Karlsruher Institut für Technologie

Efficient Methods for Solving Direct and Inverse Scattering Problems in Locally Perturbed Periodic Structures

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

von der KIT-Fakultät für Mathematik
des Karlsruher Instituts für Technologie (KIT)
genehmigte

Dissertation

von

Nasim Shafieeabyaneh

Referent:	PD. Dr. Tilo Arens
Korreferent:	Prof. Dr. Roland Griesmaier
Korreferentin:	Jun.-Prof. Dr. Ruming Zhang

Tag der mündlichen Prüfung: 26.11.2025

Acknowledgments

I would like to express my gratitude to the many people who supported me over the past few years. First, I would like to thank PD Dr. Tilo Arens for being an exceptional supervisor, for carefully reading my work, providing valuable feedback, and offering support in many ways. For sure, this work would not have been possible without his advice and encouragement. Special thanks to my second supervisor, Jun.-Prof. Dr. Ruming Zhang, for her support and contributions during her time in Karlsruhe, as well as for taking the time to meet with me online from Berlin. I would also like to thank Prof. Dr. Roland Griesmaier for carefully reviewing my thesis and providing valuable comments that greatly improved it.

I am very thankful to PD Dr. Frank Hettlich for the pleasant collaboration in teaching and for always being willing to help. Many thanks to Sonja Becker for her constant kindness, positive energy, and the warmth she brought to my daily work.

I would like to thank the rest of my working group for the many enjoyable times we spent together. Special wonderful moments were shared with Lisa and Marvin, whose kindness and friendship – whether nearby or from afar – made my PhD journey feel warm and joyful. My next thanks go to Eliane, for our weekly runs in the Schlossgarten – filled with discussions about math and life – which provided refreshing breaks throughout my thesis journey. Moreover, I highly appreciate the help of Eliane, Lisa, Leonie and Raphael, who kindly contributed by proofreading this thesis.

I am deeply grateful to my parents, brother, and uncle for their constant support, even from far away. Their love and care made the distance feel much smaller. Last but certainly not least, I would like to thank Francesco for his support and understanding throughout every challenge and for sharing countless unforgettable moments along the way.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project IDs 258734477 — SFB 1173 and 433126998.

Abstract

We consider acoustic scattering of a non-periodic incident field by locally perturbed periodic structures. Our goal is to propose an efficient, high-order numerical method for solving such direct scattering problems.

As a first step, we focus on purely periodic domains. Here, the non-periodicity of the incident field prevents classical reduction to a bounded cell. However, due to the periodicity of the domain, we can apply the Floquet–Bloch (FB) transform. This yields a decoupled family, indexed by the Floquet parameter, of periodic problems posed in a single bounded cell. As our first main result, we derive a representation of the transformed solution that highlights the structure of its singularities with respect to the Floquet parameter. This allows us to develop a tailor-made numerical scheme adapted to the singularities. For locally perturbed periodic structures, the direct application of the FB transform is not possible due to the lack of periodicity in the domain. To address this issue, we employ a coordinate transformation that eliminates the perturbation, resulting in an equation with non-constant coefficients. This reformulation enables the use of the FB transform, but introduces a coupling in the resulting family of problems. Proposing a tailored numerical method here significantly increases computational complexity due to the coupling. To improve efficiency, we approximate the solution using the perfectly matched layer (PML). We prove exponential convergence of the PML approximation of the solution, with respect to the PML parameter, on every compact set. We also show that the PML approximation of the transformed solution is analytic with respect to the Floquet parameter. Therefore, this allows us to compute solutions of original scattering problems by considering fewer members of this family. Furthermore, we propose a fast and parallel solver using recursive Schur complements.

Finally, we apply our fast direct solver to inverse scattering problems in order to reconstruct unknown perturbations. To employ iterative regularization schemes, we prove that the scattered field is Fréchet differentiable with respect to the perturbation. Through numerical examples, we demonstrate the efficacy of our methods.

Contents

CHAPTER 1

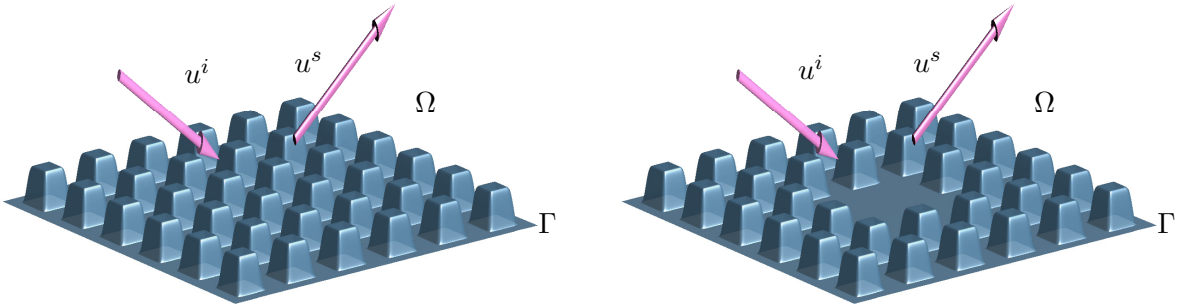
INTRODUCTION

Scattering theory generally describes how waves are affected by irregularities or obstacles in their path. In particular, an incident wave encounters an obstacle and interacts with its boundary or the medium in its interior. This interaction generates a scattered field, whose behavior depends on the incident field as well as the shape and physical properties of the obstacle. The objective in *direct scattering problems* is to determine the scattered field from this known information. Mathematically, the scattered field satisfies a partial differential equation in the exterior domain, together with boundary conditions (e.g., sound-soft or sound-hard conditions) imposed on the boundary of the obstacle. Additionally, a radiation condition must be imposed at infinity to guarantee uniqueness.

In literature, obstacles are generally classified into two main categories: bounded and unbounded. For *bounded obstacles*, the scattering problem is well-understood. The scattered field behaves asymptotically like an outgoing spherical wave. This behaviour in the acoustic case is modeled by the *Sommerfeld radiation condition*. To compute the scattered field numerically, the unbounded exterior domain can be truncated in the radial direction by standard techniques like *perfectly matched layer (PML)* or *Dirichlet-to-Neumann (DtN) map*. For an overview over these types of problems, we refer to monographs by Colton and Kress [31], Monk [90] and Kirsch and Hettlich [70].

In contrast to bounded obstacles, scattering by *unbounded structures* involves obstacles whose geometry extends to infinity in one or more directions. Such structures include examples like rough surfaces [100], open waveguides [67, 97], and periodic media [35]. Since these obstacles lack compact boundaries, the analysis and numerical treatment present additional challenges. Specifically, the classical Sommerfeld radiation condition is usually not applicable, and standard approaches for proving existence and uniqueness of solutions — such as Fredholm theory and Rellich's lemma — may be unusable, depending on the geometry and boundary conditions. To obtain a bounded computational domain for numerical simulation, the unbounded structure must be truncated typically both horizontally and vertically.

Motivated by these challenges, in this work we focus on *time-harmonic acoustic waves* interacting with *unbounded surfaces*. More precisely, let the unbounded scatterer Γ be represented as the



(a) Periodic structure (b) Perturbed structure

FIGURE 1.1. Two possible structures of the unbounded domain Ω.

graph of a function $\zeta\colon \mathbb{R}^{d-1} \to \mathbb{R}$ for $d = 2, 3$, i.e.,

$$\Gamma \coloneqq \left\{(\widetilde{x}, \zeta(\widetilde{x})) : \widetilde{x} \in \mathbb{R}^{d-1}\right\}.$$

Moreover, we define the unbounded domain Ω as the region lying above the surface Γ,

$$\Omega \coloneqq \left\{(\widetilde{x}, x_d) : \widetilde{x} \in \mathbb{R}^{d-1},\, x_d > \zeta(\widetilde{x})\right\}.$$

The acoustic scattering phenomenon is governed by the *Helmholtz equation* with the wave number $k > 0$, i.e.,

$$\Delta u^s + k^2 u^s = 0 \qquad \text{in } \Omega\,, \tag{1.1}$$

where u^s denotes the scattered field. In the case of sound-soft scattering, the scattered field corresponding to the incident field u^i satisfies the boundary condition

$$u^s = -u^i \qquad \text{on } \Gamma\,. \tag{1.2}$$

The formulation of the scattering problem is not complete without a radiation condition, which ensures that the solution is unique and physically meaningful. We consider the *upward propagating radiation condition*, which guarantees that the scattered field u^s is propagating upwards from Γ. For a detailed analysis, we refer to Chandler-Wilde and Monk [25] and Arens and Hohage [4]. This radiation condition is equivalent to a transparent boundary condition on a flat surface above the scatterer Γ. It allows waves to pass through the flat surface without any reflection. This boundary condition can be formulated using the *DtN map* and additionally enables us to truncate the domain vertically (see [23, 25]).

In this work, we concentrate specifically on wave scattering by *periodic surfaces* (see Figure 1.1a), which may include *localized perturbations* (see Figure 1.1b). The study of wave propagation in periodic media has its roots in the work of Lord Rayleigh, who conducted one of the first analyses of diffraction by gratings. Since then, it has become an important topic of modern mathematical physics, with applications in thin solar cells, photonic crystals, and organic LED optimization

(see [3, 12, 61, 62]). Various numerical approaches have been used to study scattering in such media, including recursive doubling [37, 38, 102] and propagation techniques [44, 63]. These methods are typically designed for specific periodic geometries, limiting their extension to more general or locally perturbed structures. Furthermore, in the special case where the incident field is (quasi-)periodic and the scatterer is sufficiently smooth and periodic, the scattering problem can be directly reduced to a single bounded cell of periodicity (see, e.g., [66, 96]). This cell problem can then be solved efficiently using well-established numerical techniques, such as integral equations [89] or finite element methods [10]. However, in more general situations, where either the incident field or the surface is not periodic, this direct reduction no longer works. As a result, the development of novel numerical schemes is essential to efficiently solve these problems.

One way to tackle such problems is to use the *Floquet–Bloch (FB) transform*, which was introduced in [15, 46], with further analysis by Kuchment [77] and Lechleiter [81]. This transform decomposes the original problem in the unbounded periodic domain into a family of *decoupled* periodic problems indexed by the Floquet parameter. Since these problems involve only periodic fields, they can be formulated in a *single bounded cell*. Each of them depends only on the spatial variable and can be solved by a classical numerical method. We call the solutions of these problem the *transformed fields*.

In this procedure, the numerical error is a combination of two components: the error of the spatial discretization and the error resulting from the approximation of the inverse FB transform. In literature, this transform has most often been applied to two-dimensional scattering problems. Detailed numerical results can be found in [29, 55, 83, 85], while theoretical analyses are provided in [81, 82]. However, the application of FB transform in three dimensions can only be found in [73, 84], where the convergence of the numerical method with respect to the Floquet parameter remains relatively slow. In this work, we present a high-order, efficient method for inverting the FB transform, significantly improving the convergence rate compared to existing approaches. This requires proving regularity properties of the transformed field with respect to the Floquet parameter: the inverse FB transform essentially consists of an integral of the transformed field over a bounded domain, but the integrand has a particular structure of singularities. Based on these regularity results, one of the main results presented in this work is a tailor-made quadrature rule to numerically obtain the scattered field of the original non-periodic problem.

Pure periodicity is rarely found in real media; instead, *disruptions* appear in small, localized regions (see [20, 101]). Various methods have been employed to analyze scattering in such media including the Lippmann-Schwinger equation [29], a volume integral approach [55], a perturbation technique [18, 19, 92] and a numerical scheme specially designed to obtain the exact boundary conditions on the vertical segments of a waveguide in [44, 45, 63]. These approaches are applied in the absorbing case (i.e., for complex wave number), which avoids the presence of singularities.

One question that arises is whether our proposed approach for the purely periodic case remains effective when extended to the locally perturbed case. Although approaches based on the FB of the DtN map combined with tailor-made inversion formulas are possible, they require substantial computational effort, especially in three dimensions (see, for example, [6]). Let us now outline some of the challenges.

In the locally perturbed case, applying the FB transform directly is not possible due to the lack of periodicity in the structure. Since the periodic surface and the local perturbation are considered

known, the perturbation can be removed via a *coordinate transformation*. This comes at the cost of dealing with a perturbed Helmholtz equation with non-constant coefficients. However, as the domain becomes periodic, we are able to apply the FB transform. This yields a family of periodic problems in a bounded cell, which are *coupled* because of the variable coefficients. From the regularity analysis of the non-perturbed case, we know that the structure of the DtN map leads to singularities in the transformed field with respect to the Floquet parameter. Therefore, a discretization of the inverse FB transform requires evaluating the transformed field for a large number of Floquet parameters. Furthermore, the coupling prevents solving these problems in parallel, which hence requires a high computational time. This can be understood from [83, 104] for the two-dimensional case and from [6] for the three-dimensional case. Hence the necessity of proposing a fast solver for such scattering problems becomes more pronounced.

To improve efficiency, inspired by [26, 71], we use the PML instead of the DtN map to truncate the domain in the vertical direction. The PML was originally introduced by Berenger for electromagnetic waves [13]. Since then, it has been widely applied in various wave propagation problems. These include scattering by bounded obstacles [28, 30, 33], rough surface scattering [26], electromagnetic optics [90], and seismology [43].

Applying the FB transform to the PML-truncated problem has the advantage that the transformed field depends analytically on the Floquet parameter. This enables us to evaluate the inverse FB transform accurately from relatively few values of the Floquet parameter. However, setting up and solving the discretized system directly is still time-consuming due to the coupling. Using the Schur complement recursively allows us to rewrite the complete system in such a way that the matrix-vector multiplications are reduced to sums of terms that can be evaluated independently. Therefore, we can benefit greatly from parallelizing these evaluations by solving the linear system with an iterative method.

From the theoretical point of view, the convergence rate of the PML has been proven to be globally linear with respect to the PML parameter for rough surfaces [26]. Additionally, it has been shown that for flat scatterers the convergence is exponential in every compact set. In the conclusion of [26], the question of whether the local exponential convergence can be extended to rough surfaces was stated as an open problem. A partial answer has been provided in [28, 105] by proving the local exponential convergence in the pure periodic case. We extend the exponential convergence results of [26, 105] to the locally perturbed case in two dimensions.

So far, we have assumed a known, local perturbation (or defect) in a periodic structure and focused on computing the scattered field. However, detecting and reconstructing such localized defects is critical for optimizing the performance of devices based on periodic media (see [20, 50, 101]). These defects can be viewed as perturbations of the pure periodic structure. We now consider the *inverse scattering problem*: An incident field is directed into the medium, and the scattered field is measured at multiple observation points above the scatterer. Using these measurements, the goal is to detect or reconstruct the unknown perturbation on the periodic surface.

To detect the support of the perturbation, we refer to [86] for a linear sampling method and [17] for a factorization method when the periodic background is known. Furthermore, approaches that require less a priori knowledge of periodic structures have been developed in [21, 22]. However, each of these methods requires sending and measuring waves from all directions.

In this thesis, we make the assumption that the location of the perturbation is known. Our focus is only on the reconstruction of the perturbation using the measured data from a single incident field.

This inverse problem can be formulated as an optimization problem, described as follows: Among the set of admissible perturbations X, find the optimal perturbation $\delta^* \in X$, for which

$$\delta^* = \arg\min_{\delta \in X} J(\delta)\,,$$

where $J\colon X \to \mathbb{R}$ is the objective function depending on the measured data. Computing J often involves solving a direct scattering problem, where $\delta \in X$ modifies the domain in which the equation is posed. At this point, we profit from our fast direct solver as the key component of a method to reconstruct the perturbation.

One question in inverse problems is how to ensure that the measured data uniquely determines the perturbation. As mentioned in [11, 40, 57], when the wave number is real, establishing global uniqueness using only a single incident plane wave remains an open problem. So far, uniqueness results have been proven under certain assumptions on the regularity of the structure (see [1, 11, 57, 69] for sufficiently smooth periodic structures and [39, 40, 42] for polygonal periodic structures). For local perturbations, uniqueness from measured data generated by point source waves has been established in [59], using a finite number of incident fields when the defect's size and height are known; otherwise, infinitely many incident fields are required.

Furthermore, such inverse problems are *ill-posed*, meaning that small changes in the measured data can lead to large errors in the estimated position and shape of perturbations. Consequently, the resulting optimization problems are also unstable. One common approach to solving such ill-posed problems involves *iterative regularization methods*. Some of these methods are formulated using derivatives with respect to boundary variations (see [56, 86]). Therefore, an essential preliminary step is to prove that the objective function is Fréchet differentiable with respect to the boundary. This has been established in [66] for (quasi-)periodic scattering problems with respect to periodic surfaces. In this work, we extend these results to non-periodic scattering problems.

The Fréchet derivative can be computed by solving an additional boundary value problem for each admissible perturbation. To apply a Newton–type algorithm, several admissible perturbations must be considered in each iteration. Therefore, using a fast direct solver for the scattering problem significantly improves the performance of the reconstruction algorithm. Thus, we can use the fast iterative solver proposed in an earlier section of the thesis.

1.1. Outline of the Thesis

In Chapter 2, we introduce and discuss some fundamental tools that are essential for the entire thesis. These include function spaces on unbounded structures, two vertical truncation methods — the DtN map and the PML — as well as the FB transform. We also review the properties of the FB transform and describe how it decomposes a non-periodic problem in a periodic domain into a family of periodic problems in a single bounded cell.

