{16} & -\frac{1}{8} \\ \frac{7}{16} & \frac{3}{8} \end{bmatrix}, \begin{bmatrix} \frac{3}{32} & 0 \\ -\frac{3}{32} & 0 \end{bmatrix} \right\}_{[-2,2]},$$
$$\tilde{b} = \left\{ \begin{bmatrix} -\frac{3}{32} & \frac{1}{8} \\ \frac{3}{128} & -\frac{1}{32} \end{bmatrix}, \begin{bmatrix} \frac{3}{16} & -\frac{3}{8} \\ -\frac{3}{64} & \frac{3}{32} \end{bmatrix}, \begin{bmatrix} \frac{5}{16} & -\frac{3}{8} \\ 0 & -\frac{3}{32} \end{bmatrix}, \begin{bmatrix} \frac{3}{16} & \frac{1}{8} \\ \frac{3}{64} & \frac{1}{32} \end{bmatrix}, \begin{bmatrix} -\frac{3}{32} & 0 \\ -\frac{3}{128} & 0 \end{bmatrix} \right\}_{[-2,2]}.$$

Note that $\phi = (\phi^1, \phi^2)^\mathsf{T}$ has an analytic expression. That is,

$$\phi^1(x) = (2x^2 + 3x + 1)\chi_{[-1,0)} + (2x^2 - 3x + 1)\chi_{[0,1]} \quad \text{and} \quad \phi^2(x) = (-4x^2 + 4x)\chi_{[0,1]}. \tag{6.5}$$

Furthermore, $\mathrm{sm}(a) = \mathrm{sm}(\tilde{a}) = 1.5$ and $\mathrm{sr}(a) = \mathrm{sr}(\tilde{a}) = 3$, and its matching filters $v, \tilde{v} \in (l_0(\mathbb{Z}))^{1\times 2}$ with $\widehat{v}(0)\widehat{\phi}(0) = \widehat{\tilde{v}}(0)\widehat{\tilde{\phi}}(0) = 1$ are given by $\widehat{v}(0) = (1,1)$, $\widehat{v}'(0) = \mathsf{i}(0, \frac{1}{2})$, $\widehat{v}''(0) = (0, -\frac{1}{4})$, $\widehat{\tilde{v}}(0) = (\frac{1}{3}, \frac{2}{3})$, $\widehat{\tilde{v}}'(0) = \mathsf{i}(0, \frac{1}{3})$, and $\widehat{\tilde{v}}''(0) = (\frac{1}{30}, -\frac{1}{5})$. Refer to (5.23) and the beginning of Section 5.7 for the definitions of the foregoing quantities and a matching filter. Let $\phi^L := \phi^1\chi_{[0,\infty)}$ and $\phi^{L,bc} := \phi^2\chi_{[0,\infty)}$. Note that $\phi^L = \phi^L(2\cdot) + \frac{3}{8}\phi^2(2\cdot) - \frac{1}{8}\phi^2(2\cdot - 1)$ and $\phi^{L,bc} = \frac{3}{4}\phi^{L,bc}(2\cdot) + [1, \frac{3}{4}]\phi(2\cdot - 1)$. The direct approach in Chapter 5 yields

$$\begin{aligned} \psi^L &:= \phi^L(2\cdot) - \tfrac{9}{16}\phi^2(2\cdot) + [\tfrac{3}{4}, -\tfrac{1}{16}]\phi(2\cdot - 1), \\ \psi^{L,bc} &:= \phi^{L,bc}(2\cdot) + [-\tfrac{2121}{512}, \tfrac{657}{4096}]\phi(2\cdot - 1) + [\tfrac{3877}{1024}, -\tfrac{4023}{4096}]\phi(2\cdot - 2). \end{aligned}$$

We do not include any information on the dual boundary elements, since they do not play an explicit role in the Galerkin scheme.

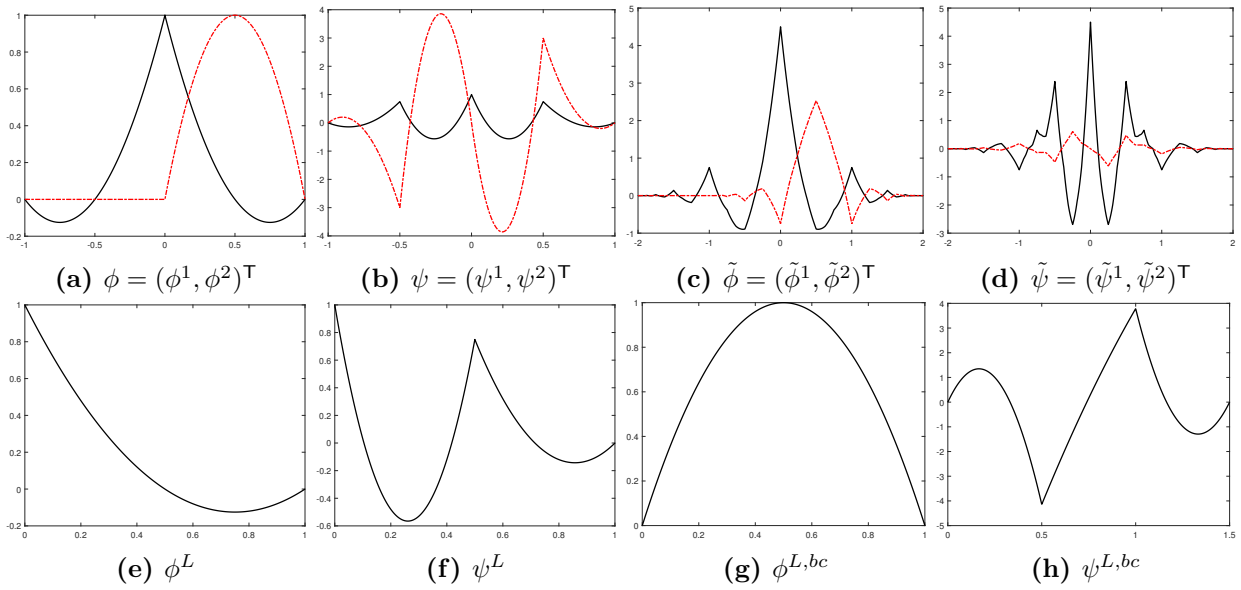
Denote $f_{j;k} := 2^{-j/2} f(2^j \cdot - k)$. For $J_0 \geqslant 1$ and $j \geqslant J_0$, define

$$\Phi^x_{J_0} := \{\phi^{L,bc}_{J_0;0}\} \cup \{\phi_{J_0;k} : 1 \leqslant k \leqslant 2^{J_0} - 1\}, \quad \Psi^x_j := \{\psi^{L,bc}_{j;0}\} \cup \{\psi_{j;k} : 1 \leqslant k \leqslant 2^j - 1\} \cup \{\psi^{R,bc}_{j;2^j-1}\},$$

where $\psi^{R,bc} = \psi^{L,bc}(1 - \cdot)$, and

$$\Phi^y_{J_0} := \Phi^x_{J_0} \cup \{\phi^R_{J_0;2^{J_0}-1}\}, \quad \Psi^y_j := \left(\Psi^x_j \backslash \{\psi^{R,bc}_{j;2^j-1}\}\right) \cup \{\psi^R_{j;2^j-1}\},$$

where $\phi^R = \phi^L(1 - \cdot)$ and $\psi^R = \psi^L(1 - \cdot)$. Then, $\Phi^x_{J_0} \cup \{\Psi^x_j : j \geqslant J_0\}$ forms a Riesz wavelet in $H^1(0,1)$ satisfying the homogeneous Dirichlet boundary condition at both endpoints, and $\Phi^y_{J_0} \cup \{\Psi^y_j : j \geqslant J_0\}$ forms a Riesz wavelet in $H^1(0,1)$ satisfying the homogeneous Dirichlet boundary condition only at the left endpoint. See Fig. 6.2 for the generators of these 1D