Chapter 3 is devoted to solving non-periodic scattering problems in unbounded periodic

structures. We truncate the domain in the vertical direction using the DtN map. Then, we apply the FB transform to obtain a decoupled family of problems in a bounded cell. From the theoretical point of view, we analyze the regularity of the transformed field with respect to the Floquet parameter. For this purpose, we provide a representation reflecting the expected structure of singularities, which arise from the DtN map. We propose a tailor-made quadrature rule adapted to the singularities of the transformed field, which allows us to evaluate the inverse FB transform with higher accuracy. We additionally obtain an error estimate of the proposed numerical approach. Numerical examples demonstrating the performance of the proposed scheme are included.

In Chapter 4, we focus on solving non-periodic scattering problems in locally perturbed periodic structures. We introduce two formulations of the truncated problem: one based on the exact DtN map and the other based on its PML approximation. Afterwards, we restore periodicity of the domain via a coordinate transformation, which allows us to apply the FB transform. This yields a coupled family of periodic problems posed in a bounded cell. We prove the unique solvability of the PML-truncated problem. Moreover, we show that, in two dimensions, the PML approximation converges exponentially to the solution of the DtN-truncated problem, with respect to the PML parameter, on every compact set. To numerically compute the PML approximation of the scattered field, we propose a fast iterative solver. At each iteration, the matrix-vector products corresponding to different Floquet parameters can be evaluated in parallel. As a conclusion, by using this technique, we are able to significantly reduce the computational time. Some numerical results illustrate the efficiency and the convergence rate of the proposed method.

Chapter 5 is concerned with solving an inverse scattering problem for compactly supported perturbations. The objective is to reconstruct the unknown perturbation from near-field observations corresponding to a non-periodic incident field. This requires inverting the scattering functional, which maps the perturbation to the observed scattered field. This inverse problem can be framed as an optimization problem. To solve it with a Newton–type method, we prove the differentiability of the scattering operator and thus obtain its Fréchet derivative. To stabilize the optimization problem, we introduce a penalty term and determine its Fréchet derivative. We bring together all these requirements to establish an efficient Gauss–Newton algorithm to reconstruct the unknown perturbation. Numerical results demonstrating the performance of the proposed method are provided.

The final part of this thesis contains three appendices: Appendix A provides the technical computations necessary for Chapters 3 and 4. Appendix B compares the computational cost of the proposed iterative method in Chapter 4 with the direct solver in [98, Thm. 2.2]. Appendix C summarizes useful properties of Green's function.

1.2. PRIOR PUBLICATION

Some results of Chapter 3 have already been published in [6].

CHAPTER 2

Fundamental Tools

In this chapter, we collect some mathematical tools which are required for analyzing and solving scattering problems in unbounded domains. We start with the definition of some useful function spaces. Afterwards, we explain some approaches to reformulate the problem in a bounded domain.

2.1. Sobolev Spaces

To construct Sobolev spaces of non-integer order, we follow [88] and begin with the *Schwartz space* of rapidly decreasing complex-valued C^∞ functions (see [88, p. 72]).

Definition 2.1. The Schwartz space $\mathcal{S}(\mathbb{R}^d)$ is defined by

$$\mathcal{S}(\mathbb{R}^d) \coloneqq \left\{ \phi \in C^\infty(\mathbb{R}^d) : \sup_{x\in\mathbb{R}^d} \left| x^\alpha \partial^\beta \phi(x) \right| < \infty \text{ for all multi-indices } \alpha, \beta \in \mathbb{N}^d \right\}.$$

For all multi-indices α and β, we consider the semi-norms $|\phi|_{\alpha,\beta} \coloneqq \sup_{x\in\mathbb{R}^d} |x^\alpha \partial^\beta \phi(x)|$. A sequence $\{\phi_j\}_{j\in\mathbb{N}} \subseteq \mathcal{S}(\mathbb{R}^d)$ is said to converge to $\phi \in \mathcal{S}(\mathbb{R}^d)$ if $|\phi_j - \phi|_{\alpha,\beta} \to 0$ as $j \to \infty$ for all α and β.

That means, this space consists of smooth functions whose derivatives, as well as the function itself, decay at infinity faster than any polynomial. The topology of this space is induced by the countable family of semi-norms $|\phi|_{\alpha,\beta}$.

Now, we introduce the Fourier transform of functions in the space $\mathcal{S}(\mathbb{R}^d)$ as in [88, p. 72].

Definition 2.2. The Fourier transform $\mathcal{F}\colon \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ is given by

$$(\mathcal{F}\phi)(\xi) \coloneqq (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i}\xi\cdot x} \phi(x)\,\mathrm{d}x \qquad \text{for all } \xi \in \mathbb{R}^d,$$

with the inverse Fourier transform

$$(\mathcal{F}^{-1}\varphi)(x) \coloneqq (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\xi\cdot x} \varphi(\xi)\,\mathrm{d}\xi \qquad \text{for all } x \in \mathbb{R}^d.$$

Since $(\mathcal{F}(\partial^\alpha \phi))(\xi) = (\mathrm{i}\xi)^\alpha (\mathcal{F}\phi)(\xi)$ and $\mathcal{F}((-\mathrm{i}x)^\alpha \phi(x)) = \partial^\alpha(\mathcal{F}\phi)$, the action of the Fourier transform on the functions in the Schwartz space is well defined and continuous.

Remark 2.3. A straightforward consequence of [88, Cor. 3.5 and Thm. 3.12] is that the Fourier transform can be extended by density to an isometry $\mathcal{F}\colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$.

The dual of the Schwartz space, denoted by $\mathcal{S}^*(\mathbb{R}^d)$, is known as the space of *temperate distributions*, which contains all continuous linear functionals on $\mathcal{S}(\mathbb{R}^d)$. The duality pairing between these spaces is written as $\langle \cdot, \cdot \rangle_{\mathcal{S}^*(\mathbb{R}^d)\times\mathcal{S}(\mathbb{R}^d)}$. For readability, we henceforth use the simplified notation $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$.

The Fourier transform can be extended by duality to an operator $\mathcal{F}\colon \mathcal{S}^*(\mathbb{R}^d) \to \mathcal{S}^*(\mathbb{R}^d)$, i.e., for $\phi \in \mathcal{S}^*(\mathbb{R}^d)$

$$\langle \mathcal{F}\phi, \psi \rangle_{\mathbb{R}^d} \coloneqq \langle \phi, \mathcal{F}\psi \rangle_{\mathbb{R}^d} \quad \text{for all } \psi \in \mathcal{S}(\mathbb{R}^d)\,.$$

Using the recalled preliminaries, we define Sobolev spaces of non-integer order as in [88, p. 76].

Definition 2.4. The Sobolev space $H^s(\mathbb{R}^d)$ of order $s \in \mathbb{R}$ is defined by

$$H^s(\mathbb{R}^d) \coloneqq \left\{ \phi \in \mathcal{S}^*(\mathbb{R}^d) : (1 + |\cdot|^2)^{s/2} \mathcal{F}\phi \in L^2(\mathbb{R}^d) \right\}.$$

This space is a Hilbert space equipped with the inner product

$$\left\langle \phi, \overline{\psi} \right\rangle_{H^s(\mathbb{R}^d)} \coloneqq \left\langle (1 + |\cdot|^2)^{s/2} \mathcal{F}\phi, \overline{(1 + |\cdot|^2)^{s/2} \mathcal{F}\psi} \right\rangle_{L^2(\mathbb{R}^d)},$$

which induces the norm

$$\|\phi\|_{H^s(\mathbb{R}^d)} \coloneqq \left\| (1 + |\cdot|^2)^{s/2} \mathcal{F}\phi \right\|_{L^2(\mathbb{R}^d)}.$$

The corresponding weighted Sobolev space can be defined based on [81, Sec. 3].

Definition 2.5. The *weighted Sobolev space* $H^s_r(\mathbb{R}^d)$, for any $s, r \in \mathbb{R}$, is given by

$$H^s_r(\mathbb{R}^d) \coloneqq \left\{ \phi \in \mathcal{S}^*(\mathbb{R}^d) : (1 + |\cdot|^2)^{r/2} \phi \in H^s(\mathbb{R}^d) \right\}$$

and it is equipped with the norm

$$\|\phi\|_{H^s_r(\mathbb{R}^d)} \coloneqq \left\| (1 + |\cdot|^2)^{r/2} \phi \right\|_{H^s(\mathbb{R}^d)}.$$

For $r \in \mathbb{R}$, $L^2_r(\mathbb{R}^d) \coloneqq H^0_r(\mathbb{R}^d)$. The dual of $H^s_r(\mathbb{R}^d)$ is $H^{-s}_{-r}(\mathbb{R}^d)$.

Remark 2.6. If $s \in \mathbb{N}$, the following is an equivalent definition for the latter norm

$$\|\phi\|^2_{H^s_r(\mathbb{R}^d)} = \sum_{m \in \mathbb{N}^d, |m| \le s} \left\| \partial^m \left((1 + |\cdot|^2)^{r/2} \phi \right) \right\|^2_{L^2(\mathbb{R}^d)}.$$

We now introduce a class of subspaces of $H^s(\mathbb{R}^d)$, namely Sobolev spaces containing α-quasiperiodic functions for $\alpha \in \mathbb{R}^d$ as in [81, Eq. (8)]. This includes periodic functions, corresponding to $\alpha = 0$.

Definition 2.7. For a given $\alpha \in \mathbb{R}^d$, a function $\phi\colon \mathbb{R}^d \to \mathbb{R}^d$ is an α-quasiperiodic function with fundamental period $L > 0$ if

$$\phi(x + Lj) = \mathrm{e}^{\mathrm{i}L\alpha\cdot j}\phi(x) \quad \text{for } x \in \mathbb{R}^d,\ j \in \mathbb{Z}^d\,.$$

The corresponding Sobolev space $H^s_\alpha(\mathbb{R}^d)$ for $s \in \mathbb{R}$ is given by

$$H^s_\alpha(\mathbb{R}^d) \coloneqq \left\{\phi \in H^s(\mathbb{R}^d) : \phi(x + Lj) = \mathrm{e}^{\mathrm{i}L\alpha\cdot j}\phi(x) \text{ for } x \in \mathbb{R}^d,\ j \in \mathbb{Z}^d\right\},$$

equipped with the norm

$$\|\phi\|^2_{H^s_\alpha(\mathbb{R}^d)} \coloneqq \sum_{j\in\mathbb{Z}^d} (1 + |j|^2)^s \left|\widehat{\phi_\alpha}(j)\right|^2,$$

where $\widehat{\phi_\alpha}(j)$ denotes the j-th Fourier coefficient of $\phi(x)\mathrm{e}^{-\mathrm{i}\alpha\cdot x}$, i.e.,

$$\widehat{\phi_\alpha}(j) \coloneqq L^{-d/2} \int_{[-L/2,L/2]^d} \phi(x)\, \mathrm{e}^{-\mathrm{i}\alpha\cdot x}\mathrm{e}^{-\mathrm{i}(2\pi/L)j\cdot x}\,\mathrm{d}x\,.$$

For $s < 0$, the integral above can be understood as a dual pairing.

Remark 2.8. Note that an α-quasiperiodic function becomes periodic when multiplied by $\mathrm{e}^{-\mathrm{i}\alpha\cdot x}$. For $\alpha = 0$, we write $H^s_{\mathrm{per}}(\mathbb{R}^d)$ instead of $H^s_0(\mathbb{R}^d)$ to emphasize the periodicity.

Lemma 2.9. *For $\alpha \in \mathbb{R}^d$, let $\mathcal{M}_\alpha\colon \phi \mapsto \phi\mathrm{e}^{\mathrm{i}\alpha\cdot x}$. If $\phi \in H^s_{\mathrm{per}}(\mathbb{R}^d)$, then $\mathcal{M}_\alpha\phi \in H^s_\alpha(\mathbb{R}^d)$ and*

$$\|\phi\|_{H^s_{\mathrm{per}}(\mathbb{R}^d)} = \|\mathcal{M}_\alpha\phi\|_{H^s_\alpha(\mathbb{R}^d)}\,.$$

Proof. Using the definition of these norms yields

$$\|\mathcal{M}_\alpha\phi\|^2_{H^s_\alpha(\mathbb{R}^d)} = L^{-d}\sum_{j\in\mathbb{Z}^d} (1 + |j|^2)^s \left|\int_{[-L/2,L/2]^d} \mathcal{M}_{-\alpha}\mathcal{M}_\alpha\phi\, \mathrm{e}^{-\mathrm{i}(2\pi/L)j\cdot x}\,\mathrm{d}x\right|^2 = \|\phi\|^2_{H^s_{\mathrm{per}}(\mathbb{R}^d)}\,.$$

□

So far, we have defined Sobolev spaces on the full space $\mathbb{R}^d$. For any non-empty open set $\Omega \subseteq \mathbb{R}^d$, we define (see [88, p. 77])

$$H^s_r(\Omega) \coloneqq \left\{\varphi = \phi|_\Omega\colon \phi \in H^s_r(\mathbb{R}^d)\right\},$$

equipped with the norm

$$\|\varphi\|_{H^s_r(\Omega)} \coloneqq \inf_{\substack{\phi\in H^s_r(\mathbb{R}^d)\\ \phi|_\Omega=\varphi}} \|\phi\|_{H^s_r(\mathbb{R}^d)}\,.$$

Let $C^{m-1,1}$ for $m \in \mathbb{N}$ be the set of functions whose $(m-1)$-th derivative is Lipschitz (see [88, p. 90]). Let the boundary of the domain Ω be denoted by $\partial\Omega = \overline{\Omega} \setminus \Omega$ and assume that it is the graph of a $C^{m-1,1}$-function ζ for $m \in \mathbb{N}$. For $\phi\colon \partial\Omega \to \mathbb{C}$, we define $\phi_\zeta\colon \mathbb{R}^{d-1} \to \mathbb{C}$ by

$$\phi_\zeta(x) \coloneqq \phi(x, \zeta(x)) \quad \text{for } x \in \mathbb{R}^{d-1}\,.$$

Now, Sobolev spaces on the boundary $\partial\Omega$ are given by (see [88, pp. 98-99])

$$L^2_r(\partial\Omega) \coloneqq \left\{\phi \in L^2_{\mathrm{loc}}(\partial\Omega) : \phi_\zeta \in L^2_r(\mathbb{R}^{d-1})\right\}, \quad \text{for } r \in \mathbb{R},$$
$$H^s_r(\partial\Omega) \coloneqq \left\{\phi \in L^2_r(\partial\Omega) : \phi_\zeta \in H^s_r(\mathbb{R}^{d-1})\right\}, \quad \text{for } 0 < s \le m \text{ and } r \in \mathbb{R},$$

equipped with the norm $\|\phi\|_{H^s_r(\partial\Omega)} = \|\phi_\zeta\|_{H^s_r(\mathbb{R}^{d-1})}$. Moreover, by

$$\|\phi\|_{H^{-s}_r(\partial\Omega)} \coloneqq \left\|\sqrt{1+|\nabla\zeta|^2}\phi_\zeta\right\|_{H^{-s}_r(\mathbb{R}^{d-1})} \qquad \text{for } 0 < s \le m,$$

the space $H^{-s}_r(\partial\Omega)$ can be defined as the completion of $L^2(\partial\Omega)$ with respect to this norm.

To define the restriction of a function to the boundary, we make use of the *trace* operator, which is defined as

$$\gamma_D \colon C^\infty_0(\Omega) \to C^\infty_0(\partial\Omega), \quad \gamma_D\phi \coloneqq \phi|_{\partial\Omega}.$$

According to [88, Thm. 3.37], γ_D has a unique bounded extension $\gamma_D \colon H^s_r(\Omega) \to H^{s-1/2}_r(\partial\Omega)$ for $r \in \mathbb{R}$ when the boundary of Ω is a graph of $C^{m-1,1}$-functions and $1/2 < s \le m$. Moreover, this extension is surjective.

Since this operator is bounded, there exists a constant $c > 0$ such that

$$\|\gamma_D\phi\|_{H^{s-1/2}_r(\partial\Omega)} \le c\|\phi\|_{H^s_r(\Omega)} \quad \text{for all } \phi \in H^s_r(\Omega). \tag{2.1}$$

The trace operator allows us to define the Sobolev spaces of functions which are zero on the boundary. For $s \ge 1$, we define

$$\widetilde{H}^s_r(\Omega) \coloneqq \{\phi \in H^s_r(\Omega) : \gamma_D\phi = 0\}.$$

We next introduce the *conormal derivative* which can be used to describe Neumann boundary conditions based on [88, Lem. 4.3, 49, Thm 2.2]. Let

$$H^1(\Delta,\Omega) \coloneqq \left\{\phi \in H^1(\Omega) : \Delta\phi \in L^2(\Omega)\right\},$$

with $\|\phi\|^2_{H^1(\Delta,\Omega)} \coloneqq \|\phi\|^2_{H^1(\Omega)} + \|\Delta\phi\|^2_{L^2(\Omega)}$. Then, there exists a unique bounded linear operator $\gamma_N \colon H^1(\Delta,\Omega) \to H^{-1/2}(\partial\Omega)$ such that Green's first identity is satisfied, i.e.,

$$\left\langle \nabla\phi, \overline{\nabla\psi}\right\rangle_\Omega = \left\langle \Delta\phi, \overline{\psi}\right\rangle_\Omega + \left\langle \gamma_N\phi, \overline{\gamma_D\psi}\right\rangle_{\partial\Omega}$$

for all $\psi \in H^1(\Omega)$. Note that for $\phi \in C^1(\overline{\Omega})$, $\gamma_N\phi = n \cdot \nabla\phi|_{\partial\Omega}$, where n denotes the outward unit normal vector on the surface $\partial\Omega$.

2.2. Upward Propagating Radiation Condition

A common radiation condition for the scattering problem (1.1)–(1.2) in the unbounded domain Ω is to assume that the scattered field is propagating away from the scatterer Γ (represented by

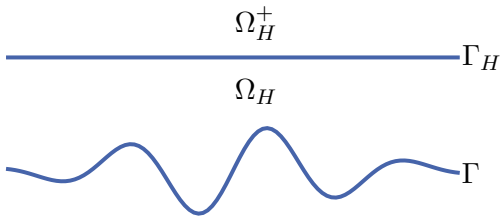


Figure 2.1. A two-dimensional sketch of the unbounded domains Ω_H and Ω_H^+.

the graph of ζ) and does not reflect back into the computational domain Ω.

As explained in [9, 25], a classical approach to deriving this radiation condition is to first introduce a flat surface $\Gamma_H \coloneqq \mathbb{R}^{d-1} \times \{H\}$ for some $H > \|\zeta\|_\infty$. This surface divides the domain Ω into two unbounded regions: the interior domain Ω_H and the exterior domain Ω_H^+, as depicted in Figure 2.1. The next step is to solve the exterior problem in Ω_H^+ together with a boundary condition on Γ_H. This means, for $g \in H_r^{1/2}(\Gamma_H)$ with $|r| < 1$, we need to solve the exterior problem in the weak sense

$$\Delta u^s + k^2 u^s = 0 \qquad \text{in } \Omega_H^+, \tag{2.2a}$$
$$u^s = g \qquad \text{on } \Gamma_H, \tag{2.2b}$$
$$u^s \text{ is outgoing}. \tag{2.2c}$$

Let $\widetilde{x}$ denote the first $d-1$ components of $x \in \mathbb{R}^d$ for $d = 2, 3$. The solution to (2.2) can formally be obtained by applying the Fourier transform with respect to the first $d-1$ variables:

$$(\mathcal{F}u^s)(\xi, x_d) \coloneqq (2\pi)^{-(d-1)/2} \int_{\mathbb{R}^{d-1}} \mathrm{e}^{-\mathrm{i}\widetilde{x}\cdot\xi} u^s(\widetilde{x}, x_d)\,\mathrm{d}\widetilde{x}\,, \quad \xi \in \mathbb{R}^{d-1},\ x_d \geq H\,.$$

Applying the Fourier transform to the Helmholtz equation (2.2a) yields the following ordinary differential equation

$$\partial_{x_d}^2 (\mathcal{F}u^s)(\xi, x_d) + (k^2 - |\xi|^2)(\mathcal{F}u^s)(\xi, x_d) = 0\,, \quad \xi \in \mathbb{R}^{d-1},\ x_d \geq H\,, \tag{2.3}$$

together with the boundary value $u^s(\xi, H) = g(\xi, H)$. The general solution to this ordinary differential equation is given by

$$(\mathcal{F}u^s)(\xi, x_d) = C_1(\xi)\mathrm{e}^{\mathrm{i}\sqrt{k^2-|\xi|^2}x_d} + C_2(\xi)\mathrm{e}^{-\mathrm{i}\sqrt{k^2-|\xi|^2}x_d}\,, \quad \xi \in \mathbb{R}^{d-1},\ x_d \geq H \tag{2.4}$$

for some complex-valued coefficients C_1 and C_2. Note that when $|\xi| < k$, the value of $\sqrt{k^2 - |\xi|^2}$ is a positive real number, hence both $\exp(\mathrm{i}\sqrt{k^2 - |\xi|^2}x_d)$ and $\exp(-\mathrm{i}\sqrt{k^2 - |\xi|^2}x_d)$ are oscillatory. On the other hand, when $|\xi| > k$, the value of $\sqrt{k^2 - |\xi|^2}$ is purely imaginary, which implies that $\exp(\mathrm{i}\sqrt{k^2 - |\xi|^2}x_d)$ is exponentially decaying and $\exp(-\mathrm{i}\sqrt{k^2 - |\xi|^2}x_d)$ is exponentially growing as x_d increases. Since the solution with respect to x_d is assumed to be outgoing and $\exp(-\mathrm{i}\sqrt{k^2 - |\xi|^2}x_d)$ is incoming, we conclude that $C_2 = 0$. By imposing the boundary condition,

we obtain the following representation

$$(\mathcal{F}u^s)(\xi, x_d) = (\mathcal{F}g)(\xi, H)\mathrm{e}^{\mathrm{i}\sqrt{k^2-|\xi|^2}(x_d-H)}\,, \quad \xi \in \mathbb{R}^{d-1},\ x_d \geq H\,,$$

in which $\sqrt{k^2-|\xi|^2}$ has a non-negative imaginary part. Thus, by applying the inverse Fourier transform, we recover the solution for $x_d \geq H$ by

$$u^s = \mathcal{D}g\,,$$

where the operator $\mathcal{D}\colon H_r^{1/2}(\Gamma_H) \to H^1_{\mathrm{loc}}(\Omega_H^+)$ is defined by

$$(\mathcal{D}g)(x) := (2\pi)^{-(d-1)/2} \int_{\mathbb{R}^{d-1}} \mathrm{e}^{\mathrm{i}\widetilde{x}\cdot\xi+\mathrm{i}\sqrt{k^2-|\xi|^2}(x_d-H)}(\mathcal{F}g)(\xi, H)\,\mathrm{d}\xi \quad x \in \Omega_H^+\,. \tag{2.5}$$

This operator is well defined (see [23, Sec. 2]) for $|r| < 1$. Moreover, it indicates that the radiating solution u^s is a superposition of the homogeneous and non-homogeneous upward propagating plane waves $\exp(\mathrm{i}\widetilde{x}\cdot\xi + \mathrm{i}\sqrt{k^2-|\xi|^2}(x_d-H))$ for $|\xi| \leq k$ and evanescent waves $\exp(\mathrm{i}\widetilde{x}\cdot\xi - \sqrt{|\xi|^2-k^2}(x_d-H))$ for $|\xi| > k$.

What follows is a summary of some results from [23, Sec. 2] to clarify why the input of the operator $\mathcal{D}$ is a function in the Sobolev space $H_r^{1/2}(\Gamma_H)$ with the weight r restricted to $|r| < 1$. We will first point out in Lemma 2.10 that the integral on the right-hand side of (2.5) exists only for $g \in H_r^{1/2}(\Gamma_H)$ with $r > -1$. However, we will show that the interior boundary value problem in Ω_H together with the upward propagating radiation condition is not solvable in general for $r \geq 1$.

Lemma 2.10. *The integral on the right-hand side of (2.5) exists for all $g \in H_r^{1/2}(\mathbb{R}^{d-1})$ if and only if $r > -1$.*

Proof. We first focus on $r \geq 0$. In this case, it is clear that $H_r^{1/2}(\mathbb{R}^{d-1}) \subseteq H^{1/2}(\mathbb{R}^{d-1}) \subseteq L^2(\mathbb{R}^{d-1})$. Since the Fourier transform is an isometry on $L^2(\mathbb{R}^{d-1})$, we have $\mathcal{F}g \in L^2(\mathbb{R}^{d-1})$ for any $g \in H_r^{1/2}(\mathbb{R}^{d-1})$. Hence, the integral on the right-hand side of (2.5) is well defined in the Lebesgue sense if $f_x \in L^2(\mathbb{R}^{d-1})$, where f_x is defined by

$$f_x(\xi) := \mathrm{e}^{\mathrm{i}\widetilde{x}\cdot\xi+\mathrm{i}\sqrt{k^2-|\xi|^2}(x_d-H)}\,, \quad \xi \in \mathbb{R}^{d-1}\,.$$

This holds because using polar coordinates leads to

$$\begin{aligned}\|f_x\|^2_{L^2(\mathbb{R}^{d-1})} = \int_{\mathbb{R}^{d-1}} \left|f_x\overline{f_x}\right| \mathrm{d}\xi &= \int_{|\xi|\leq k} \mathrm{d}\xi + \int_{|\xi|>k} \left|\mathrm{e}^{-2\sqrt{|\xi|^2-k^2}(x_d-H)}\right| \mathrm{d}\xi \\ &\leq C(k,d) + \int_0^\infty \mathrm{e}^{-2\rho(x_d-H)}\rho^{d-2}\,\mathrm{d}\rho < \infty\,.\end{aligned}$$

Now, it remains to analyze the existence of the integral on the right-hand side of (2.5) for $g \in H_r^{1/2}(\Gamma_H)$ with $r < 0$. In this case, for fixed $x \in \Omega_H^+$, we interpret the mapping $g \mapsto (\mathcal{D}g)(x)$ as a bounded linear functional on $H_r^{1/2}(\Gamma_H)$. Now, we have to establish for which r is possible. As $g \in H_r^{1/2}(\mathbb{R}^{d-1})$, we have $\mathcal{F}g \in H^r_{1/2}(\mathbb{R}^{d-1})$. From the definition of the operator $\mathcal{D}$, we need to prove that $f_x \in H^{-r}_{-1/2}(\mathbb{R}^{d-1})$ which holds only for $r > -1$. To show it, it is sufficient to prove

that $\mathcal{F}f_x \in H^{-1/2}_{-r}(\mathbb{R}^{d-1})$ if and only if $r > -1$.

For this purpose, we first recall from [23, Sec. 2] that for $g \in L^2(\Gamma_H)$, (2.5) is equivalent to

$$(\mathcal{D}g)(x) = 2\int_{\Gamma_H} \partial_{y_d}\Phi(x,y)g(y)\,\mathrm{d}s(y) = 2\int_{\mathbb{R}^{d-1}} \partial_{y_d}\Phi(x,\widetilde{y},H)g(\widetilde{y},H)\,\mathrm{d}\widetilde{y}\,, \tag{2.6}$$

where Φ is the fundamental solution of the Helmholtz equation given by

$$\Phi(x,y) := \begin{cases} \dfrac{\mathrm{i}}{4}H_0^{(1)}(k|x-y|) & \text{if } d=2\,,\\[2ex] \dfrac{1}{4\pi}\dfrac{\mathrm{e}^{\mathrm{i}k|x-y|}}{|x-y|} & \text{if } d=3 \end{cases}$$

and $H_0^{(1)}$ is the Hankel function of the first kind of order zero.

By taking into account that the Fourier transform is unitary for functions in L^2-space and comparing the definition of $\mathcal{D}$ with (2.6), we have

$$(\mathcal{F}f_x)(y) = 2(2\pi)^{(d-1)/2}\partial_{y_d}\Phi(x,y)|_{y_d=H}\,,$$

and according to [27, Eq. (2.4)]

$$|(\mathcal{F}f_x)(y)| \sim c(x_d - H)|y|^{-(1+d)/2} \quad \text{as } |y|\to\infty\,, \tag{2.7}$$

where the constant c depends on the wave number k and the dimension d. Since for $r<0$ we have $L^2_{-r}(\mathbb{R}^{d-1}) \subset H^{-1/2}_{-r}(\mathbb{R}^{d-1})$, it is sufficient to prove that

$$\mathcal{F}f_x \in L^2_{-r}(\mathbb{R}^{d-1}) \quad \text{for } r > -1\,.$$

By using the definition of $L^2_{-r}(\mathbb{R}^{d-1})$ and polar coordinates, we obtain

$$\begin{aligned}\|\mathcal{F}f_x\|^2_{L^2_{-r}(\mathbb{R}^{d-1})} &= C_1 + C_2\int_{\mathbb{R}^{d-1}\setminus B(0,1)}\left|(1+|y|^2)^{-r/2}|y|^{-(1+d)/2}\right|^2\mathrm{d}y\\ &\le C_3\int_{\mathbb{R}^{d-1}\setminus B(0,1)}\left|y^{-r}y^{-(1+d)/2}\right|^2\mathrm{d}y\\ &\le C_3\int_{\mathbb{R}^{d-1}\setminus B(0,1)}|y|^{-2r-1-d}\,\mathrm{d}y \le C_3\int_1^\infty \rho^{-2r-1-d}\rho^{d-2}\,\mathrm{d}\rho\end{aligned}$$

for some constants C_1, C_2 and C_3. This integral exists when $-2r-3<-1$, which is equivalent to $r>-1$.

We still need to show that $\mathcal{F}f_x \notin H^{-1/2}_{-r}(\mathbb{R}^{d-1})$ for $r \le -1$. As

$$H^{-1/2}_{-r}(\mathbb{R}^{d-1}) \subseteq H^{-1}_{-r}(\mathbb{R}^{d-1}) \subseteq H^{-1}_{1}(\mathbb{R}^{d-1})\,,$$

it is enough to prove $\mathcal{F}f_x \notin H^{-1}_1(\mathbb{R}^{d-1})$. The proof is done by contradiction. We assume that

$\mathcal{F}f_x \in H_1^{-1}(\mathbb{R}^{d-1})$. Using the operator norm, we have

$$\|\mathcal{F}f_x\|_{H_1^{-1}(\mathbb{R}^{d-1})} = \sup_{0\neq v\in H_{-1}^1(\mathbb{R}^{d-1})} \frac{|\langle \mathcal{F}f_x, \overline{v}\rangle_{\mathbb{R}^{d-1}}|}{\|v\|_{H_{-1}^1(\mathbb{R}^{d-1})}} \geq \sup_{n\in\mathbb{N}} \frac{|\langle \mathcal{F}f_x, \overline{v_n}\rangle_{\mathbb{R}^{d-1}}|}{\|v_n\|_{H_{-1}^1(\mathbb{R}^{d-1})}},$$

where the sequence $(v_n)_{n\in\mathbb{N}}$ is chosen such that

$$\begin{aligned} v_n(y) &= |y|^{(3-d)/2-1/n} \quad \text{for } |y|>1\,,\\ \|v_n\|_{H_{-1}^1(\mathbb{R}^{d-1})} &< c \quad \text{for } |y|\leq 1\,. \end{aligned}$$

By a straightforward computation, we obtain

$$\|v_n\|^2_{H_{-1}^1(\mathbb{R}^{d-1})} \leq c\int_{\mathbb{R}^{d-1}\setminus B(0,1)} |y|^{-2}|y|^{3-d-2/n}\,\mathrm{d}y = c\int_1^\infty \rho^{-1-2/n}\,\mathrm{d}\rho = c_1 n\,,$$

for a positive constant c_1 independent of $n\in\mathbb{N}$. By using the asymptotic behaviour of $\mathcal{F}f_x$ given in (2.7) and polar coordinate, we arrive at

$$\begin{aligned} |\langle \mathcal{F}f_x, \overline{v_n}\rangle_{\mathbb{R}^{d-1}}| &= \left|\langle \mathcal{F}f_x, \overline{v_n}\rangle_{\mathbb{R}^{d-1}\setminus B(0,1)} + \langle \mathcal{F}f_x, \overline{v_n}\rangle_{B(0,1)}\right|\\ &\geq \left|\langle \mathcal{F}f_x, \overline{v_n}\rangle_{\mathbb{R}^{d-1}\setminus B(0,1)}\right| - \left|\langle \mathcal{F}f_x, \overline{v_n}\rangle_{B(0,1)}\right|\\ &\geq \left|c\int_{\mathbb{R}^{d-1}\setminus B(0,1)} |y|^{-(1+d)/2}\,\overline{v_n(y)}\,\mathrm{d}y\right| - \|\mathcal{F}f_x\|_{H_1^{-1}(\mathbb{R}^{d-1})}\|v_n\|_{H_{-1}^1(B(0,1))}\\ &\geq \widetilde{c} + c\left|\int_1^\infty \rho^{-1-1/n}\,\mathrm{d}\rho\right| \geq c_2 n\,, \end{aligned}$$

for positive constants $\widetilde{c}$ and c_2 independent of n. This yields

$$\|\mathcal{F}f_x\|_{H_1^{-1}(\mathbb{R}^{d-1})} \geq \sup_{n\in\mathbb{N}} \frac{c_2 n}{\sqrt{c_1 n}} = +\infty\,.$$

This shows that $\mathcal{F}f_x \notin H_1^{-1}(\mathbb{R}^{d-1})$. Consequently, $\mathcal{F}f_x \notin H_{-r}^{-1/2}(\mathbb{R}^{d-1})$ for $r\leq -1$. □

So far, we have shown that the operator $\mathcal{D}g$ is well defined for $g\in H_r^{1/2}(\mathbb{R}^{d-1})$ for all $r>-1$. Now we explain why we restrict the weight to $r<1$. This is because the interior boundary value problem in Ω_H is not solvable in general for $r\geq 1$. To show this, we focus on a simple case by selecting $\Gamma = \mathbb{R}^{d-1}\times\{c\}$ and $\Gamma_H = \mathbb{R}^{d-1}\times\{2c\}$. We consider the incident field $u^i(x,y) = \Phi(x,y) - \Phi(x,y')$ generated by two point sources $y = (0,y_2)^\top$ between Γ and Γ_H, and $y' = (0,y_2-2c)^\top$ below Γ. The corresponding scattered field satisfies

$$\begin{aligned} \Delta u^s + k^2 u^s &= 0 \quad \text{in } \Omega_H\,,\\ u^s &= -u^i \quad \text{on } \Gamma\,, \end{aligned}$$

together with the radiation condition $u^s(x) = (\mathcal{D}\,u^s|_{\Gamma_H})(x)$ for all x above Γ_H. The exact solution of this problem is $-G(x,\widehat{y})$, where $G(x,\widehat{y}) = \Phi(x,\widehat{y}) - \Phi(x,y')$ with $\widehat{y} = (0,2c-y_2)^\top$. Now we show that $g := G(\cdot,\widehat{y})|_{\Gamma_H} \notin H_r^{1/2}(\Gamma_H)$ when $r\geq 1$. Since $H_r^{1/2}(\Gamma_H)\subseteq H_1^{1/2}(\Gamma_H)\subseteq L_1^2(\Gamma_H)$, it is sufficient to prove $g\notin L_1^2(\Gamma_H)$.