(a) $\phi = (\phi^1, \phi^2)^\mathsf{T}$ **(b)** $\psi = (\psi^1, \psi^2)^\mathsf{T}$ **(c)** $\tilde{\phi} = (\tilde{\phi}^1, \tilde{\phi}^2)^\mathsf{T}$ **(d)** $\tilde{\psi} = (\tilde{\psi}^1, \tilde{\psi}^2)^\mathsf{T}$

(e) ϕ^L **(f)** ψ^L **(g)** $\phi^{L,bc}$ **(h)** $\psi^{L,bc}$

Figure 6.2: The generators of Riesz wavelets $\Phi^x_{J_0} \cup \{\Psi^x_j : j \geqslant J_0\}$ and $\Phi^y_{J_0} \cup \{\Psi^y_j : j \geqslant J_0\}$ of $H^1(0,1)$ for $J_0 \geqslant 1$. The black (solid) and red (dotted dashed) lines correspond to the first and second components of a vector function.

Riesz wavelets. By the refinability property, the following relations hold

$$\Phi^x_j = A_{j,j'}\Phi^x_{j'}, \quad \Psi^x_j = \begin{bmatrix} B_{j,j'} \\ B^{R,bc}_{j,j'} \end{bmatrix} \Phi^x_{j'}, \quad \Phi^y_j = \begin{bmatrix} A_{j,j'} \\ A^R_{j,j'} \end{bmatrix} \Phi^y_{j'}, \quad \text{and} \quad \Psi^y_j = \begin{bmatrix} B_{j,j'} \\ B^R_{j,j'} \end{bmatrix} \Phi^y_{j'} \quad \forall j < j', \tag{6.6}$$

where $A_{j,j'}$, $A^R_{j,j'}$, $B_{j,j'}$, $B^{R,bc}_{j,j'}$, and $B^R_{j,j'}$ are well-defined matrices.

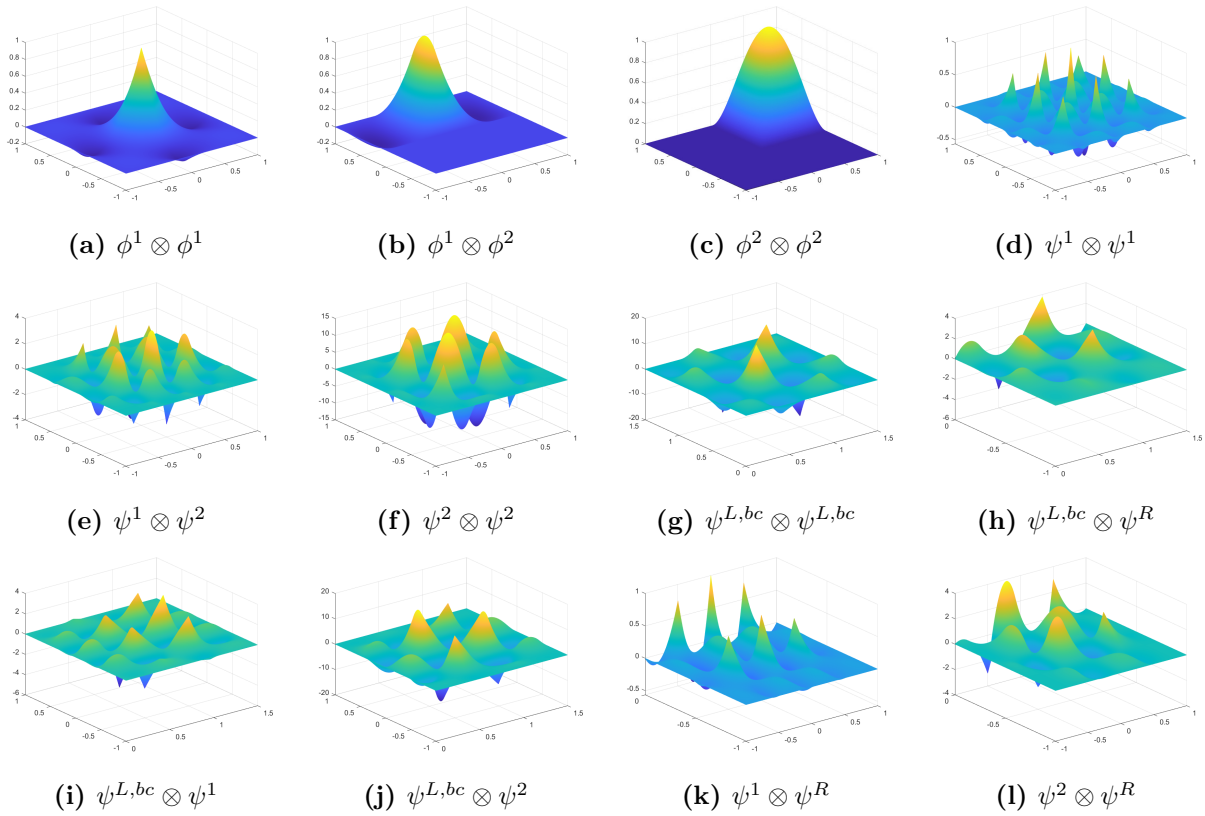
Given one-dimensional functions $f_1, f_2 : \mathbb{R} \to \mathbb{C}$, the two-dimensional function $f_1 \otimes f_2$ is defined by $(f_1 \otimes f_2)(x,y) := f_1(x)f_2(y)$, where $x, y \in \mathbb{R}$. Furthermore, if F_1, F_2 are sets containing one-dimensional functions, then $F_1 \otimes F_2 := \{f_1 \otimes f_2 : f_1 \in F_1, f_2 \in F_2\}$. Define

$$\mathcal{B}_{J_0,J} := (\Phi^x_{J_0} \otimes \Phi^y_{J_0}) \cup \{\Phi^x_j \otimes \Psi^y_j \cup \Psi^x_j \otimes \Phi^y_j \cup \Psi^x_j \otimes \Psi^y_j : J_0 \leqslant j \leqslant J-1\}.$$

Note that when $J = J_0$, $\mathcal{B}_{J_0,J_0} = \Phi^x_{J_0} \otimes \Phi^y_{J_0}$. The 2D Riesz wavelet in $\mathcal{H}$ we shall employ is $\mathcal{B}_{J_0} := \mathcal{B}_{J_0,\infty}$, where $J_0 \geqslant 1$. See Fig. 6.3 for some generators of $\mathcal{B}_{J_0}$.