According to [23, Eq. (2.9)], the asymptotic behaviour of Green's function is given by

$$|G(x,\widehat{y})| \sim c(k,d)\left(x_d|x|^{-(1+d)/2}\right) \text{ as } |x|\to\infty\,.$$

By using the definition of the L^2_r-norm, we obtain

$$\begin{aligned}\|g\|^2_{L^2_1(\Gamma_H)} &\geq C\int_{\mathbb{R}^{d-1}}\left(1+|x|^2\right)|x|^{-(1+d)}\,\mathrm{d}x\\ &\geq C\int_{\mathbb{R}^{d-1}\setminus B(0,1)}\left(1+|x|^2\right)|x|^{-(1+d)}\,\mathrm{d}x\\ &\geq \widehat{C}\int_{\mathbb{R}^{d-1}\setminus B(0,1)}|x|^{1-d}\,\mathrm{d}x\\ &\geq \widehat{C}\int_1^\infty \rho^{-1}\,\mathrm{d}\rho = +\infty\,,\end{aligned}$$

for some constants C and $\widehat{C}$. This shows that $g\notin L^2_1(\Gamma_H)$ and consequently $g\notin H^1_r(\Gamma_H)$ for $r\geq 1$.

In this section, we have described how to obtain the upward propagating radiation condition. However, this condition is imposed on the unbounded domain Ω^+_H above Γ_H. In the following section, we use this condition to derive a transparent boundary condition on the flat surface Γ_H for the scattering problem (1.1). This boundary condition allows us to truncate the computational domain in the vertical direction without reflecting the scattered field back into the domain artificially.

2.2.1. Vertical Domain Truncation via DtN Map

We are going to show that the outgoing solution given by (2.5) can be expressed as the trace of the solution on Γ_H. That means, equation (2.5) can be equivalently formulated by a *transparent boundary condition* on Γ_H. Taking the normal derivative of u^s with respect to x_d and evaluating it on Γ_H leads to

$$(\partial_{x_d}u^s)(\widetilde{x},H) \coloneqq (\mathcal{T}^+u^s)(\widetilde{x},H)\,, \tag{2.8}$$

where the *Dirichlet-to-Neumann (DtN) map* $\mathcal{T}^+\colon H^{1/2}_r(\Gamma_H)\to H^{-1/2}_r(\Gamma_H)$ is given by

$$(\mathcal{T}^+\varphi)(\widetilde{x},H) = \mathrm{i}(2\pi)^{-(d-1)/2}\int_{\mathbb{R}^{d-1}}\sqrt{k^2-|\xi|^2}\,\mathrm{e}^{\mathrm{i}\widetilde{x}\cdot\xi}(\mathcal{F}\varphi)(\xi,H)\,\mathrm{d}\xi\,. \tag{2.9}$$

Lemma 2.11. *For $|r|<1$, the DtN operator $\mathcal{T}^+\colon H^{1/2}_r(\Gamma_H)\to H^{-1/2}_r(\Gamma_H)$ is well defined and continuous.*

Proof. See [23, Lem. 3.3]. □

Remark 2.12. The DtN operator $\mathcal{T}^+$ can be written as

$$\mathcal{T}^+ = \mathcal{F}^{-1}\mathcal{M}_\gamma\mathcal{F}\,,$$

where $\mathcal{F}$ is the Fourier operator and the operator $\mathcal{M}_\gamma$ is the multiplication by $\gamma(\xi)$ defined by

$$\gamma(\xi) \coloneqq \begin{cases} \mathrm{i}\sqrt{k^2 - |\xi|^2} & \text{if } |\xi| \le k, \\ -\sqrt{|\xi|^2 - k^2} & \text{if } |\xi| > k\,. \end{cases} \tag{2.10}$$

The transparent boundary condition (2.8) can be used in place of the radiation condition. This results in the following boundary value problem, now posed in the vertically bounded domain Ω_H

$$\Delta u^s + k^2 u^s = 0 \qquad \text{in } \Omega_H, \tag{2.11a}$$
$$u^s = -u^i \qquad \text{on } \Gamma, \tag{2.11b}$$
$$\partial_{x_d} u^s = \mathcal{T}^+ u^s \qquad \text{on } \Gamma_H\,. \tag{2.11c}$$

Since the incident field u^i satisfies the Helmholtz equation, by using the total field $u = u^i + u^s$, we can recast problem (2.11) into

$$\Delta u + k^2 u = 0 \qquad \text{in } \Omega_H, \tag{2.12a}$$
$$u = 0 \qquad \text{on } \Gamma, \tag{2.12b}$$
$$(\partial_{x_d} - \mathcal{T}^+)u = (\partial_{x_d} - \mathcal{T}^+)u^i \qquad \text{on } \Gamma_H\,. \tag{2.12c}$$

In the rest of this work, the main focus is on the variational form of problem (2.12), which is stated below. Before stating the problem, let $\widetilde{H}^1_r(\Omega_H) \coloneqq \{\phi \in H^1_r(\Omega_H) : \phi|_\Gamma = 0\}$ for $|r| < 1$.

Variational Problem: For $u^i \in H^1_r(\Omega_H)$ with $|r| < 1$, find $u \in \widetilde{H}^1_r(\Omega_H)$ such that

$$a_r(u, v) = \left\langle (\partial_{x_d} - \mathcal{T}^+)u^i, \overline{v} \right\rangle_{\Gamma_H} \quad \text{for all } v \in \widetilde{H}^1_{-r}(\Omega_H)\,, \tag{2.13}$$

where $a_r : \widetilde{H}^1_r(\Omega_H) \times \widetilde{H}^1_{-r}(\Omega_H) \to \mathbb{C}$ is defined by

$$a_r(u, v) \coloneqq \left\langle \nabla u, \overline{\nabla v} \right\rangle_{\Omega_H} - k^2 \langle u, \overline{v} \rangle_{\Omega_H} - \left\langle \mathcal{T}^+ u, \overline{v} \right\rangle_{\Gamma_H}. \tag{2.14}$$

The above sesquilinear form is well defined and continuous on $\widetilde{H}^1_r(\Omega_H) \times \widetilde{H}^1_{-r}(\Omega_H)$ for $|r| < 1$. This is a direct consequence of Lemma 2.11.

Problem (2.13) is considered the general framework of the more specific cases studied in the subsequent chapters. We present here the existence and uniqueness results established in [23, Sec. 4] for the general case. We now elaborate on the details of the proof.

2.2.2. Existence and Uniqueness of Solutions to the Truncated Problem

The *Variational Problem* stated in (2.13) in the non-weighted Sobolev space $\widetilde{H}^1_0(\Omega_H)$ has a unique solution. This result was established in [25, Cor. 4.3] using the generalized Lax–Milgram theorem (see, e.g., [60, Thm. 2.15]).

Lemma 2.13. *Let the sesquilinear form a_0 be defined as in (2.14) for $r = 0$ which satisfies the inf-sup condition, i.e.,*

$$C_{\text{infsup}} \coloneqq \inf_{0\neq u\in\widetilde{H}^1_0(\Omega_H)} \sup_{0\neq v\in\widetilde{H}^1_0(\Omega_H)} \frac{|a_0(u,v)|}{\|u\|_{H^1_0(\Omega_H)}\|v\|_{H^1_0(\Omega_H)}} > 0\,. \tag{2.15}$$

Then, for $u^i \in H^1_0(\Omega_H)$, the Variational Problem (2.13) *has a unique solution $u \in \widetilde{H}^1_0(\Omega_H)$, which satisfies*

$$\|u\|_{H^1_0(\Omega_H)} \leq \frac{1}{C_{\text{infsup}}}\|\mathcal{G}\|_{(H^{1/2}_0(\Gamma_H))^*}\,, \tag{2.16}$$

with $\mathcal{G} \coloneqq \langle(\partial_{x_d} - \mathcal{T}^+)u^i, \cdot\rangle_{\Gamma_H} \in (H^{1/2}_0(\Gamma_H))^*$.

Proof. See [25, Cor. 4.3]. □

To extend this result to the weighted Sobolev spaces $\widetilde{H}^1_r(\Omega_H)$ for $r \neq 0$, the main idea is to use a perturbation argument involving a commutator (see [23, Sec. 2]). This reduces the theorem to a form involving only the non-weighted spaces, i.e., $r = 0$. Thus, the existence and uniqueness result presented in Lemma 2.13 can be applied.

In the following lemma, the commutator estimate is given, which is an essential tool to make a connection to the non-weighted case.

Lemma 2.14. *Let $C \coloneqq \mathcal{T}^+ - (b^2 + |x|^2)^{r/2}\mathcal{T}^+(b^2 + |x|^2)^{-r/2}$ with parameter $b > 0$. Then, for $kb \geq 1$ and $|r| < 1$,*

$$\|C\|_{H^{-1/2}(\Gamma_H)\leftarrow H^{1/2}(\Gamma_H)} \leq c(r)\sqrt{\frac{k}{b}}\,.$$

Proof. See [23, Thm. 6.1]. □

The sesquilinear form (2.14) defines a continuous linear operator $\mathcal{A}_r\colon \widetilde{H}^1_r(\Omega_H) \to (\widetilde{H}^1_r(\Omega_H))^*$ for $|r| < 1$. The invertibility of $\mathcal{A}_0$ was established in Lemma 2.13. The following theorem states that the operator $\mathcal{A}_r$ is also invertible, as shown in [23, Thm. 4.1].

Theorem 2.15. *For $|r| < 1$, the operator $\mathcal{A}_r$ is invertible. Hence, the* Variational Problem (2.13) *has a unique solution for all $u^i \in H^1_r(\Omega_H)$.*

Proof. For the parameter $b > 0$, we define the following norms

$$\|u\|_{L^2_{r,b}(\Omega_H)} \coloneqq \left\|(b^2 + |x|^2)^{r/2}u\right\|_{L^2(\Omega_H)}\,,$$
$$\|u\|^2_{H^1_{r,b}(\Omega_H)} \coloneqq \int_{\Omega_H}\left(\left|(b^2 + |x|^2)^{r/2}u\right|^2 + \left|\nabla\left((b^2 + |x|^2)^{r/2}u\right)\right|^2\right)\mathrm{d}x\,.$$

Let $a > 0$ be sufficiently large. For $u \in \widetilde{H}^1_{r,b}(\Omega_H)$ and $v \in \widetilde{H}^1_{-r,b}(\Omega_H)$, we consider

$$\varphi \coloneqq (b^2 + |x|^2)^{r/2}u \in \widetilde{H}^1(\Omega_H)\,,$$
$$\psi \coloneqq (b^2 + |x|^2)^{-r/2}v \in \widetilde{H}^1(\Omega_H)\,.$$

Substituting $u = (b^2 + |x|^2)^{-r/2}\varphi$ and $v = (b^2 + |x|^2)^{r/2}\psi$ into the sesquilinear form (2.14) yields

$$a_r(u, v) = a_0(\varphi, \psi) + l_b(\varphi, \psi)\,, \tag{2.17}$$

where $a_0\colon \widetilde{H}^1(\Omega_H) \times \widetilde{H}^1(\Omega_H) \to \mathbb{C}$ has the same representation as a_r given in (2.14). Moreover, $l_b \coloneqq l_{b,1} + l_{b,2}$ with

$$\begin{aligned} l_{b,1}(\varphi, \psi) \coloneqq &\left\langle \nabla\left((b^2 + |x|^2)^{-r/2}\right)\varphi, \overline{\nabla\left((b^2 + |x|^2)^{r/2}\right)\psi}\right\rangle_{\Omega_H} \\ &+ \left\langle (b^2 + |x|^2)^{r/2}\left(\nabla(b^2 + |x|^2)^{-r/2}\right)\varphi, \overline{\nabla\psi}\right\rangle_{\Omega_H} \\ &+ \left\langle \nabla\varphi, \overline{(b^2 + |x|^2)^{-r/2}\left(\nabla(b^2 + |x|^2)^{r/2}\right)\psi}\right\rangle_{\Omega_H} \end{aligned}$$

and

$$l_{b,2}(\varphi, \psi) \coloneqq \left\langle (b^2 + |x|^2)^{r/2}\,\mathcal{T}^+(b^2 + |x|^2)^{-r/2}\varphi - \mathcal{T}^+\varphi, \overline{\psi}\right\rangle_{\Gamma_H} = -\left\langle C\varphi, \overline{\psi}\right\rangle_{\Gamma_H}.$$

For the term $l_{b,1}$, we can obtain the following estimate

$$\begin{aligned} |l_{b,1}(\varphi, \psi)| \leq &\left|\left\langle \left((b^2 + |x|^2)^{r/2}\nabla(b^2 + |x|^2)^{-r/2}\right)\varphi, \overline{\left((b^2 + |x|^2)^{-r/2}\nabla(b^2 + |x|^2)^{r/2}\right)\psi}\right\rangle_{\Omega_H}\right| \\ &+ \left|\left\langle \left((b^2 + |x|^2)^{r/2}\nabla(b^2 + |x|^2)^{-r/2}\right)\varphi, \overline{\nabla\psi}\right\rangle_{\Omega_H}\right| \\ &+ \left|\left\langle \nabla\varphi, \overline{\left((b^2 + |x|^2)^{-r/2}\nabla(b^2 + |x|^2)^{r/2}\right)\psi}\right\rangle_{\Omega_H}\right|. \end{aligned}$$

Since

$$\begin{aligned} \sup_{x\in\Omega_H}\left|\nabla(b^2 + |x|^2)^{r/2}\right|(b^2 + |x|^2)^{-r/2} &= |r| \sup_{x\in\Omega_H}(b^2 + |x|^2)^{r/2-1}|x|(b^2 + |x|^2)^{-r/2} \\ &= |r| \sup_{x\in\Omega_H}\left(\frac{b^2}{|x|} + |x|\right)^{-1} \\ &= \frac{|r|}{b}\left(\inf_{x\in\Omega_H}\left(\frac{b}{|x|} + \frac{|x|}{b}\right)\right)^{-1} \leq \frac{|r|}{2b}\,, \end{aligned}$$

the previous estimate can be written as

$$\begin{aligned} |l_{b,1}(\varphi, \psi)| &\leq \left(\frac{|r|}{2b}\right)^2 \|\varphi\|_{L^2(\Omega_H)}\|\psi\|_{L^2(\Omega_H)} \\ &\quad + \frac{|r|}{2b}\left(\|\nabla\varphi\|_{L^2(\Omega_H)}\|\psi\|_{L^2(\Omega_H)} + \|\varphi\|_{L^2(\Omega_H)}\|\nabla\psi\|_{L^2(\Omega_H)}\right) \\ &\leq \frac{|r|}{2b}\max\left\{1, \frac{|r|}{2b}\right\}\|\varphi\|_{H^1(\Omega_H)}\|\psi\|_{H^1(\Omega_H)}\,. \end{aligned} \tag{2.18}$$

Moreover, using Lemma 2.14 yields

$$\begin{aligned} |l_{b,2}(\varphi,\psi)| = \left|-\left\langle C\varphi, \overline{\psi}\right\rangle_{\Gamma_H}\right| &\leq \|C\varphi\|_{H^{-1/2}(\Gamma_H)}\|\psi\|_{H^{1/2}(\Gamma_H)} \\ &\leq c(r)\sqrt{\frac{k}{b}}\|\varphi\|_{H^{1/2}(\Gamma_H)}\|\psi\|_{H^{1/2}(\Gamma_H)} \\ &\leq c(r,k)\sqrt{\frac{1}{b}}\|\varphi\|_{H^1(\Omega_H)}\|\psi\|_{H^1(\Omega_H)}\,, \end{aligned} \tag{2.19}$$

where the last inequality is obtained by using (2.1).

Considering the operator $\mathcal{L}_b\colon \widetilde{H}^1(\Omega_H) \to (\widetilde{H}^1(\Omega_H))^*$ generated by l_b and using (2.18) and (2.19), we conclude that $\mathcal{L}_b$ tends to zero when b tends to infinity.

Finally, the operator $\mathcal{A}_r$ generated by (2.17) can be written as

$$\mathcal{A}_r = (b^2+|x|^2)^{-r/2}(\mathcal{A}_0+\mathcal{L}_b)(b^2+|x|^2)^{r/2}\,,$$

where the operator $\mathcal{A}_0$ corresponds to the sesquilinear form a_0. According to Lemma 2.13, the operator $\mathcal{A}_0$ is invertible. Furthermore, the operator $\mathcal{A}_0+\mathcal{L}_b$ is a small perturbation of the operator $\mathcal{A}_0$ when b is sufficiently large. Therefore, by applying the perturbation theorem (see [75, Thm. 10.1]), we conclude that the operator $\mathcal{A}_r$ is invertible. □

2.3. Perfectly Matched Layer

Another approach to truncate the domain vertically away from the scatterer is the perfectly matched layer (PML). The main idea is to add an absorbing layer with finite thickness above the computational domain. Absorption is obtained by stretching the vertical coordinate into the complex plane. Since the outgoing waves are absorbed by the PML, the problem can be truncated by imposing a boundary condition at the top of the layer. In this work, we choose the homogeneous Dirichlet boundary condition. This section elaborates on how the PML can be used as a truncation method, based on [26, Sec. 2].