In the Galerkin scheme, our approximated solution is of the form $u_J = \sum_{\eta \in \mathcal{B}_{J_0,J}} c_\eta \eta$. Plugging it into the weak formulation (6.4), using test functions in $\mathcal{B}_{J_0,J}$, and recalling the relations in (6.6), we obtain the linear system

$$\left(R\left([\langle v, w\rangle_{(0,1)}]_{v,w\in\Phi^x_J} \otimes [\langle v, w\rangle_{(0,1)}]_{v,w\in\Phi^y_J}\right) R^\mathsf{T} - T\right) C = F, \tag{6.7}$$



(a) $\phi^1 \otimes \phi^1$ **(b)** $\phi^1 \otimes \phi^2$ **(c)** $\phi^2 \otimes \phi^2$ **(d)** $\psi^1 \otimes \psi^1$

(e) $\psi^1 \otimes \psi^2$ **(f)** $\psi^2 \otimes \psi^2$ **(g)** $\psi^{L,bc} \otimes \psi^{L,bc}$ **(h)** $\psi^{L,bc} \otimes \psi^R$

(i) $\psi^{L,bc} \otimes \psi^1$ **(j)** $\psi^{L,bc} \otimes \psi^2$ **(k)** $\psi^1 \otimes \psi^R$ **(l)** $\psi^2 \otimes \psi^R$

Figure 6.3: Some generators of the Riesz wavelet $\mathcal{B}_{J_0}$ of $\mathcal{H}$, where $J_0 \geqslant 1$.

where $R := [R_1^{\mathsf{T}}, R_2^{\mathsf{T}}, \cdots, R_{2(J-J_0)+1}^{\mathsf{T}}]^{\mathsf{T}}$ with $R_1 := A_{J_0,J} \otimes A_{J_0,J}$,

$$R_\ell := \begin{bmatrix} B_{J_0+\ell-2,J} \otimes A_{J_0+\ell-2,J} \\ B^{R,bc}_{J_0+\ell-2,J} \otimes A_{J_0+\ell-2,J} \\ A_{J_0+\ell-2,J} \otimes B_{J_0+\ell-2,J} \\ B_{J_0+\ell-2,J} \otimes B_{J_0+\ell-2,J} \\ B^{R,bc}_{J_0+\ell-2,J} \otimes B_{J_0+\ell-2,J} \end{bmatrix}, \quad R_{J-J_0+\ell} := \begin{bmatrix} B_{J_0+\ell-2,J} \otimes A^R_{J_0+\ell-2,J} \\ B^{R,bc}_{J_0+\ell-2,J} \otimes A^R_{J_0+\ell-2,J} \\ A_{J_0+\ell-2,J} \otimes B^R_{J_0+\ell-2,J} \\ B_{J_0+\ell-2,J} \otimes B^R_{J_0+\ell-2,J} \\ B^{R,bc}_{J_0+\ell-2,J} \otimes B^R_{J_0+\ell-2,J} \end{bmatrix},$$

$$2 \leqslant \ell \leqslant J - J_0 + 1,$$

$$S := [S_1^{\mathsf{T}}, \cdots, S_{J-J_0}^{\mathsf{T}}]^{\mathsf{T}}, \quad S_\ell := 2^{-(J_0+\ell-1)/2} \begin{bmatrix} B_{J_0+\ell-1,J} \\ B^{R,bc}_{J_0+\ell-1,J} \\ A_{J_0+\ell-1,J} \\ B_{J_0+\ell-1,J} \\ B^{R,bc}_{J_0+\ell-1,J} \end{bmatrix}, \quad 1 \leqslant \ell \leqslant J - J_0,$$

$$T := \begin{bmatrix} 0_{(\mathrm{rows}(R)-\mathrm{rows}(S))\times(\mathrm{rows}(R)-\mathrm{rows}(S))} & 0_{(\mathrm{rows}(R)-\mathrm{rows}(S))\times\mathrm{rows}(S)} \\ 0_{\mathrm{rows}(S)\times(\mathrm{rows}(R)-\mathrm{rows}(S))} & S\left[\langle \mathcal{T}(\eta), \zeta\rangle_\Gamma\right]_{\eta,\zeta\in\Phi_J^x} S^{\mathsf{T}} \end{bmatrix},$$

$$F := \begin{bmatrix} 0_{(\mathrm{rows}(R)-\mathrm{rows}(S))\times 1} \\ S\left[\langle g, v\rangle_\Gamma\right]_{v\in\Phi_J^x} \end{bmatrix} - R\mathrm{vec}\left(\left[\langle f, vw\rangle_\Omega\right]_{w\in\Phi_J^y, v\in\Phi_J^x}\right),$$

$\otimes$ denotes the kronecker product, C denotes the coefficients $\{c_\eta\}_{\eta\in\mathcal{B}_{J_0,J}}$ properly arranged in a vector form, $0_{m\times n}$ denotes an $m\times n$ zero matrix, $\mathrm{rows}(\cdot)$ denotes the number of rows of a given matrix, and $\mathrm{vec}(\cdot)$ denotes the standard vectorization operation. We make some important remarks regarding the assembly of the linear system. First, we further normalize each element in $\mathcal{B}_{J_0,J}$ by $|a(\cdot,\cdot)|^{-1/2}$, where $a(\cdot,\cdot)$ is defined in (6.4). This makes the modulus of all diagonal entries of the coefficient matrix on the left-hand side of (6.7) equal to 1. Second, we note that the assembly of the linear system can be done efficiently by exploiting the refinability structure. The inner products are computed only for the refinable functions at the highest scale level (i.e., elements of Φ_J^x and Φ_J^y). Third, following [98, Remark 4.1], we rewrite the non-local boundary condition as

$$\mathcal{T}(v) = \int_0^1 \ln(|x-x'|)q_0(x-x')v(x')dx' + \int_0^1 q_1(x-x')v(x')dx' + \frac{1}{\pi}\fint_0^1 \frac{v(x')}{|x-x'|^2}dx', \quad (6.8)$$

where

$$q_0(s) := \frac{\mathrm{i}\kappa H_1^{(1)}(\kappa|s|)}{2|s|} + \frac{\kappa J_1(\kappa|s|)}{\pi|s|}\ln(|s|) - \frac{1}{\pi|s|^2}, \quad q_1(s) := -\frac{\kappa J_1(\kappa|s|)}{\pi|s|},$$

and J_1 is the first order Bessel function of the first kind. Note that $q_0(s)$ and $q_1(s)$ are even analytic functions. The first integral in (6.8) is only weakly singular. After properly partitioning this integral so that the weak singularity appears on an endpoint, we can use a combination of the Gauss-Legendre and double exponential quadratures to compute it. The second integral in (6.8) can be handled by the Gauss-Legendre quadrature. Recall that

$$\fint_0^1 \frac{v(x')}{(x-x')^2}dx' := \lim_{\epsilon\to 0}\left(\int_0^{x-\epsilon}\frac{v(x')}{(x-x')^2}dx' + \int_{x+\epsilon}^1 \frac{v(x')}{(x-x')^2}dx' - \frac{2v(x)}{\epsilon}\right). \quad (6.9)$$

For the right side to exist, a sufficient condition is $v\in C^{1,\alpha}(0,1)$ (i.e., the first derivative of v is α-Hölder continuous on the unit interval with $0<\alpha\leqslant 1$). Then, the third integral of (6.8) can be exactly computed by (6.9), since the Riesz wavelet we employ has an analytic expression.

6.3 Numerical experiments

In what follows, we present several numerical experiments to compare the performance of the wavelet and standard Galerkin schemes. The relative errors reported below are in terms of 2-norm. Assuming that the exact solution u exists, we define

$$\|u - u_J\|_2^2 := 2^{-22} \sum_{i=1}^{2^{11}} \sum_{j=1}^{2^{11}} |u(x_i, y_j) - u_J(x_i, y_j)|^2,$$

where (x_i, y_j) for $i, j = 0, \ldots, 2^{11}$, and $x_{i+1} - x_i = y_{j+1} - y_j = 2^{-11}$ for all $i, j = 0, \ldots, 2^{11} - 1$. Note that the above error is just an approximation of the error in the L_2 norm. In each table below, we report the relative errors $\|u - u_J\|_2 / \|u\|_2$ in the 'Rel. err' column. We also report the condition numbers (i.e., the ratio of the largest and smallest singular values) of the coefficient matrices coming from the wavelet and standard Galerkin methods, which are respectively denoted by $\mathcal{B}_{J_0,J}$ and $\Phi_J := \Phi_J^x \otimes \Phi_J^y$ in each table below. Their ratios are reported in the column 'CN Ratio.' The convergence rate reported in the 'Order' column is obtained by calculating $\log_2(\|u - u_J\|_2 / \|u - u_{J+1}\|_2)$.

Example 6.1. Consider the model problem (6.1), where $\mathcal{T}$ is defined in (6.2), and f and g are chosen such that $u = \exp(xy)\sin(\kappa x)\sin((\kappa + \pi/2)y)$. Additionally, we let $\kappa = 4\pi, 8\pi, 16\pi$. See Table 6.1 for the numerical results.

Example 6.2. Consider the model problem (6.1), where $\mathcal{T}$ is defined in (6.2), $\kappa = 32\pi$, and f and g are chosen such that $u = \sin(\pi x)\sin(\sqrt{\kappa^2 - \pi^2}y)$. See Table 6.2 for the numerical results.

In all cases, we observe that the condition numbers of the coefficient matrices associated with the standard Galerkin method are around 2 to 800 times worse than those associated with the wavelet Galerkin method. The rapid growth in these condition numbers is primarily caused by the decreasing smallest singular values; the largest singular values, on the other hand, behave like a constant at various scale levels. One can expect that this ratio continues to increase dramatically as the scale level J increases. Also, since $\operatorname{span}(\mathcal{B}_{J_0,J}) = \operatorname{span}(\Phi_J)$ for all $J \geqslant J_0 \geqslant 1$, it is not surprising that the errors are essentially identical. As a final remark we note that the choice of ϕ in our biorthogonal wavelet in (6.5) actually belongs to a special family of interpolating refinable functions. These interpolating refinable functions have been well studied. To achieve a higher order convergence rate, we may replace the current ϕ with another one in the same family but with a higher multiplicity. For this family of interpolating refinable functions, a higher multiplicity means the refinable function has a higher polynomial reproduction order and consequently a higher convergence rate.

$\kappa = 4\pi$							
J	CN of $\mathcal{B}_{2,J}$	CN of Φ_J	CN Ratio	Rel. err of $\mathcal{B}_{2,J}$	Order	Rel. err of Φ_J	Order
3	73.6	184.0	2.5	3.89E-2		3.89E-2	
4	145.3	853.7	5.9	4.90E-3	2.99	4.90E-3	2.99
5	197.0	3464.4	17.6	6.11E-4	3.00	6.11E-4	3.00
6	232.2	13876.3	59.8	7.63E-5	3.00	7.63E-5	3.00
7	256.2	55503.0	216.7	9.53E-6	3.00	9.53E-6	3.00
8	273.0	221978.6	813.0	1.19E-6	3.00	1.19E-6	3.00
$\kappa = 8\pi$							
J	CN of $\mathcal{B}_{3,J}$	CN of Φ_J	CN Ratio	Rel. err of $\mathcal{B}_{3,J}$	Order	Rel. err of Φ_J	Order
3	92.2	92.2	1	2.03E-1		2.03E-1	
4	185.2	491.5	2.7	3.51E-2	2.53	3.51E-2	2.53
5	314.9	3015.3	9.6	4.39E-3	3.00	4.39E-3	3.00
6	342.6	12556.1	36.7	5.46E-4	3.00	5.46E-4	3.00
7	355.0	50397.3	142.0	6.82E-5	3.00	6.82E-5	3.00
8	364.0	201647.9	554.0	8.53E-6	3.00	8.53E-6	3.00
$\kappa = 16\pi$							
J	CN of $\mathcal{B}_{4,J}$	CN of Φ_J	CN Ratio	Rel. err of $\mathcal{B}_{4,J}$	Order	Rel. err of Φ_J	Order
4	297.3	297.3	1	1.91E-1		1.91E-1	
5	415.2	908.9	2.2	3.36E-2	2.51	3.36E-2	2.51
6	1061.6	10192.8	9.6	4.17E-3	3.01	4.17E-3	3.01
7	1266.4	46561.9	36.8	5.28E-4	2.98	5.28E-4	2.98
8	1320.9	188128.2	142.4	1.11E-4	2.25	1.11E-4	2.25

Table 6.1: Condition numbers and relative errors for Example 6.1.

$\kappa = 32\pi$							
J	CN of $\mathcal{B}_{5,J}$	CN of Φ_J	CN Ratio	Rel. err of $\mathcal{B}_{5,J}$	Order	Rel. err of Φ_J	Order
5	1285.9	1285.9	1	1.72		1.72	
6	1258.5	2673.7	2.1	4.18E-1	2.04	4.18E-1	2.04
7	2979.5	28615.4	9.6	2.97E-2	3.81	2.97E-2	3.81
8	4738.9	174325.0	36.8	1.93E-3	3.95	1.93E-3	3.95
9	5066.7	722029.3	142.5	1.26E-4	3.93	1.26E-4	3.95

Table 6.2: Condition numbers and relative errors for Example 6.2.