2.3.1. Vertical Domain Truncation via PML

Recall the surface Γ, which is the graph of the function ζ introduced in Chapter 1. To describe the PML, we first define some notations.

We introduce two flat surfaces $\Gamma_H \coloneqq \mathbb{R}^{d-1}\times\{H\}$ and $\Gamma_{H+\lambda} \coloneqq \mathbb{R}^{d-1}\times\{H+\lambda\}$ for some $H > \|\zeta\|_\infty$ and $\lambda > 0$. The PML, denoted by $\Omega_{\mathrm{PML}} \coloneqq \mathbb{R}^{d-1}\times(H, H+\lambda)$, is the region between these two surfaces with the physical width λ. Moreover, we define $\Omega_{H+\lambda} \coloneqq \Omega_H \cup \Omega_{\mathrm{PML}}$. A sketch of these domains is presented in Figure 2.2.

To derive the PML problem, we select an integrable function $s\colon (-\infty, H+\lambda] \to \mathbb{C}$ such that $s(t) = 1$ for $t \leq H$ and for $t > H$, $\mathrm{Re}(s(t)) > 0$ and $\mathrm{Im}(s(t)) > 0$. The *complex stretched coordinate* Ξ is defined by

$$\Xi(x_d) \coloneqq \int_0^{x_d} s(t)\,\mathrm{d}t\,. \tag{2.20}$$

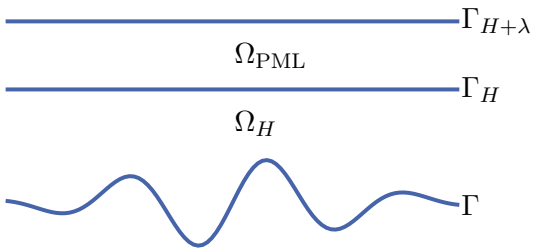


FIGURE 2.2. A two-dimensional sketch of the PML.

Clearly Ξ is the identity below H. Physically, this means that waves below the PML propagate freely, as they would in an unbounded medium, with no modifications to absorb or attenuate them. On the other hand, above H, the coordinate stretching introduces a complex transformation. This makes the PML act as an absorbing layer, in which the outgoing solutions are gradually damped without being reflected back into the computational domain Ω_H. As described in [26, Sec. 2], a common function to use in the complex stretched coordinate is a power law, namely

$$s(t) = \begin{cases} 1 & \text{if } t < H\,, \\ 1 + \rho\, \mathrm{e}^{\mathrm{i}\pi/4} \left(\dfrac{t-H}{\lambda} \right)^2 & \text{if } t \geq H\,, \end{cases} \tag{2.21}$$

where ρ is a positive parameter.

Since the radiating solution (2.5) is an analytic function with respect to x_d, we can analytically continue it to a function defined for complex coordinates. We still denote this extension by u^s.

The analytic continuation of the solution satisfies the Helmholtz equation in the complex coordinates, i.e.,

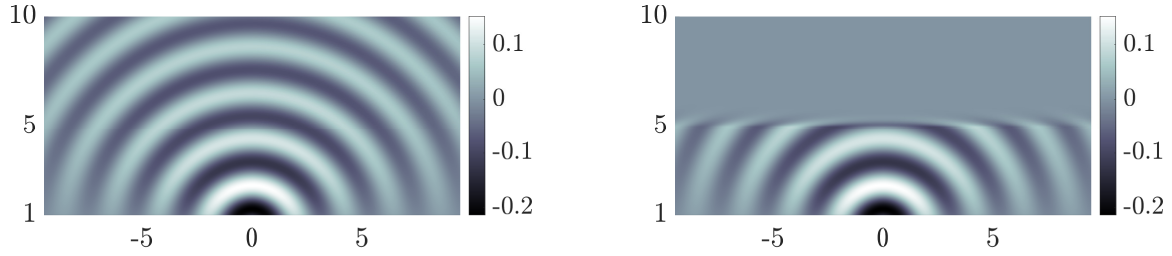
$$\Delta u^s + k^2 u^s = 0 \qquad \text{in } \mathbb{R}^{d-1} \times \Xi([H, H+\lambda])\,. \tag{2.22}$$

Considering the complex coordinate $\Xi(x_d)$ modifies the behavior of the general solution inside the PML such that it decays as $|x_d|$ increases. The absorption strength of the PML is determined by the virtual width of the layer, given by

$$\sigma := \int_H^{H+\lambda} s(t)\,\mathrm{d}t = \lambda \left(1 + \frac{\rho\, \mathrm{e}^{\mathrm{i}\pi/4}}{3} \right). \tag{2.23}$$

Remark 2.16. To illustrate the influence of σ on the absorbing strength of the PML, we fix the physical thickness λ and assume that the function u^s is sufficiently regular. Extending the vertical direction to the complex coordinate in (2.5), we have for $\widetilde{x} \in \mathbb{R}^{d-1}$

$$\begin{aligned} u^s(\widetilde{x}, H+\lambda) &= \mathcal{F}^{-1}\left(\mathrm{e}^{\mathrm{i}\sqrt{k^2-|\xi|^2}(\Xi(H+\lambda)-H)} \mathcal{F}u^s(\xi, H) \right) \\ &= \mathcal{F}^{-1}\left(\mathrm{e}^{\mathrm{i}\sigma\sqrt{k^2-|\xi|^2}} \mathcal{F}u^s(\xi, H) \right) \\ &= \mathcal{F}^{-1}\left(\mathrm{e}^{\mathrm{i}k\sigma\sqrt{1-|\xi/k|^2}} \mathcal{F}u^s(\xi, H) \right). \end{aligned}$$



(a) Oscillatory behaviour without PML in $[-\pi, \pi] \times [1, 10]$

(b) Absorbing effect of the PML in $[-\pi, \pi] \times \Xi([1, 10])$

Figure 2.3. Behaviour of Green's function for $y = (0, 0.5)$.

We see that a large $\mathrm{Im}(k\sigma)$ enhances absorption of the propagating waves (i.e., $|\xi| < k$) entering the PML. On the other hand, by a large $\mathrm{Re}(k\sigma)$, the PML effectively absorbs evanescent modes (i.e., $|\xi| > k$).

In the following example, we illustrate the absorbing effect of the PML.

Example 2.17. We consider the Dirichlet Green's function in the upper half plane as the outgoing wave

$$G(x, y) = \frac{\mathrm{i}}{4}\left(H_0^{(1)}(k|x - y|) - H_0^{(1)}(k|x - \widehat{y}|)\right), \quad x \in \mathbb{R}^2_+ \coloneqq \{x \in \mathbb{R}^2 : x_2 > 0\},$$

with the point source $y = (0, y_2)$ and the reflected source $\widehat{y} = (0, -y_2)$, for $y_2 > 0$. Defining $R \coloneqq |x - y|$ and $\widehat{R} \coloneqq |x - \widehat{y}|$ and using the fact that $(H_0^{(1)})' = -H_1^{(1)}$, the Dirichlet Green's function can be written as

$$G(x, y) = \frac{\mathrm{i}}{4}\int_{kR}^{k\widehat{R}} H_1^{(1)}(\tau)\,\mathrm{d}\tau\,, \quad x \in \mathbb{R}^2_+. \tag{2.24}$$

According to [93, Eq. (10.2.5)], we have

$$H_1^{(1)}(\tau) \sim \sqrt{\frac{2}{\pi\tau}}\mathrm{e}^{\mathrm{i}(\tau - \frac{3\pi}{4})} \quad \text{as } \tau \to \infty\,.$$

By substituting the equation above in (2.24), we obtain

$$\begin{aligned} G(x, y) &\sim \frac{\mathrm{i}}{4}\sqrt{\frac{2}{\pi}}\mathrm{e}^{-\mathrm{i}\frac{3\pi}{4}}\left(\int_{kR}^{k\widehat{R}} \frac{\mathrm{e}^{\mathrm{i}kR}}{\sqrt{kR}}\,\mathrm{d}\tau + \int_{kR}^{k\widehat{R}}\left(\frac{\mathrm{e}^{\mathrm{i}\tau}}{\sqrt{\tau}} - \frac{\mathrm{e}^{\mathrm{i}kR}}{\sqrt{kR}}\right)\mathrm{d}\tau\right) \\ &= \frac{\mathrm{i}}{4}k(\widehat{R} - R)\sqrt{\frac{2}{\pi kR}}\mathrm{e}^{\mathrm{i}(kR - \frac{3\pi}{4})} + \frac{\mathrm{i}}{4}\sqrt{\frac{2}{\pi}}\mathrm{e}^{-\mathrm{i}\frac{3\pi}{4}}\int_{kR}^{k\widehat{R}}\left(\frac{\mathrm{e}^{\mathrm{i}\tau}}{\sqrt{\tau}} - \frac{\mathrm{e}^{\mathrm{i}kR}}{\sqrt{kR}}\right)\mathrm{d}\tau\,. \end{aligned}$$

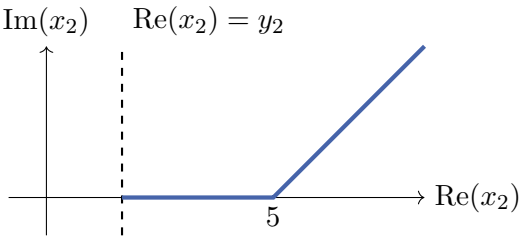


FIGURE 2.4. Complex stretched coordinate in the vertical direction x_2.

Since $t \in [kR, k\widehat{R}]$ and

$$\begin{aligned}\left|\frac{\mathrm{i}}{4}\sqrt{\frac{2}{\pi}}\mathrm{e}^{-\mathrm{i}\frac{3\pi}{4}}\int_{kR}^{k\widehat{R}}\left(\frac{\mathrm{e}^{\mathrm{i}\tau}}{\sqrt{\tau}}-\frac{\mathrm{e}^{\mathrm{i}kR}}{\sqrt{kR}}\right)\mathrm{d}\tau\right| &= \mathcal{O}\left(\int_{kR}^{k\widehat{R}}\left|\frac{\mathrm{e}^{\mathrm{i}\tau}}{\sqrt{\tau}}-\frac{\mathrm{e}^{\mathrm{i}kR}}{\sqrt{kR}}\right|\mathrm{d}\tau\right)\\ &= \mathcal{O}\left(\int_{kR}^{k\widehat{R}}\left(\left|\frac{1}{\sqrt{\tau}}\right|+\left|\frac{1}{\sqrt{kR}}\right|\right)\mathrm{d}\tau\right)\\ &= \mathcal{O}\left(\int_{kR}^{k\widehat{R}}\frac{1}{\sqrt{R}}\,\mathrm{d}\tau\right)\\ &= \mathcal{O}\left(\frac{\widehat{R}-R}{\sqrt{R}}\right),\end{aligned}$$

it is concluded that

$$G(x,y) \sim \frac{\mathrm{i}}{4}k(\widehat{R}-R)\sqrt{\frac{2}{\pi kR}}\mathrm{e}^{\mathrm{i}(kR-\frac{3\pi}{4})} \quad \text{as } x \to \infty\,.$$

The oscillatory behaviour of this function is plotted in Figure 2.3(a).

Green's function is analytic with respect to $x_2 > y_2$. Hence, it can be analytically continued into the complex half plane $\{x_2 \in \mathbb{C} : \operatorname{Re} x_2 > y_2\}$, while x_1 is a real number. The analytic continuation of R is thus $R = \sqrt{x_1^2 + (x_2 - y_2)^2}$ with the branch cut on the real negative axis. As $\operatorname{Re}(x_2 - y_2) > 0$, then $\operatorname{Re} R > 0$ and $G \sim \mathcal{O}(\sqrt{|R|})\mathrm{e}^{\mathrm{i}kR}$ as $|R| \to \infty$. This shows that Green's function is exponentially decaying as $\operatorname{Im}(x_2) \to \infty$.

We evaluate Green's function along the complex coordinate in the vertical direction x_2, that means, for $\operatorname{Re} x_2 > 5$ we have added a linearly growing imaginary part (depicted in Figure 2.4). In Figure 2.3(b), we see that the Green's function is decaying for $\operatorname{Re} x_2 > 5$, which is consistent with the previous computation.

Since the solution decays in the PML, we can expect to obtain a good approximation by truncating the domain and imposing a homogeneous Dirichlet boundary condition on the artificial boundary $\Gamma_{H+\lambda}$.

Solving the differential equation (2.22) along the complex stretched coordinate directly is challenging. Instead, we transform the complex coordinate to the real standard coordinates by using the change of variables Ξ. This yields

$$u_\sigma^s(\widetilde{x}, x_d) \coloneqq u^s(\widetilde{x}, \Xi(x_d))\,. \tag{2.25}$$

Due to the definition of Ξ in (2.20), we have

$$\partial_{x_d} u_\sigma^s(x) = s(x_d)\partial_{x_d} u^s(\widetilde{x}, \Xi(x_d))\,.$$

By substituting the above change of variables into (2.22), we obtain

$$\sum_{j=1}^{d-1} \partial_{x_j}\left(s(x_d)\partial_{x_j} u_\sigma^s\right) + \partial_{x_d}\left(\frac{1}{s(x_d)}\partial_{x_d} u_\sigma^s\right) + k^2 s(x_d) u_\sigma^s = 0 \quad \text{in } \Omega_{H+\lambda}\,.$$

For simplification, we define the PML operator Δ_{PML} as follows

$$\Delta_{\mathrm{PML}} \coloneqq \sum_{j=1}^{d-1} \partial_{x_j}\left(s(x_d)\partial_{x_j}\right) + \partial_{x_d}\left(\frac{1}{s(x_d)}\partial_{x_d}\right) = \nabla\cdot(\mathbf{S}(x_d)\nabla) \tag{2.26}$$

with the matrix $\mathbf{S}(x_d) \coloneqq \mathrm{diag}(s(x_d), \ldots, s(x_d), 1/s(x_d)) \in \mathbb{C}^{d\times d}$.

Remark 2.18. Below the PML, the operator Δ_{PML} is equal to the Laplace operator, since $s(x_d) = 1$ for $x_d \le H$ and $\mathbf{S}$ is the identity.

The truncated PML problem can now be formulated as follows: For $u^i \in H^{1/2}(\Gamma)$, we seek the weak solution $u_\sigma^s \in H^1(\Omega_{H+\lambda})$ such that

$$\Delta_{\mathrm{PML}} u_\sigma^s + k^2 s(x_d) u_\sigma^s = 0 \qquad \text{in } \Omega_{H+\lambda}\,, \tag{2.27a}$$
$$u_\sigma^s = -u^i \qquad \text{on } \Gamma\,, \tag{2.27b}$$
$$u_\sigma^s = 0 \qquad \text{on } \Gamma_{H+\lambda}\,. \tag{2.27c}$$

The solution of problem (2.27) in the PML is not related to the actual scattered field. In the following section, we explain how to reformulate the PML problem with an artificial boundary condition on Γ_H.