Chapter 7

Future Work

We conclude this thesis by outlining some directions of some future work.

Generalizing DAT for solving a larger class of the 2D Helmholtz equation is a problem we may consider in the future. It is a challenging multifaceted problem, which requires new ideas. There are two critical issues that need to be resolved. The source term would still be partitioned by shifted square hat functions and their refinability can still be used to give rise to the tree structure. The first issue comes from the 2D Dirac assisted local problems. In contrast to the 1D Dirac assisted local problem whose source term is a Dirac distribution at a single point, the 2D Dirac assisted local problems would have the source term consisting of weighted Dirac distributions defined along the boundary of a rectangular subdomain. Obtaining a highly accurate numerical solution of such 2D Dirac assisted local problem and accurately estimating its outward fluxes are challenging, because its weak solution involves highly singular functions caused by the singular distribution source term. The second issue is on formulating the linking problems and generalizing Theorem 2.2 for stitching all the local solutions into a global solution. In a multidimensional setting, this is considerably more difficult than in 1D due to more complicated topology and boundaries of subdomains.

Another problem is to explore the possibility of obtaining a sixth order compact FDM for the 2D Helmholtz equation with variable coefficients and reduced pollution effect. This can be considered as a generalization of Chapter 3. So far, there already exists a fourth order FDM for the 2D Helmholtz equation with smooth variable coefficients [16]; however, the authors did not further reduce the pollution effect. This work would be useful in solving the 2D Helmholtz equation with the perfectly matched layer boundary condition, which is commonly used in geophysics. Throughout this thesis, we have mostly used the first order absorbing boundary condition in our model problems.

Still in the context of FDM, we may also perform a rigorous error analysis for our proposed FDM. The major difficulty comes from the fact that the coefficient matrix is sign-indefinite.

Thus, new techniques may be needed to carry out the analysis.

The stability bounds in Chapter 4 motivate us to develop a new numerical scheme, which presently is a combination of a Fourier method and DAT. We use our stability bounds to strategically pick dominant Fourier coefficients in the solution, which are then computed by a Filon-type quadrature due to its accuracy and adaptivity. Our initial experiments suggest that these bounds play a critical role, especially when we have high frequency boundary data. At this moment, we are trying to figure out what high frequency boundary data are typically encountered in practice and whether they have any structures.

In terms of the wavelet Galerkin method we proposed, we may apply the same methodology in Chapter 6 to wavelets whose primal refinable function belongs to the same family as (6.5) but with a higher multiplicity. As discussed earlier, a higher multiplicity in this case ultimately translates to a higher rate of convergence. One challenge we encounter is finding an appropriate compactly biorthogonal wavelet on the real line, since there are many freedoms and they take the form of a nonlinear system. We may also consider using a biorthogonal wavelet whose dual has infinite support, and see what extra benefits it can offer (e.g., a further reduction of the condition number). Unfortunately, the theory presented in Chapter 5 does not apply to this situation. If a biorthogonal wavelet with an infinitely supported dual does offer extra advantages, then we may also consider building a theory to construct/adapt such wavelets on a bounded interval similar to the content of Chapter 5. In Chapter 6, we performed an exhaustive parameter search to reduce the condition number as much as possible. A question that may be worthwhile addressing is if there is a systematic or more efficient way to do this. Finally, similar to before, we are also interested in performing a rigorous error analysis for our proposed wavelet Galerkin method.

Bibliography

[1] G. Alessandrini, Strong unique continuation for general elliptic equations in 2D. *J. Math. Anal. Appl.* **386** (2012), 669-676.

[2] A. Altürk and F. Keinert, Regularity of boundary wavelets. *Appl. Comput. Harmon. Anal.* **32** (2012), 65–85.

[3] H. Ammari, G. Bao, and A. W. Wood, Analysis of the electromagnetic scattering from a cavity. *Japan J. Indust. Appl. Math.* **19** (2002), 301-310.

[4] L. Andersson, N. Hall, B. Jawerth, and G. Peters, Wavelets on closed subsets of the real line. Recent advances in wavelet analysis, 1–61, *Wavelet Anal. Appl.*, 3, Academic Press, Boston, MA, 1994.

[5] E. Ashpazzadeh, B. Han, and M. Lakestani, Biorthogonal multiwavelets on the interval for numerical solutions of Burgers' equation. *J. Comput. Appl. Math.* **317** (2017), 510–534.

[6] P. Auscher, Ondelletes á support compact et conditions aux limites. *J. Funct. Anal.* **111** (1993), 29–43.

[7] A. K. Aziz, R. B. Kellogg, and A. B. Stephens. A two point boundary value problem with a rapidly oscillating solution. *Numer. Math.* **53** (1988), 107-121.

[8] I. M. Babuška and J. M. Melenk, The partition of unity method. *Internat. J. Numer. Methods Engrg.* **40** (1997), no. 4, 727–758.

[9] I. M. Babuška and S. A. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Rev.* **42** (2000), no. 3, 451-484.

[10] G. Bao and J. Lai, Radar cross section reduction of a cavity in the ground plane. *Commun. Comput. Phys.* **15** (2014), no. 4, 895-910.

[11] G. Bao and W. Sun, A fast algorithm for the electromagnetic scattering from a large cavity. *SIAM J. Sci. Comput.* **27** (2005), no. 2, 553-574.

[12] G. Bao and K. Yun, Stability for the electromagnetic scattering from large cavities. *Arch. Rational Mech. Anal.* **220** (2016), 1003-1044.

[13] G. Bao, K. Yun, and Z. Zou, Stability of the scattering from a large electromagnetic cavity in two dimensions. *SIAM J. Math. Anal.* **44** (2012), no.1, 383-404.

[14] A. Bendalli and Ph. Guillaume, Non-reflecting boundary conditions for waveguides. *Math. Comp.* **68** (1999), no. 225, 123-144.

[15] S. Britt, S. Tsynkov, and E. Turkel, A compact fourth order scheme for the Helmholtz equation in polar coordinates. *J. Sci. Comput.* **45** (2010), 26-47.

[16] S. Britt, S. Tsynkov, and E. Turkel, Numerical simulation of time-harmonic waves in inhomogeneous media using compact high order schemes. *Commun. Comput. Phys.* **9** (2011), no. 3, 520-541.

[17] D. L. Brown, D. Gallistl, and D. Peterseim, Multiscale Petrov-Galerkin method for high-frequency heterogeneous Helmholtz equations, in Meshfree Methods for Partial Differential Equations VIII, M. Griebel, M. Schweitzer, *Lect. Notes Comput. Sci. Eng.* **115**, Springer, Cham, 85-115.

[18] D. Černá, Wavelets on the interval and their applications, Habilitation thesis at Masaryk University, (2019).

[19] D. Černá and V. Finěk, Construction of optimally conditioned cubic spline wavelets on the interval. *Adv. Comput. Math.* **34** (2011),219–252.

[20] D. Černá and V. Finěk, Cubic spline wavelets with complementary boundary conditions. *Appl. Math. Comput.* **219** (2012), 1853–1865.

[21] T. Chaumont-Frelet, Approximations par l'elements finis de problemes d'Helmholtz pour la propagation d'ondes sismiques, PhD Thesis at Inria, (2015).

[22] Z. Chen, D. Cheng, W. Feng, and T. Wu, An optimal 9-point finite difference scheme for the Helmholtz equation with PML. *Int. J. Numer. Anal. Mod.* **10** (2013), no. 2, 389-410.

[23] Z. Chen, T. Wu, and H. Yang, An optimal 25-point finite difference scheme for the Helmholtz equation with PML. *J. Comput. Appl. Math.* **236** (2011), 1240-1258.

[24] C. K. Chui and J. Z. Wang, On compactly supported wavelets and a duality principle, *Trans. Amer. Math. Soc.* **330** (1992) 903–916.

[25] C. K. Chui and E. Quak, Wavelets on a bounded interval. Numerical methods in approximation theory, Vol. 9, 53–75, *Internat. Ser. Numer. Math.*, 105, Birkhäuser, Basel, 1992.

[26] P.-H. Cocquet, M. J. Gander, and X. Xiang, A finite difference method with optimized dispersion correction for the Helmholtz equation. Domain decomposition methods in science and engineering XXIV, *Lecture Notes in Computational Science and Engineering* **125**, Springer, Cham, 2018, 205-213.

[27] P.-H. Cocquet, M. J. Gander, and X. Xiang, Dipersion correction for Helmholtz in 1D with piecewise constant wavenumber. Domain decomposition methods in science and engineering XXV, *Lecture Notes in Computational Science and Engineering* **138**, Springer, Cham, 2020, 359-366.

[28] P.-H. Cocquet, M. J. Gander, and X. Xiang, Closed form dispersion corrections including a real shifted wavenumber for finite difference discretizations of 2D constant coefficient Helmholtz problems. *SIAM J. Sci. Comput.* **43** (2021), no. 1, A278-A308.

[29] A. Cohen, I. Daubechies, and J. C. Feauveau, Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **45** (1992), 485–560.

[30] A. Cohen, I. Daubechies, and G. Plonka, Regularity of refinable function vectors. *J. Fourier Anal. Appl.* **3** (1997), 295–324.

[31] A. Cohen, I. Daubechies, and P. Vial, Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** (1993), 54–81.

[32] P. Cummings and X. Feng, Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Math. Models Methods. Appl. Sci.* **16** (2006), no. 1, 139-160.

[33] W. Dahmen, B. Han, R.-Q. Jia, and A. Kunoth, Biorthogonal multiwavelets on the interval: cubic Hermite splines. *Constr. Approx.* **16** (2000), 221–259.

[34] W. Dahmen, A. Kunoth and K. Urban, Biorthogonal spline wavelets on the interval—stability and moment conditions. *Appl. Comput. Harmon. Anal.* **6** (1999), 132–196.

[35] W. Dahmen and R. Schneider, Wavelets with complementary boundary conditions–function spaces on the cube. *Results Math.* **34** (1998), 255–293.

[36] H. Dastour and W. Liao, A fourth-order optimal finite difference scheme for the Helmholtz equation with PML. *Comput. Math. Appl.* **78** (2019), no. 6, 2147-2165.

[37] H. Dastour and W. Liao, An optimal 13-point finite difference scheme for a 2D Helmholtz equation with a perfectly matched layer boundary condition. *Numer. Algorithms* **86** (2021), 1109-1141.

[38] I. Daubechies, Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41** (1988), 909–996.

[39] T. J. Dijkema and R. Stevenson. A sparse Laplacian in tensor product wavelet corrdinates. *Numer. Math.* **115** (2010), 433–449.

[40] G. Donovan, J. Geronimo, and D. Hardin, Intertwining multiresolution analyses and the construction of piecewise-polynomial wavelets. *SIAM J. Math. Anal.* **27** (1996), 1791–1815.

[41] K. Du, B. Li, and W. Sun. A numerical study on the stability of a class of Helmholtz problems. *J. Comput. Phys.* **287** (2015), 46-59.

[42] K. Du, W. Sun, and X. Zhang, Arbitrary high-order C^0 tensor product Galerkin finite element methods for the electromagnetic scattering from a large cavity. *J. Comput. Phys.* **242** (2013), 181-195.

[43] Y. Du and H. Wu, Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number. *SIAM J. Numer. Anal.* **53** (2015), no. 2, 782-804.

[44] Y. A. Erlangga, C. W. Oosterlee, and C. Vuik, A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comput.* **27** (2006), no. 4, 1471-1492.

[45] O. G. Ernst and M. J. Gander, Why is it difficult to solve Helmholtz problems with classical iterative methods. Numerical analysis of multiscale problems, *Lecture Notes in Computational Science and Engineering* **83**, Springer, Berlin, Heidelberg, 2011, 325-363.

[46] O. G. Ernst and M. J. Gander, Multigrid methods for Helmholtz problems: A convergent scheme in 1D using standard components. Direct and inverse problems in wave propagation and applications, *Radon Series on Computational and Applied Mathematics* **14**, De Gruyter, Berlin, 2013, 135-186.

[47] S. Esterhazy and J. M. Melenk, On stability of discretizations of the Helmholtz equation, in Numerical Analysis of Multiscale Problems, I. G. Graham, T. Y. Hou, O. Lakkis, and R. Scheil, *Lect. Notes Comput. Sci. Eng.* **83**, Springer-Verlag, Berlin, 2012, 285-324.

[48] Q. Fang, J. Shen, and L. Wang, An efficient and accurate spectral method for acoustic scattering in elliptic domains. *Numer. Math. Theor. Meth. Appl.* **2** (2009), no. 3, 258-274.

[49] Q. Feng, B. Han, and P. Minev, Sixth order compact finite difference schemes for Poisson interface problems with singular sources. *Comp. Math. Appl.* **99** (2021), 2-25.

[50] Q. Feng, B. Han, and P. Minev, A high order compact finite difference scheme for elliptic interface problems with discontinuous and high-contrast coefficients. *Appl. Math. Comput.* **431** (2022), 127314.

[51] Q. Feng, B. Han, and M. Michelle, Sixth order compact finite difference method for 2D Helmholtz equations with singular sources and reduced pollution effect. Submitted to *SIAM J. Sci. Comput.* (2021), arXiv:2112.07154v1, 20 pages.

[52] X. Feng and D. Sheen, An elliptic regularity coefficient estimate for a problem arising from the frequency domain treatment of waves. *Trans. Amer. Math. Soc.* **346** (1994), 475-487.

[53] X. Feng and H. Wu, Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM J. Numer. Anal.* **47** (2009), no. 4, 2872-2896.

[54] X. Feng and H. Wu, *hp*-discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Math. Comp.* **80** (2011), no. 276, 1997-2024.