2.3.2. PML Approximation of the DtN Map

We are now going to obtain an approximation of the DtN map by using the PML. To this end, we need to solve the Helmholtz equation in Ω_{PML} for $g \in H^{1/2}(\Gamma_H)$ such that

$$\Delta_{\mathrm{PML}} u_\sigma^s + k^2 s(x_d) u_\sigma^s = 0 \qquad \text{in } \Omega_{\mathrm{PML}}\,, \tag{2.28a}$$
$$u_\sigma^s = g \qquad \text{on } \Gamma_H\,, \tag{2.28b}$$
$$u_\sigma^s = 0 \qquad \text{on } \Gamma_{H+\lambda}\,. \tag{2.28c}$$

By considering the definition of u_σ^s in (2.25), we can express its Fourier transform using (2.4). Therefore, there exist two constants C_1, C_2 depending on ξ but not on x_d such that

$$(\mathcal{F}u_\sigma^s)(\xi, x_d) = C_1(\xi)\mathrm{e}^{\gamma(\xi)(\Xi(x_d)-H)} + C_2(\xi)\mathrm{e}^{-\gamma(\xi)(\Xi(x_d)-H)}\,, \quad \xi \in \mathbb{R}^{d-1},\ x_d \in [H, H+\lambda]\,,$$

where $\gamma(\xi)$ is given by (2.10). By imposing the boundary conditions (2.28b) and (2.28c) and solving the resulting linear system, the constant C_1 and C_2 are obtained as

$$C_1(\xi) = \frac{\mathrm{e}^{-\gamma(\xi)\sigma}}{\mathrm{e}^{-\gamma(\xi)\sigma} - \mathrm{e}^{\gamma(\xi)\sigma}}(\mathcal{F}g)(\xi, H) \quad \text{and} \quad C_2(\xi) = -\frac{\mathrm{e}^{\gamma(\xi)\sigma}}{\mathrm{e}^{-\gamma(\xi)\sigma} - \mathrm{e}^{\gamma(\xi)\sigma}}(\mathcal{F}g)(\xi, H)\,,$$

where σ is the virtual width of the PML as defined in (2.23). Hence, we have

$$(\mathcal{F}u_\sigma^s)(\xi, x_d) = \frac{\mathrm{e}^{\gamma(\xi)(\Xi(x_d)-H-\sigma)} - \mathrm{e}^{-\gamma(\xi)(\Xi(x_d)-H-\sigma)}}{\mathrm{e}^{-\gamma(\xi)\sigma} - \mathrm{e}^{\gamma(\xi)\sigma}}(\mathcal{F}g)(\xi, H)\,, \quad \xi \in \mathbb{R}^{d-1}\,.$$

Taking the derivative of $\mathcal{F}u_\sigma^s$ with respect to x_d, evaluating it on Γ_H and using $\Xi(H) = H$ and $\Xi'(H) = 1$, we obtain

$$\partial_{x_d}(\mathcal{F}u_\sigma^s)(\xi, H) = \gamma(\xi)\left(\frac{\mathrm{e}^{-\gamma(\xi)\sigma} + \mathrm{e}^{\gamma(\xi)\sigma}}{\mathrm{e}^{-\gamma(\xi)\sigma} - \mathrm{e}^{\gamma(\xi)\sigma}}\right)(\mathcal{F}g)(\xi, H) = \gamma(\xi)\coth\left(\gamma(\xi)\sigma\right)(\mathcal{F}g)(\xi, H)\,.$$

Using the inverse Fourier transform, we can define $\mathcal{T}_\sigma^+ \colon H^{1/2}(\Gamma_H) \to H^{-1/2}(\Gamma_H)$ as the Neumann data of the solution on Γ_H:

$$(\mathcal{T}_\sigma^+ u_\sigma^s)(\widetilde{x}, H) \coloneqq \partial_{x_d} u_\sigma^s(\widetilde{x}, H) = (2\pi)^{-\frac{d-1}{2}} \int_{\mathbb{R}^{d-1}} \gamma(\xi)\coth\left(\gamma(\xi)\sigma\right)(\mathcal{F}g)(\xi, H)\mathrm{e}^{\mathrm{i}\widetilde{x}\cdot\xi}\,\mathrm{d}\xi\,. \tag{2.29}$$

Since the solution u_σ^s satisfies the above boundary condition on Γ_H, we can use it to obtain an equivalent version of (2.27), namely

$$\begin{aligned} \Delta u_\sigma^s + k^2 u_\sigma^s &= 0 && \text{in } \Omega_H\,, \\ u_\sigma^s &= -u^i && \text{on } \Gamma\,, \\ \partial_{x_d} u_\sigma^s &= \mathcal{T}_\sigma^+ u_\sigma^s && \text{on } \Gamma_H\,. \end{aligned}$$

As the incident field u^i also satisfies the Helmholtz equation, the above problem can be further reformulated by using the total field $u_\sigma = u^i + u_\sigma^s$.

PML Problem: For $u^i \in H^1(\Omega_H)$, we seek the weak solution $u_\sigma \in \widetilde{H}^1(\Omega_H)$ such that

$$\Delta u_\sigma + k^2 u_\sigma = 0 \qquad \text{in } \Omega_H, \tag{2.30a}$$

$$u_\sigma = 0 \qquad \text{on } \Gamma, \tag{2.30b}$$

$$(\partial_{x_d} - \mathcal{T}_\sigma^+)u_\sigma = (\partial_{x_d} - \mathcal{T}_\sigma^+)u^i \qquad \text{on } \Gamma_H\,. \tag{2.30c}$$

The corresponding variational form is to find $u_\sigma \in \widetilde{H}^1(\Omega_H)$ such that

$$a_{\mathrm{PML}}(u_\sigma, v) = \left\langle (\partial_{x_d} - \mathcal{T}_\sigma^+)u^i, \overline{v}\right\rangle_{\Gamma_H} \quad \text{for all } v \in \widetilde{H}^1(\Omega_H)\,,$$

where $a_{\mathrm{PML}} \colon \widetilde{H}^1(\Omega_H) \times \widetilde{H}^1(\Omega_H) \to \mathbb{C}$ is defined by

$$a_{\mathrm{PML}}(\phi, \psi) \coloneqq \left\langle \nabla\phi, \overline{\nabla\psi}\right\rangle_{\Omega_H} - k^2\left\langle \phi, \overline{\psi}\right\rangle_{\Omega_H} - \left\langle \mathcal{T}_\sigma^+\phi, \overline{\psi}\right\rangle_{\Gamma_H}. \tag{2.31}$$

In Chapter 4, we study the *PML Problem* posed in a domain with a special structure. To provide a foundation for this analysis, it is useful to prove the existence and uniqueness for the general case, which has been shown in [26, Sec. 3]. In the following section, we elaborate on the details.

2.3.3. Existence and Uniqueness of Solutions to the PML Problem

The sesquilinear form a_{PML} defined in (2.31) generates the operator $\mathcal{A}_\sigma\colon \widetilde{H}^1(\Omega_H) \to (\widetilde{H}^1(\Omega_H))^*$ such that

$$\langle \mathcal{A}_\sigma u, \overline{v}\rangle_{\Omega_H} \coloneqq a_{\mathrm{PML}}(u,v) \qquad \text{for all } v \in \widetilde{H}^1(\Omega_H)\,.$$

Before presenting the main result, we need some preliminary lemmas.

Lemma 2.19. *Let the operator $\mathcal{A}$ and $\mathcal{A}_\sigma$ be induced by (2.14) and (2.31), respectively. Then,*

$$\|\mathcal{A}-\mathcal{A}_\sigma\|_{H^{-1}(\Omega_H)\leftarrow H^1(\Omega_H)} \leq 2\left\|\mathcal{T}^+ - \mathcal{T}_\sigma^+\right\|_{H^{-1/2}(\Gamma_H)\leftarrow H^{1/2}(\Gamma_H)}\,,$$

where $\mathcal{T}^+$ and $\mathcal{T}_\sigma^+$ are given by (2.9) and (2.29), respectively.

Proof. For $u, v \in \widetilde{H}^1(\Omega_H)$, a straightforward computation yields

$$\begin{aligned}
\left|\langle(\mathcal{A}-\mathcal{A}_\sigma)u,\overline{v}\rangle_{\Omega_H}\right| &= |a(u,v)-a_{\mathrm{PML}}(u,v)| \\
&= \left|\int_{\Gamma_H} \overline{v}(\mathcal{T}^+ - \mathcal{T}_\sigma^+)u\,\mathrm{d}s\right| \\
&\leq \left\|\mathcal{T}^+ - \mathcal{T}_\sigma^+\right\|_{H^{-1/2}(\Gamma_H)\leftarrow H^{1/2}(\Gamma_H)} \|u\|_{H^{1/2}(\Gamma_H)}\|v\|_{H^{1/2}(\Gamma_H)} \\
&\leq 2\left\|\mathcal{T}^+ - \mathcal{T}_\sigma^+\right\|_{H^{-1/2}(\Gamma_H)\leftarrow H^{1/2}(\Gamma_H)} \|u\|_{H^1(\Omega_H)}\|v\|_{H^1(\Omega_H)}\,,
\end{aligned}$$

where the last inequality is obtained using $\|v\|_{H^{1/2}(\Gamma_H)} \leq \sqrt{2}\|v\|_{H^1(\Omega_H)}$ from [25, Lem. 3.4]. □

Lemma 2.20. *Let σ denote the virtual width of the PML as in (2.23). Then,*

$$\left\|\mathcal{T}^+ - \mathcal{T}_\sigma^+\right\|_{H^{-1/2}(\Gamma_H)\leftarrow H^{1/2}(\Gamma_H)} \leq C_u(k\sigma)\,,$$

where

$$C_u(z) \coloneqq \frac{1}{\mathrm{e}}\max\left\{\frac{1}{\operatorname{Re} z} + \frac{\operatorname{Im} z}{\pi(\operatorname{Re} z)^2}, \frac{1}{\operatorname{Im} z} + \frac{\operatorname{Re} z}{\pi(\operatorname{Im} z)^2}\right\} \quad \textit{for } \operatorname{Re} z, \operatorname{Im} z > 0\,.$$

Proof. See [26, Thm. 3.1]. □

Theorem 2.21. *Let C_u be as in the previous lemma and let C_{infsup} be as in (2.15). Moreover, assume that $2C_u(k\sigma) < C_{\mathrm{infsup}}$ and $u^i \in H^1(\Omega_H)$. Then, the* PML problem *(2.30) has a unique solution $u_\sigma \in \widetilde{H}^1(\Omega_H)$ and the following error estimation holds*

$$\|u-u_\sigma\|_{H^1(\Omega_H)} \leq \frac{2C_u(k\sigma)}{C_{\mathrm{infsup}} - 2C_u(k\sigma)}\left(\|u\|_{H^1(\Omega_H)} + \left\|u^i\right\|_{H^1(\Omega_H)}\right)\,,$$

where u is the solution to (2.12).

Proof. Let $\mathcal{A}$ and $\mathcal{A}_\sigma$ be defined by the sesquilinear forms (2.14) and (2.31). According to Theorem 2.15, the operator $\mathcal{A}$ is invertible and, by using [60, Thm. 2.15], it satisfies

$$\left\|\mathcal{A}^{-1}\right\|_{H^1(\Omega_H)\leftarrow H^{-1}(\Omega_H)} \leq \frac{1}{C_{\text{infsup}}}\,. \tag{2.32}$$

To show existence and uniqueness of the *PML Problem*, it suffices to show that $\mathcal{A}_\sigma$ is boundedly invertible. Using the perturbation theorem [75, Thm. 10.1], the inverse of the operator $\mathcal{A}_\sigma$ exists provided that

$$\left\|\mathcal{A}^{-1}(\mathcal{A}-\mathcal{A}_\sigma)\right\|_{H^1(\Omega_H)\leftarrow H^1(\Omega_H)} < 1\,.$$

By straightforward computations and using (2.32) and Lemmas 2.19 and 2.20, we have

$$\begin{aligned}\left\|\mathcal{A}^{-1}(\mathcal{A}-\mathcal{A}_\sigma)\right\|_{H^1(\Omega_H)\leftarrow H^1(\Omega_H)} &\leq \left\|\mathcal{A}^{-1}\right\|_{H^1(\Omega_H)\leftarrow H^{-1}(\Omega_H)}\|\mathcal{A}-\mathcal{A}_\sigma\|_{H^{-1}(\Omega_H)\leftarrow H^1(\Omega_H)}\\ &\leq \frac{2}{C_{\text{infsup}}}\left\|\mathcal{T}^+-\mathcal{T}_\sigma^+\right\|_{H^{-1/2}(\Gamma_H)\leftarrow H^{1/2}(\Gamma_H)}\\ &\leq \frac{2}{C_{\text{infsup}}}C_u(k\sigma) < 1\,.\end{aligned} \tag{2.33}$$

To compute the error bound, we again use the perturbation theorem [75, Thm. 10.1] as well as the above results given in (2.32) and (2.33). These give us

$$\begin{aligned}\|u-u_\sigma\|_{H^1(\Omega_H)} &\leq \frac{\|\mathcal{A}^{-1}\|_{H^1(\Omega_H)\leftarrow H^{-1}(\Omega_H)}}{1-\|\mathcal{A}^{-1}(\mathcal{A}-\mathcal{A}_\sigma)\|_{H^1(\Omega_H)\leftarrow H^1(\Omega_H)}}\Big(\|(\mathcal{A}-\mathcal{A}_\sigma)u\|_{H^{-1}(\Omega_H)}\\ &\qquad + \left\|(\mathcal{T}^+-\mathcal{T}_\sigma^+)u^i\right\|_{H^{-1/2}(\Gamma_H)}\Big)\\ &\leq \frac{1}{C_{\text{infsup}}-2C_u(k\sigma)}\Big(\|(\mathcal{A}-\mathcal{A}_\sigma)u\|_{H^{-1}(\Omega_H)} + \left\|(\mathcal{T}^+-\mathcal{T}_\sigma^+)u^i\right\|_{H^{-1/2}(\Gamma_H)}\Big)\\ &\leq \frac{\|(\mathcal{T}^+-\mathcal{T}_\sigma^+)\|_{H^{-1/2}(\Gamma_H)\leftarrow H^{1/2}(\Gamma_H)}}{C_{\text{infsup}}-2C_u(k\sigma)}\left(2\|u\|_{H^1(\Omega_H)} + \sqrt{2}\left\|u^i\right\|_{H^1(\Omega_H)}\right)\\ &\leq \frac{2C_u(k\sigma)}{C_{\text{infsup}}-2C_u(k\sigma)}\left(\|u\|_{H^1(\Omega_H)} + \left\|u^i\right\|_{H^1(\Omega_H)}\right)\,,\end{aligned}$$

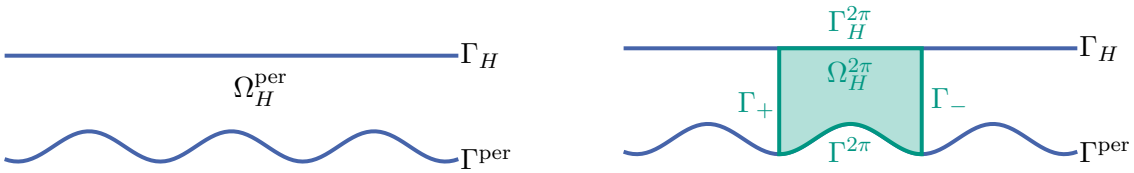
where the last inequality is obtained by Lemma 2.20. □

Remark 2.22. Without prior knowledge of the properties of the total field u, an a priori estimate for $(\mathcal{T}^+-\mathcal{T}_\sigma^+)u$ is difficult. Therefore, determining the optimal value for σ is not a straightforward task.

Remark 2.23. By assuming that $1/c \leq \operatorname{Re}(k\sigma)/\operatorname{Im}(k\sigma) \leq c$ for some constant $c > 1$, Theorem 2.21 leads to

$$\|u-u_\sigma\|_{H^1(\Omega_H)} = \mathcal{O}(1/\operatorname{Re}(k\sigma)) \quad \text{as } \operatorname{Re}(k\sigma)\to\infty\,.$$

This indicates that the global error decreases at least linearly as $\operatorname{Re}(k\sigma)\to\infty$. Moreover, it has been shown in [26, Thm. 4.2] that, for a flat scatterer, the global error decreases no faster than $|k\sigma|^{-2}\log(|k\sigma|)^{-1}$ as $|k\sigma|\to\infty$. This means that the global exponential convergence is



(a) Periodic structure (b) Bounded cell

Figure 2.5. A two-dimensional sketch of an unbounded periodic domain and its fundamental cell.

unachievable for the unbounded scatterer. This contrasts with the exponential rate of the PML for the bounded scatterers, as proven in [28, 79, 80]. However, it has been shown in [105] that the PML solution of the source problem converges exponentially on a compact subset of the unbounded purely periodic domains. In Chapter 4, we will extend these results to the scattering problems in locally perturbed domains.

2.4. Floquet–Bloch Transform

Up to this point, we have described two methods for truncating the generic domain Ω in the vertical direction. However, the resulting truncated domain Ω_H remains unbounded in the horizontal directions. Under the assumption of periodicity of the domain in the horizontal directions, a widely used tool is the Floquet–Bloch (FB) transform, which decomposes a non-periodic function defined on an unbounded periodic domain into a family of periodic functions, each defined in a bounded cell [2, 7, 77, 81].

We consider a 2π-periodic function $\zeta^{\mathrm{per}}\colon \mathbb{R}^{d-1} \to \mathbb{R}$ whose graph is the periodic surface Γ^{per}. For $H \geq \|\zeta^{\mathrm{per}}\|_\infty$, we define the periodic domain Ω_H^{per}, depicted in Figure 2.5a, by

$$\Omega_H^{\mathrm{per}} \coloneqq \left\{(\widetilde{x}, x_d) : \widetilde{x} \in \mathbb{R}^{d-1}, \zeta^{\mathrm{per}}(\widetilde{x}) < x_d < H\right\}.$$

Note that the domain $\Omega_H^{\mathrm{per}} \subseteq \mathbb{R}^d$ is 2π-periodic with respect to the first $d-1$ variables. Moreover, the corresponding fundamental cell of periodicity is denoted by

$$\Omega_H^{2\pi} \coloneqq \left\{(\widetilde{x}, x_d) \in \Omega_H^{\mathrm{per}} : \widetilde{x} \in (-\pi, \pi)^d\right\}$$

whose boundary is the union of the vertical boundaries

$$\Gamma^{2\pi} \coloneqq \left\{x \in \Gamma^{\mathrm{per}} \colon \widetilde{x} \in [-\pi, \pi]^{d-1}\right\} \quad \text{and} \quad \Gamma_H^{2\pi} \coloneqq \left\{x \in \Gamma_H \colon \widetilde{x} \in [-\pi, \pi]^{d-1}\right\}$$

and the lateral boundaries

$$\Gamma_- \coloneqq \left\{(\widetilde{x}, x_d) \in \partial\Omega_H^{2\pi} : x_1 = -\pi \text{ or } x_2 = -\pi \text{ or } \ldots \text{ or } x_{d-1} = -\pi\right\},$$
$$\Gamma_+ \coloneqq \left\{(\widetilde{x}, x_d) \in \partial\Omega_H^{2\pi} : x_1 = \pi \text{ or } x_2 = \pi \text{ or } \ldots \text{ or } x_{d-1} = \pi\right\}.$$

The bounded cell $\Omega_H^{2\pi}$ and its boundaries are plotted in two dimensions in Figure 2.5.

To introduce the FB transform as presented in [81, Sec. 6], we first denote by $C_0^\infty(\Omega_H^{\mathrm{per}})$ all the smooth functions that are compactly supported on Ω_H^{per}.

Definition 2.24. For a given function $f \in C_0^\infty(\Omega_H^{\mathrm{per}})$, the *FB transform*, denoted by $\mathcal{J}f$, is defined by

$$(\mathcal{J}f)(\alpha; x) \coloneqq \sum_{j\in\mathbb{Z}^{d-1}} f(\widetilde{x} + 2\pi j, x_d)\, \mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi j)}\,, \quad \alpha \in \mathbb{R}^{d-1},\, x \in \Omega_H^{\mathrm{per}}\,, \tag{2.34}$$

where α is called *Floquet parameter*.

The summation is well defined because the function f is compactly supported. Therefore, the above series reduces to a finite sum. The FB transform is similar to the Fourier series. However, the coefficients in the FB transform are not constant and still depend on the original variable x.

Remark 2.25. In general, the FB transform is defined on fully periodic domains and applied in all spatial directions (see [81, Sec. 2]). However, in our setting, the computational domain is only periodic with respect to the first $d-1$ variables. Therefore, we apply the FB transform only to these variables.

Below, we mention some important properties of the FB transform.

Proposition 2.26. *Let $f \in C_0^\infty(\Omega_H^{\mathrm{per}})$ and $\Lambda \coloneqq [-1/2, 1/2]^{d-1}$. Then, $\mathcal{J}f$ satisfies the following conditions:*

(a) For each fixed $x \in \Omega_H^{2\pi}$, $\alpha \mapsto \mathrm{e}^{\mathrm{i}\alpha\cdot\widetilde{x}}(\mathcal{J}f)(\alpha; x)$ is 1-periodic with fundamental cell of periodicity Λ.

(b) For each fixed $\alpha \in \Lambda$, $x \mapsto (\mathcal{J}f)(\alpha; x)$ is 2π-periodic in the first $d-1$ variables with fundamental cell of periodicity $\Omega_H^{2\pi}$.

Proof. (a) For each fixed x, the function $\alpha \mapsto \mathrm{e}^{\mathrm{i}\alpha\cdot\widetilde{x}}(\mathcal{J}f)(\alpha; x)$ is 1-periodic, since for $\ell \in \mathbb{Z}^{d-1}$

$$\begin{aligned}
\mathrm{e}^{\mathrm{i}(\alpha+\ell)\cdot\widetilde{x}}(\mathcal{J}f)(\alpha+\ell; x) &= \mathrm{e}^{\mathrm{i}(\alpha+\ell)\cdot\widetilde{x}} \sum_{j\in\mathbb{Z}^{d-1}} f(\widetilde{x} + 2\pi j, x_d)\mathrm{e}^{-\mathrm{i}(\alpha+\ell)\cdot(\widetilde{x}+2\pi j)} \\
&= \mathrm{e}^{\mathrm{i}(\alpha+\ell)\cdot\widetilde{x}} \sum_{j\in\mathbb{Z}^{d-1}} f(\widetilde{x} + 2\pi j, x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi j)}\mathrm{e}^{-\mathrm{i}\ell\cdot(\widetilde{x}+2\pi j)} \\
&= \mathrm{e}^{\mathrm{i}\alpha\cdot\widetilde{x}} \sum_{j\in\mathbb{Z}^{d-1}} f(\widetilde{x} + 2\pi j, x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi j)}\mathrm{e}^{-\mathrm{i}\ell\cdot(2\pi j)} \\
&= \mathrm{e}^{\mathrm{i}\alpha\cdot\widetilde{x}}(\mathcal{J}f)(\alpha; x)\,,
\end{aligned}$$

where for the last equality we used the fact that $\mathrm{e}^{-\mathrm{i}\ell\cdot(2\pi j)} = 1$ for all $j, \ell \in \mathbb{Z}^{d-1}$.

(b) For each fixed α, the function $x \mapsto (\mathcal{J}f)(\alpha; x)$ is 2π-periodic, since for $\ell \in \mathbb{Z}^{d-1}$

$$\begin{aligned}
(\mathcal{J}f)(\alpha; \widetilde{x} + 2\pi\ell, x_d) &= \sum_{j\in\mathbb{Z}^{d-1}} f(\widetilde{x} + 2\pi(j+\ell), x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi(j+\ell))} \\
&= \sum_{m\coloneqq j+\ell\in\mathbb{Z}^{d-1}} f(\widetilde{x} + 2\pi m, x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi m)} = (\mathcal{J}f)(\alpha; x)\,.
\end{aligned}$$

□

To analyze the mapping properties of the FB transform, we first introduce the space $L^2(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))$, which contains square-integrable functions from Λ to $H^s_{\mathrm{per}}(\Omega_H^{2\pi})$ equipped with the following inner product (see [99, Sec. 39.2])

$$\langle f, \overline{g} \rangle_{L^2(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))} = \int_\Lambda \left\langle f(\alpha), \overline{g(\alpha)} \right\rangle_{H^s_{\mathrm{per}}(\Omega_H^{2\pi})} \mathrm{d}\alpha \, .$$

To simplify notation, we write $f(\alpha) \in H^s_{\mathrm{per}}(\Omega_H^{2\pi})$ for $f \in L^2(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))$, while continuing to use the notation $f(\alpha; x)$ instead of $f(\alpha)(x)$.

Definition 2.27. The space $H^r_\alpha(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))$ for $r \geq 0$ and $s \in \mathbb{R}$ consists of the functions $f \in L^2(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))$ such that $\alpha \mapsto f(\alpha; \cdot)\mathrm{e}^{\mathrm{i}\alpha\cdot}$ is 1-periodic with fundamental cell of periodicity Λ and the following norm is finite

$$\|f\|_{H^r_\alpha(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))} \coloneqq \left\| \alpha \mapsto \|f(\alpha)\|_{H^s_{\mathrm{per}}(\Omega_H^{2\pi})} \right\|_{H^r_\alpha(\Lambda)} .$$

For $r < 0$, $H^r_\alpha(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))$ is the dual of $H^{-r}_\alpha(\Lambda; H^{-s}_{\mathrm{per}}(\Omega_H^{2\pi}))$ and

$$\|f\|_{H^r_\alpha(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))} \coloneqq \sup_{g \in H^{-r}_\alpha(\Lambda; H^{-s}_{\mathrm{per}}(\Omega_H^{2\pi}))} \langle f, \overline{g} \rangle_{\Lambda \times \Omega_H^{2\pi}} \, .$$

Theorem 2.28. *Let Λ as in Proposition 2.26.*

(a) The FB transform from $C_0^\infty(\Omega_H^{\mathrm{per}})$ can be extended to an isometry between $L^2(\Omega_H^{\mathrm{per}})$ and $L^2(\Lambda, L^2_{\mathrm{per}}(\Omega_H^{\mathrm{per}}))$ and its inverse transform is obtained by

$$\mathcal{J}^{-1} f(\widetilde{x} + 2\pi j, x_d) = \int_\Lambda f(\alpha; x) \mathrm{e}^{\mathrm{i}\alpha\cdot(\widetilde{x} + 2\pi j)} \, \mathrm{d}\alpha \, , \quad x \in \Omega_H^{2\pi} \text{ and } j \in \mathbb{Z}^{d-1}. \tag{2.35}$$

(b) For $s, r \in \mathbb{R}$, the FB transform from $C_0^\infty(\Omega_H^{\mathrm{per}})$ can be extended to an isomorphism between $H^s_r(\Omega_H^{\mathrm{per}})$ and $H^r_\alpha(\Lambda; H^s_{\mathrm{per}}(\Omega_H^{2\pi}))$ and its inverse transform is obtained by (2.35).

(c) For any $f, g \in L^2(\Omega_H^{\mathrm{per}})$, the Plancherel formula holds, i.e.,

$$\int_{\Omega_H^{\mathrm{per}}} f(x) \overline{g(x)} \, \mathrm{d}x \, = \int_\Lambda \int_{\Omega_H^{2\pi}} (\mathcal{J} f)(\alpha; x) \overline{(\mathcal{J} g)(\alpha; x)} \, \mathrm{d}x \, \mathrm{d}\alpha \, . \tag{2.36}$$

Proof. See [81, Thm. 8] for parts (a) and (b) and [12, p. 220] for (c). □

Note that the Plancherel formula also holds for $f \in H^s_r(\Omega_H^{\mathrm{per}})$ and $g \in H^{-s}_{-r}(\Omega_H^{\mathrm{per}})$ due to the density of $L^2(\Omega_H^{\mathrm{per}})$ in $L^2_{-r}(\Omega_H^{\mathrm{per}})$ for $r \geq 0$ and of $L^2_{-r}(\Omega_H^{\mathrm{per}})$ in $H^{-s}_{-r}(\Omega_H^{\mathrm{per}})$ for $s \geq 0$ (see the proof of [81, Thm. 4]).

Remark 2.29. The mapping properties of the FB transform when operating on functions defined on a flat surface Γ_H or a periodic surface Γ^{per} are analogous (see [81, Sec. 5]).

Theorem 2.30. *Let $q : \Omega_H^{\mathrm{per}} \to \mathbb{C}$ be a 2π-periodic function in the first $d-1$ variables. Then*

$$(\mathcal{J}(qf))(\alpha; x) = q(x)(\mathcal{J} f)(\alpha; x) \, .$$

Proof. By using the definition of the FB transform given in (2.34), a direct calculation yields

$$(\mathcal{J}(qf))(\alpha;x) = \sum_{j\in\mathbb{Z}^{d-1}} (qf)(\widetilde{x}+2\pi j, x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi j)} = q(x)(\mathcal{J}f)(\alpha;x)\,.$$

□

Theorem 2.31. *For any $f \in H^1(\Omega_H^{\mathrm{per}})$, it holds*

$$(\mathcal{J}\partial_{x_\ell} f)(\alpha;x) = \begin{cases} (\partial_{x_\ell} + \mathrm{i}\alpha_\ell)(\mathcal{J}f)(\alpha;x) & \text{for } \ell \in \{1,\dots,d-1\}\,, \\ (\partial_{x_\ell}\mathcal{J}f)(\alpha;x) & \text{for } \ell = d\,. \end{cases}$$

Proof. For $\ell \in \{1,\dots,d-1\}$, the definition of the FB transform given in (2.34) and a straightforward computation yield

$$\begin{aligned}(\mathcal{J}\partial_{x_\ell} f)(\alpha;x) &= \sum_{j\in\mathbb{Z}^{d-1}} (\partial_{x_\ell} f)(\widetilde{x}+2\pi j, x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi j)} \\ &= \sum_{j\in\mathbb{Z}^{d-1}} \partial_{x_\ell}\left(f(\widetilde{x}+2\pi j, x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi j)}\right) \\ &\quad + \mathrm{i}\alpha_\ell \sum_{j\in\mathbb{Z}^{d-1}} f(\widetilde{x}+2\pi j, x_d)\mathrm{e}^{-\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi j)} \\ &= (\partial_{x_\ell} + \mathrm{i}\alpha_\ell)(\mathcal{J}f)(\alpha;x)\,.\end{aligned}$$

The statement for $\ell = d$ follows from the fact that the FB transform acts only on the first $d-1$ variables. □

After having defined the FB transform and its properties, we focus on its effect on differential operators. We consider the equation $\mathcal{L}u = f$ in Ω_H^{per} together with the boundary condition $u = g$ on $\Gamma^{\mathrm{per}} \cup \Gamma_H$, where the differential operator is $\mathcal{L} \coloneqq \nabla\cdot(p\nabla) + q$ with periodic functions p and q.

By applying the FB transform to $\mathcal{L}u$ and using Theorems 2.30 and 2.31, we obtain the following family of boundary value problems indexed by the Floquet parameter $\alpha \in \Lambda$

$$\begin{aligned}\mathcal{L}_\alpha(\mathcal{J}u)(\alpha;x) &= (\mathcal{J}f)(\alpha;x) && \text{for } x \in \Omega_H^{\mathrm{per}}\,, \\ (\mathcal{J}u)(\alpha;x) &= (\mathcal{J}g)(\alpha;x) && \text{for } x \in \Gamma^{\mathrm{per}} \cup \Gamma_H, \\ (\mathcal{J}u)(\alpha;x) &= (\mathcal{J}u)(\alpha;\widetilde{x}+2\pi j, x_d) && \text{for } j \in \mathbb{Z}^d, x \in \Omega_H^{\mathrm{per}}\,,\end{aligned}$$

where $\mathcal{L}_\alpha$ is acting like a shifted operator, defined by

$$\mathcal{L}_\alpha \coloneqq (\nabla_{\widetilde{x}} + \mathrm{i}\alpha)\cdot\left(p(x)(\nabla_{\widetilde{x}} + \mathrm{i}\alpha)\right) + \partial_{x_d}\left(p(x)\partial_{x_d}\right) + q(x)$$

and $\nabla_{\widetilde{x}}$ is the gradient with respect to the first $d-1$ variables.

Due to the periodicity of $\mathcal{J}u$ with respect to $\widetilde{x}$, the above problem can be reduced to the bounded cell $\Omega_H^{2\pi}$ (depicted in Figure 2.5b) as follows

$$\begin{aligned}\mathcal{L}_\alpha(\mathcal{J}u)(\alpha;x) &= (\mathcal{J}f)(\alpha;x) && \text{for } x \in \Omega_H^{2\pi}\,, \\ (\mathcal{J}u)(\alpha;x) &= (\mathcal{J}g)(\alpha;x) && \text{for } x \in \Gamma^{2\pi} \cup \Gamma_H^{2\pi}\,,\end{aligned}$$

together with the periodic boundary conditions on the lateral boundaries

$$(\mathcal{J}u)(\alpha;x)|_{\Gamma_+} = (\mathcal{J}u)(\alpha;x)|_{\Gamma_-},$$
$$\partial_{\widetilde{x}}(\mathcal{J}u)(\alpha;x)|_{\Gamma_+} = \partial_{\widetilde{x}}(\mathcal{J}u)(\alpha;x)|_{\Gamma_-}.$$

CHAPTER 3

Scattering in Unbounded Periodic Structures

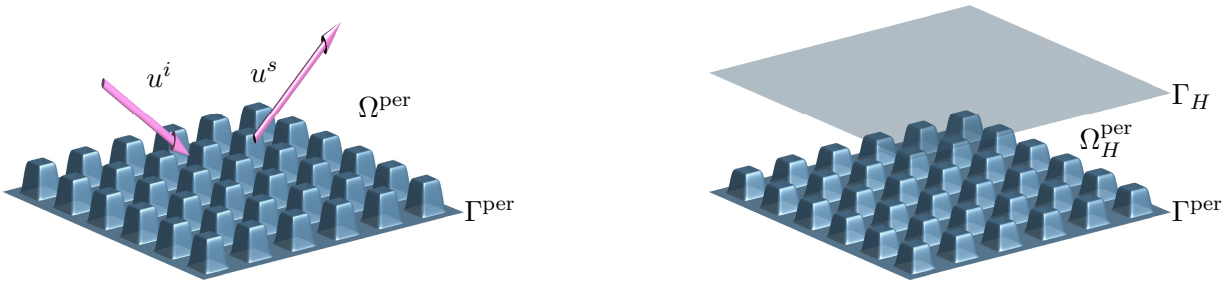
In this chapter, we study acoustic wave scattering from unbounded periodic surfaces. The general setting is described in Chapter 1; however, we place particular emphasis here on the periodicity of the geometry, using the notations Γ^{per}, ζ^{per} and Ω^{per}. For simplicity, we assume the fundamental period of ζ^{per} is 2π. Consequently, the unbounded domain $\Omega^{\mathrm{per}} \subseteq \mathbb{R}^d$ for $d = 2, 3$ is 2π-periodic with respect to the first $d-1$ variables. This setting is depicted in Figure 3.1(a).

We first employ the DtN map and the FB transform to derive a family of periodic problems posed in a single bounded cell. In order to propose a high-order numerical method, we analyze the regularity of the transformed field with respect to the Floquet parameter. It should be pointed out that the regularity of the transformed field in two dimensions is less complicated than in three dimensions. In [104, Thm. 11], it has been proven that the transformed field in two dimensions is analytic except for at most two singular points. However, in three dimensions, singularities of the transformed field no longer consist of a finite number of points. Rather, they form a set that is the union of a finite number of circular arcs. Therefore, a direct extension of the high-order numerical methods used for the two-dimensional case (in [7, Sec. 5, 104, Sec. 4]) is not possible for the three-dimensional case.

Our first main result, in Theorem 3.6, is a local representation of the transformed field mirroring the expected structure of singularities. This significantly extends similar representations found in [73, Satz 3.11, 74, Thm. 22]. Moreover, we obtain a globally valid representation in Theorem 3.9. Based on the regularity results, we construct a tailor-made quadrature rule, adapted to the singularity structure of the transformed field, to compute the inversion of the FB transform. We present some numerical examples illustrating the performance of this scheme.

3.1. Formulation in a Bounded Cell

To truncate the unbounded domain Ω^{per} in the vertical direction, we impose a transparent boundary condition on a flat surface Γ_H at height H. As explained in Section 2.2, this condition is expressed by using the DtN map. The resulting truncated domain is the unbounded periodic domain Ω_H^{per} between Γ^{per} and Γ_H (depicted in Figure 3.1(b)).



(a) Unbounded domain Ω^{per} (b) Unbounded domain Ω_H^{per}

FIGURE 3.1. Sketch of the unbounded periodic domains.