[55] S. Fu and K. Gao, A fast solver for the Helmholtz equation based on the generalized multiscale finite-element method. *Geophys. J. Int.* **211** (2017), no. 2, 797-813.

[56] S. Fu, G .Li, R. Craster, and S. Guenneau, Wavelet-based edge multiscale finite element method for Helmholtz problems in perforated domains. *Multiscale Model. Simul.* **19** (2021), no. 4, 1684-1709.

[57] M. J. Gander and H. Zhang, A class of iterative solvers for the Helmholtz equation: factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods. *SIAM Rev.* **61** (2019), no. 1, 3-76.

[58] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, Fractal functions and wavelet expansions based on several scaling functions. *J. Approx. Theory* **78** (1994), 373–401.

[59] S. S. Goh, Q. T. Jiang, and T. Xia, Construction of biorthogonal multiwavelets using the lifting scheme. *Appl. Comput. Harmon. Anal.* **9** (2000), 336–352.

[60] I. G. Graham, O. R. Pembery, and E. A. Spence, The Helmholtz equation in heterogeneous media: a priori bounds, well-posedness, and resonances. *J. Differential Equations* **266** (2019), 2869–2923.

[61] I. G. Graham and S. A. Sauter, Stability and finite element error analysis for the Helmholtz equation with variable coefficients. *Math. Comp.* **89** (2020), no. 321, 105-138.

[62] P. Grisvard, Elliptic problems in nonsmooth domains. Pitman Publishing Inc., Massachusetts, US, 1985. xx + 410 pp.

[63] T. Hagstrom and S. Kim, Complete radiation boundary conditions for the Helmholtz equation I: waveguides. *Numer. Math.* **141** (2019), 917-966.

[64] B. Han and R.-Q. Jia, Multivariate refinement equations and convergence of subdivision schemes. *SIAM J. Math. Anal.* **29** (1998), 1177–1199.

[65] B. Han, Approximation properties and construction of Hermite interpolants and biorthogonal multiwavelets. *J. Approx. Theory* **110** (2001), 18–53.

[66] B. Han, Vector cascade algorithms and refinable function vectors in Sobolev space, *J. Approx. Theory.* **124** (2003), 44–88.

[67] B. Han, Compactly supported tight wavelet frames and orthonormal wavelets of exponential decay with a general dilation matrix. *J. Comput. Appl. Math.* **155** (2003), 43–67.

[68] B. Han, Solutions in Sobolev spaces of vector refinement equations with a general dilation matrix. *Adv. Comput. Math.* **24** (2006), 375–403.

[69] B. Han, Pairs of frequency-based nonhomogeneous dual wavelet frames in the distribution space. *Appl. Comput. Harmon. Anal.* **29** (2010), 330–353.

[70] B. Han, Nonhomogeneous wavelet systems in high dimensions. *Appl. Comput. Harmon. Anal.* **32** (2012), 169–196.

[71] B. Han, Framelets and wavelets: Algorithms, analysis, and applications. *Applied and Numerical Harmonic Analysis*. Birkhäuser/Springer, Cham, 2017. xxxiii + 724 pp.

[72] B. Han and Q. T. Jiang, Multiwavelets on the interval. *Appl. Comput. Harmon. Anal.* **12** (2002), 100–127.

[73] B. Han and M. Michelle, Construction of wavelets and framelets on a bounded interval. *Anal. Appl.* **16** (2018), 807–849.

[74] B. Han and M. Michelle, Derivative-orthogonal Riesz wavelets in Sobolev spaces with applications to differential equations. *Appl. Comp. Harmon. Anal.* **47** (2019), no. 3, 759-794.

[75] B. Han and M. Michelle, Wavelets on intervals derived from arbitrary compactly supported biorthogonal multiwavelets. *Appl. Comp. Harmon. Anal.* **53** (2021), 270-331.

[76] B. Han and M. Michelle, Sharp wavenumber-explicit stability bounds for 2D Helmholtz equations. Accepted for publication in *SIAM J. Numer. Anal.* (2022), arXiv:2108.06469, 28 journal pages.

[77] B. Han, M. Michelle, and Y. S. Wong, Dirac assisted tree method for 1D heterogeneous Helmholtz equations with arbitrary variable wave numbers. *Comput. Math. Appl.* **97** (2021), 416-438.

[78] B. Han and Q. Mo, Analysis of optimal bivariate symmetric refinable Hermite interpolants. *Commun. Pure Appl. Anal.* **6** (2007), 689–718.

[79] B. Han and X. Zhuang, Matrix extension with symmetry and its application to symmetric orthonormal multiwavelets. *SIAM J. Math. Anal.* **42** (2010), 2297–2317.

[80] D. P. Hardin and S. A. Marasovich, Biorthogonal multiwavelets on $[-1, 1]$. *Appl. Comput. Harmon. Anal.* **7** (1999), 34–53.

[81] Y. He, P. Li, and J. Shen, A new spectral method for numerical solution of the unbounded rough surface scattering problem. *J. Comput. Phys.* **275** (2014), 608-625.

[82] Y. He, D. P. Nicholls, and J. Shen, An efficient and stable spectral method for electromagnetic scattering from a layered periodic structure. *J. Comput. Phys.* **231** (2012), 3007-3022.

[83] U. Hetmaniuk, Stability estimates for a class of Helmholtz problems. *Commun. Math. Sci.* **5** (2007), no. 3, 665-678.

[84] R. Hiptmair, A. Moiola, and I. Perugia, A survey of Trefftz methods for the Helmholtz equation. Building bridges: connections and challenges in modern approaches to numerical partial differential equations, *Lecture Notes in Computational Science and Engineering* **114**, Springer, Cham, 2016, 237-279.

[85] L. Hu, L. Ma, and J. Shen, Efficient spectral-Galerkin method and analysis for elliptic PDEs with non-local boundary conditions. *J. Sci. Comput.* **68** (2016), 417-437.

[86] F. Ihlenburg, Finite Element Analysis of Acoustic Scattering, *Applied Mathematical Sciences*. Springer-Verlag New York, Inc., 1998, xiv + 226 pp.

[87] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, Computational Ocean Acoustics, *Modern Acoustics and Signal Processing*. Springer, New York, 2011. xviii+794 pp.

[88] R.-Q. Jia, Spline wavelets on the interval with homogeneous boundary conditions. *Adv. Comput. Math.* **30** (2009), 177–200.

[89] R.-Q. Jia and Q. T. Jiang, Spectral analysis of the transition operator and its applications to smoothness analysis of wavelets. *SIAM J. Matrix Anal. Appl.* **24** (2003), 1071–1109.

[90] R.-Q. Jia and S.-T. Liu, Wavelet bases of Hermite cubic splines on the interval. *Adv. Comput. Math.* **25** (2006), no. 1-3, 23–39.

[91] R.-Q. Jia, S. D. Riemenschneider, and D.-X. Zhou, Smoothness of multiple refinable functions and multiple wavelets. *SIAM J. Matrix Anal. Appl.* **21** (1999), 1–28.

[92] R.-Q. Jia and W. Zhao, Riesz bases of wavelets and applications to numerical solutions of elliptic equations. *Math. Comp.* **80** (2011), no. 275, 1525–1556.

[93] Q. T. Jiang, Multivariate matrix refinable functions with arbitrary matrix dilation. *Trans. Amer. Math. Soc.* **351** (1999), 2407–2438.

[94] A. Jouini and P. G. Lemarie-Rieusset, Analyse multi-résolution bi-orthogonale sur l'intervalle et applications. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **10** (1993), 453–476.

[95] F. Keinert, Wavelets and multiwavelets. *Studies in Advanced Mathematics*. Chapman & Hall/CRC, Boca Raton, FL, 2004. xii+275 pp.

[96] F. Keinert, Regularity and construction of boundary multiwavelets. *Poincaré J. Anal. Appl.* **2**, (2015), 1–12.

[97] O. Lagrouche, P. Bettess, E. Perrey-Debain, and J. Trevelyan, Wave interpolation finite elements for Helmholtz problems with jumps in the wave speed. *Comput. Methods Appl. Mech. Engrg.* **194** (2005), no. 2-5, 367-381.

[98] H. Li, H. Ma, and W. Sun, Legendre spectral Galerkin method for electromagnetic scattering from large cavities. *SIAM J. Numer. Anal.* **51** (2013), no. 1, 353-376.

[99] W. R. Madych, Finite orthogonal transforms and multiresolution analyses on intervals. *J. Fourier Anal. Appl.* **3** (1997), 257–294.

[100] R. Masson, Biorthogonal spline wavelets on the interval for the resolution of boundary problems. *Math. Models Methods Appl. Sci.* **6** (1996), 749–791.

[101] W. McLean, Strongly elliptic systems and boundary integral equations. Cambridge University Press, Cambridge, UK, 2000.

[102] Y. Meyer, Wavelets on the interval. *Rev. Mat. Iberoamericana* **7** (1991), 115–133.

[103] J. M. Melenk, On generalized finite-element methods, PhD thesis at the University of Maryland, College Park, 1995.

[104] J. M. Melenk and I. M. Babuška, The partition of unity finite element method: basic theory and applications. *Comput. Methods Appl. Mech. Engrg.* **139** (1996), no. 1-4, 289-314.

[105] J. M. Melenk and S. Sauter, Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation. *SIAM J. Numer. Anal.* **49** (2011), no. 3, 1210-1243.

[106] D. A. Mitsoudis, Ch. Makridakis, and M. Plexousakis, Helmholtz equation with artificial boundary conditions in a two-dimensional waveguide. *SIAM J. Math. Anal.* **44** (2012), no. 6, 4320-4344.

[107] A. Moiola and E. A. Spence, Acoustic transmission problems: wavenumber-explicit bounds and resonance-free regions. *Math. Models Methods Appl. Sci.* **29** (2019), no. 2, 317-354.

[108] J.-C. Nédélec, Acoustic and electromagnetic equations. Integral representations for harmonic problems, *Applied Mathematical Sciences* **144**. Springer-Verlag, New York, 2001. x+316 pp.

[109] D. P. Nicholls and J. Shen, A stable high-order method for two-dimensional bounded-obstacle scattering, *SIAM J. Sci. Comput.* **28** (2006), no. 4, 1398-1419.

[110] K. Pan, D. He, and Z. Li, A high order compact FD framework for elliptic BVPs involving singular sources, interfaces, and irregular domains. *J. Sci. Comput.* **88** (2021), no. 67, 1-25.

[111] O. Pembery, The Helmholtz Equation in Heterogeneous and Random Media: Analysis and Numerics, PhD Thesis at the University of Bath, (2020).

[112] D. Peterseim, Eliminating the pollution effect in Helmholtz problems by local subscale correction. *Math. Comp.* **86** (2017), no. 305, 1005-1036.

[113] G. Plonka, K. Selig, and M. Tasche, On the construction of wavelets on a bounded interval. *Adv. Comput. Math.* **4** (1995), 357–388.

[114] M. Primbs, New stable biorthogonal spline-wavelets on the interval. *Results Math.* **57** (2010), 121–162.

[115] J. Shen, T. Tao, and L. Wang, Spectral methods: algorithms, analysis, and applications. *Springer Series in Computational Mathematics*, **41**. Springer-Verlag, Berlin, Heidelberg, 2011. xvi + 470 pp.

[116] J. Shen and L. Wang. Spectral approximation of the Helmholtz equation with high wave numbers. *SIAM J. Numer. Anal.* **43** (2005), no. 2, 623-644.

[117] J. Shen and L. Wang, Analysis of a spectral-Galerkin approximation to the Helmholtz equation in exterior domains. *SIAM J. Numer. Anal.* **45** (2005), no. 5, 1954-1978.

[118] E. A. Spence, Wavenumber-explicit bounds in time-harmonic acoustic scattering. *SIAM J. Math. Anal.* **46** (2014), no. 4, 2987-3024.

[119] C. C. Stolk, M. Ahmed, and S. K. Bhowmik, A multigrid method for the Helmholtz equation with optimized coarse grid corrections. *SIAM J. Sci. Comput.* **36** (2014), no. 6, A2819-A2841.

[120] E. Turkel, D. Gordon, R. Gordon, and S. Tsynkov, Compact 2D and 3D sixth order schemes for the Helmholtz equation with variable wave number. *J. Comp. Phys.* **232** (2013), no. 1, 272-287.

[121] D. Wang, R. Tezaur, J. Toivanen, and C. Farhat, Overview of the discontinuous enrichment method, the ultra-weak variational formulation, and the partition of unity method for acoustic scattering in the medium frequency regime and performance comparisons. *Int. J. Numer. Meth. Engng.* **89** (2012), no. 4, 403-417.

[122] K. Wang and Y. S. Wong, Pollution-free finite difference schemes for non-homogeneous Helmholtz equation. *Int. J. Numer. Anal. Mod.* **11** (2014), no. 4, 787-815.

[123] K. Wang and Y. S. Wong, Is pollution effect of finite difference schemes avoidable for multi-dimensional Helmholtz equations with high wave numbers? *Commun. Comput. Phys.* **21** (2017), no. 2, 490-514.

[124] Y. S. Wong and G. Li, Exact finite difference schemes for solving Helmholtz equation at any wave number. *Int. J. Numer. Anal Mod.* **2** (2011), no. 1, 91-108.

[125] T. Wu and R. Xu, An optimal compact sixth-order finite difference scheme for the Helmholtz equation. *Comput. Math. Appl.* **75** (2018), no. 7, 2520-2537.

[126] Y. Zhang, K. Wang, and R. Guo, Sixth-order finite difference scheme for the Helmholtz equation with inhomogeneous Robin boundary condition. *Adv. Differ. Equ.* **362** (2019), 1-15.