For a given incident field $u^i \in H^1_r(\Omega_H^{\mathrm{per}})$ with $|r| < 1$, we seek the unknown total field $u \in \widetilde{H}^1_r(\Omega_H^{\mathrm{per}})$, which satisfies

$$\Delta u + k^2 u = 0 \qquad \text{in } \Omega_H^{\mathrm{per}}, \tag{3.1a}$$

$$u = 0 \qquad \text{on } \Gamma^{\mathrm{per}}, \tag{3.1b}$$

$$(\partial_{x_d} - \mathcal{T}^+)u = (\partial_{x_d} - \mathcal{T}^+)u^i \qquad \text{on } \Gamma_H\,, \tag{3.1c}$$

where the DtN map $\mathcal{T}^+$ is defined as in (2.9). Note that (3.1a) is understood in the variational sense and (3.1b) and (3.1c) in the trace sense.

The variational formulation of this boundary value problem is similar to (2.13) but posed in the periodic domain Ω_H^{per}. To simplify the notation, we will omit writing the subscript r for the sesquilinear form a_r given in (2.14). More precisely, we aim to find $u \in \widetilde{H}^1_r(\Omega_H^{\mathrm{per}})$ such that

$$a(u,v) = \left\langle (\partial_{x_d} - \mathcal{T}^+)u^i, \overline{v} \right\rangle_{\Gamma_H} \quad \text{for all } v \in \widetilde{H}^1_{-r}(\Omega_H^{\mathrm{per}})\,, \tag{3.2}$$

where the sesquilinear form $a\colon \widetilde{H}^1_r(\Omega_H^{\mathrm{per}}) \times \widetilde{H}^1_{-r}(\Omega_H^{\mathrm{per}}) \to \mathbb{C}$ is defined by

$$a(\phi,\psi) := \left\langle \nabla\phi, \overline{\nabla\psi} \right\rangle_{\Omega_H^{\mathrm{per}}} - k^2 \left\langle \phi, \overline{\psi} \right\rangle_{\Omega_H^{\mathrm{per}}} - \left\langle \mathcal{T}^+\phi, \overline{\psi} \right\rangle_{\Gamma_H}.$$

This problem is uniquely solvable as shown in Theorem 2.15.

From a numerical point of view, the variational problem (3.2) is not yet adequate as it is still posed in the unbounded domain Ω_H^{per}. Since this domain is periodic with respect to its first $d-1$ variables (denoted by $\widetilde{x}$), we can apply the FB transform only with respect to $\widetilde{x}$, as in Definition 2.24. This leads to a decomposed formulation of (3.2) consisting of a family of periodic problems (indexed by the Floquet parameter α) posed in a single bounded unit cell of the periodicity. We recall from Section 2.4 the notation $\Omega_H^{2\pi}$ for the unit bounded cell whose bottom and top surfaces are denoted by $\Gamma^{2\pi}$ and $\Gamma_H^{2\pi}$, respectively. We depict a sketch of the bounded cell in Figure 3.2 for the three-dimensional case.

Let the FB transform of the total field u be denoted by $w := \mathcal{J}u$. According to Proposition 2.26, $w(\alpha;x)$ is 2π-periodic in $\widetilde{x}$ and $\mathrm{e}^{\mathrm{i}\alpha\cdot\widetilde{x}} w(\alpha;x)$ is 1-periodic in α. Therefore, the fundamental cell of the periodic function w is assumed to be $\Lambda \times \Omega_H^{2\pi}$, where $\Lambda := [-1/2, 1/2]^{d-1}$.

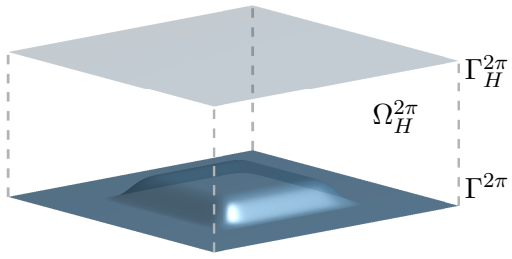


Figure 3.2. A three-dimensional bounded unit cell $\Omega_H^{2\pi}$.

By applying the FB transform, using the Plancherel formula (2.36) and Theorem 2.31, we obtain the following variational problem for $w \in H^r_\alpha(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$

$$\int_\Lambda a_\alpha(w(\alpha), z(\alpha))\,\mathrm{d}\alpha = \int_\Lambda \Big\langle (\partial_{x_d} - \mathcal{T}_\alpha^+)\mathcal{J}u^i(\alpha), \overline{z(\alpha)} \Big\rangle_{\Gamma_H^{2\pi}} \mathrm{d}\alpha \tag{3.3}$$

for all $z \in H^{-r}_\alpha(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, where

$$a_\alpha(\phi,\psi) \coloneqq \Big\langle \nabla\phi, \overline{\nabla\psi} \Big\rangle_{\Omega_H^{2\pi}} - 2\mathrm{i}\Big\langle \alpha\cdot\nabla_{\widetilde{x}}\phi, \overline{\psi} \Big\rangle_{\Omega_H^{2\pi}} - (k^2 - |\alpha|^2)\Big\langle \phi, \overline{\psi} \Big\rangle_{\Omega_H^{2\pi}} - \Big\langle \mathcal{T}_\alpha^+\phi, \overline{\psi} \Big\rangle_{\Gamma_H^{2\pi}}, \tag{3.4}$$

with $\nabla_{\widetilde{x}} \coloneqq (\partial_{x_1}, \dots, \partial_{x_{d-1}})^\top$. Note that the periodic version of the DtN map, denoted by $\mathcal{T}_\alpha^+ \colon H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi}) \to H^{-1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})$, is defined by

$$(\mathcal{T}_\alpha^+\varphi)(\widetilde{x}, H) \coloneqq \mathrm{i}\sum_{j\in\mathbb{Z}^{d-1}} \sqrt{k^2 - |\alpha - j|^2}\,\widehat{\varphi}(j)\,\mathrm{e}^{\mathrm{i}\widetilde{x}\cdot j} \quad \text{for} \quad \varphi(\widetilde{x}, H) = \sum_{j\in\mathbb{Z}^{d-1}} \widehat{\varphi}(j)\,\mathrm{e}^{\mathrm{i}\widetilde{x}\cdot j}, \tag{3.5}$$

where $\widehat{\phi}(j)$ denotes the j-th Fourier coefficient of ϕ (see [84, Eq. (11)]).

Remark 3.1. The right-hand side of (3.3) is understood as the dual pairing in $H^r_\alpha(\Lambda; H^{-1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})) \times H^{-r}_\alpha(\Lambda; H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi}))$.

Remark 3.2. The strong form of (3.3) is

$$\Delta_x w(\alpha) + 2\mathrm{i}\,\alpha\cdot\nabla_{\widetilde{x}} w(\alpha) + (k^2 - |\alpha|^2)w(\alpha) = 0 \qquad \text{in } \Omega_H^{2\pi}, \tag{3.6a}$$

$$w(\alpha) = 0 \qquad \text{on } \Gamma^{2\pi}, \tag{3.6b}$$

$$(\partial_{x_d} - \mathcal{T}_\alpha^+)w(\alpha) = (\partial_{x_d} - \mathcal{T}_\alpha^+)\mathcal{J}u^i(\alpha) \qquad \text{on } \Gamma_H^{2\pi}. \tag{3.6c}$$

Equation (3.6a) is understood in the variational sense and (3.6b) and (3.6c) in the trace sense.

Theorem 3.3. *Let $|r| < 1$ and $u^i \in H^1_r(\Omega_H^{\mathrm{per}})$. A function $u \in \widetilde{H}^1_r(\Omega_H^{\mathrm{per}})$ satisfies (3.2) if and only if $w \in H^r_\alpha(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ is a solution to the variational problem (3.3). Moreover, if $\mathcal{J}u^i$ is*

continuous with respect to α, then w is also continuous with respect to α, and for every $\alpha \in \Lambda$,

$$a_\alpha(w(\alpha), z) = \left\langle (\partial_{x_d} - \mathcal{T}_\alpha^+)\mathcal{J}u^i(\alpha), \overline{z} \right\rangle_{\Gamma_H^{2\pi}} \quad \text{for all } z \in \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}), \tag{3.7}$$

where the sesquilinear form a_α is defined in (3.4).

Proof. See [84, Thm. 2]. □

Unique solvability of the variational problem (3.7) in $\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})$ has been proven in [16, Sec. 3.5] and [41, Cor. 3.4] for any arbitrary but fixed $\alpha \in \Lambda = [-1/2, 1/2]^{d-1}$. We can thus compute numerical approximations to the transformed field $w(\alpha)$ for every $\alpha \in \Lambda$ by using some standard numerical method. Afterwards, these transformed fields are combined by means of the inverse FB transform (2.35), which yields an approximation to the solution of (3.3). This essentially amounts to the evaluation of an integral of w over the domain Λ. The accuracy of the numerical solution of (3.3) depends not only on the selected numerical method for solving (3.7), but also on the accuracy of the numerical integration method employed for this integral. In order to construct a high-order numerical scheme, requiring few quadrature points for high accuracy, it is necessary to precisely know the regularity of the transformed field with respect to the Floquet parameter α.

3.2. Regularity of the Transformed Solution

Let us heuristically motivate the results that we shall make rigorous in Theorem 3.6. From the definition of a_α in (3.4), we see that all terms in the variational formulation (3.7) depend analytically on α except for the square root functions in $\mathcal{T}_\alpha^+$ as defined in (3.5). Hence, we may expect the transformed field w to depend analytically on α, except for points where (the derivatives of) these functions have singularities, i.e., except for points located in the set Σ defined by

$$\Sigma := \left\{ \alpha \in \Lambda = [-1/2, 1/2]^{d-1} \,:\, |\alpha - j| = k \text{ for some } j \in \mathbb{Z}^{d-1} \right\}. \tag{3.8}$$

In the two-dimensional case ($d = 2$), when k is a half-integer, the set Σ includes at most two singular points, whereas for non-half-integer k, the set Σ has exactly two singular points [104, Sec. 3.2]. By increasing k, the number of the singular points in Σ does not change. However, we show in the following that the structure of singularity in the three-dimensional case ($d = 3$) is much more complicated.

For $d = 3$, the set Σ is a union of circular arcs formed by the intersection of Λ and circles with center j and radius k. We will refer to this set as the *curves of singular points*. Figure 3.3 illustrates possible structures of Σ on Λ for different wave numbers k. Any high-order method for approximately inverting the FB transform needs to take the structure of Σ into account, as it becomes more and more complex as k increases.

For any $\alpha \in \Sigma$, we also define

$$\mathbf{J}(\alpha) := \{ j \in \mathbb{Z}^2 \,:\, |\alpha - j| = k \}, \tag{3.9}$$

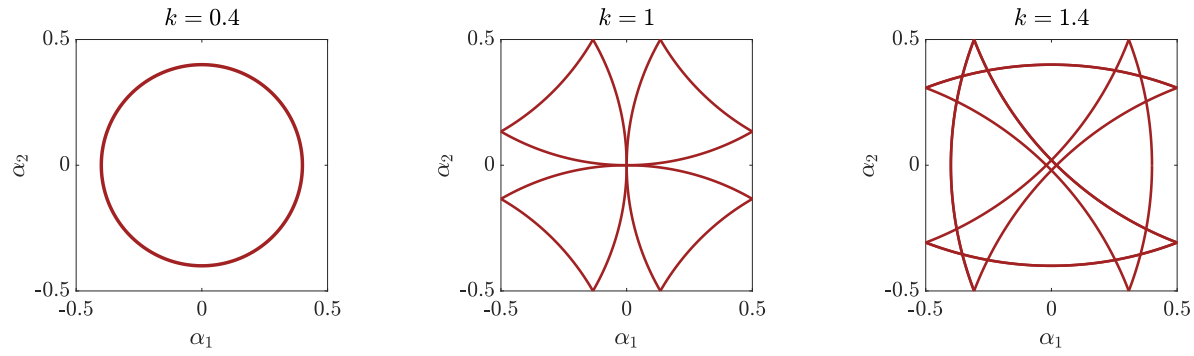


Figure 3.3. Structure of Σ for different values of k on Λ.

which is a finite set with cardinality $\#\mathbf{J}(\alpha)$.

Remark 3.4. When $k < 1/2$, $\#\mathbf{J}(\alpha) = 1$ for all $\alpha \in \Sigma$. When $k \geq 1/2$, there exists a finite number of $\alpha \in \Sigma$ with $\#\mathbf{J}(\alpha) > 1$.

For the later analysis of the numerical inversion of the FB transform, we require a particular regularity of both the transformed incident and the transformed total field. To formulate these requirements, we introduce the following definitions:

Definition 3.5. For some open set $U \subseteq \mathbb{R}^2$ and a Hilbert space V, we denote by $C^\omega(U; V)$ the space of functions from U to V that depend analytically on $\alpha \in U$. For a Hilbert space V, let

$$\mathcal{X}(V) \coloneqq \{f \colon \Lambda \to V : f \text{ satisfies (C1) and (C2)}\}, \tag{3.10}$$

where

(C1) for every open subdomain $U \subseteq \Lambda \setminus \Sigma$, $f \in C^\omega(U; V)$,

(C2) for any $\alpha_0 \in \Sigma$, there exists a neighborhood U_0 of α_0 such that

$$f(\alpha) = \sum_{\mathcal{I} \subseteq \mathbf{J}(\alpha_0)} \left(\prod_{j \in \mathcal{I}} \sqrt{k^2 - |\alpha - j|^2} \right) f_{\mathcal{I}}(\alpha) \tag{3.11}$$

for some $f_{\mathcal{I}} \in C^\omega(U_0; V)$ for each $\mathcal{I} \subseteq \mathbf{J}(\alpha_0)$.

Theorem 3.6. *Let $u^i \in H^1_r(\Omega_H^{\mathrm{per}})$ for some $|r| < 1$ and additionally $\mathcal{J}u^i \in \mathcal{X}(H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi}))$. Then, the transformed total field w that solves (3.3) satisfies $w \in \mathcal{X}(\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$.*

Proof. Let $\alpha_0 \in \Lambda$. The sesquilinear form (3.4) can be written as

$$a_\alpha(\cdot, \cdot) = b_\alpha(\cdot, \cdot) - \sum_{j \in \mathbb{Z}^2} \sqrt{k^2 - |\alpha - j|^2} c_j(\cdot, \cdot),$$

where

$$b_\alpha(\phi, \psi) \coloneqq \left\langle \nabla\phi, \overline{\nabla\psi} \right\rangle_{\Omega_H^{2\pi}} - 2\mathrm{i}\left\langle \alpha \cdot \nabla_{\widetilde{x}} \phi, \overline{\psi} \right\rangle_{\Omega_H^{2\pi}} - (k^2 - |\alpha|^2)\left\langle \phi, \overline{\psi} \right\rangle_{\Omega_H^{2\pi}},$$
$$c_j(\phi, \psi) \coloneqq \mathrm{i}\left\langle \widehat{\phi}(j)\, \mathrm{e}^{\mathrm{i}\widetilde{x}\cdot j}, \overline{\psi(x)} \right\rangle_{\Gamma_H^{2\pi}}.$$

We may define the operators $\mathcal{A}(\alpha)\colon \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}) \to (\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))^*$ and $\mathcal{C}(j)\colon H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi}) \to H^{-1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})$ by

$$\left\langle \mathcal{C}(j)\phi, \overline{\psi} \right\rangle_{\Gamma_H^{2\pi}} := c_j(\phi,\psi)$$

and

$$\begin{aligned}\left\langle \mathcal{A}(\alpha)\phi, \overline{\psi} \right\rangle_{\Omega_H^{2\pi}} &:= b_\alpha(\phi,\psi) \; - \sum_{j\in\mathbb{Z}^2\setminus\mathbf{J}(\alpha_0)} \sqrt{k^2-|\alpha-j|^2}\left\langle \mathcal{C}(j)\phi, \overline{\psi} \right\rangle_{\Gamma_H^{2\pi}} \\ &= a_\alpha(\phi,\psi) + \sum_{j\in\mathbf{J}(\alpha_0)} \sqrt{k^2-|\alpha-j|^2}\left\langle \mathcal{C}(j)\phi, \overline{\psi} \right\rangle_{\Gamma_H^{2\pi}}.\end{aligned}$$

Clearly, in a neighborhood of α_0, $\mathcal{A}(\alpha)$ depends analytically on α. Using these operators and also the antilinear form $\mathcal{G}(\alpha)$ induced by the right-hand side of (3.7), this equation can be reformulated as

$$\left[\mathcal{A}(\alpha) - \sum_{j\in\mathbf{J}(\alpha_0)} \sqrt{k^2-|\alpha-j|^2}\,\mathcal{C}(j)\right] w(\alpha) = \mathcal{G}(\alpha)\,. \tag{3.12}$$

If $\alpha_0 \notin \Sigma$, then $\mathbf{J}(\alpha_0) = \emptyset$ and as $\mathcal{J}u^i$ satisfies (C1) with $V = H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})$, so w also satisfies (C1) with $V = \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})$. We now assume $\alpha_0 \in \Sigma$, i.e., $|\alpha_0 - j| = k$ for some $j \in \mathbf{J}(\alpha_0)$. Moreover, let $B(\alpha_0,\rho)$ denote an open ball centred at α_0 with radius ρ. Then, for any $j \in \mathbf{J}(\alpha_0)$, there holds

$$\left\| \sqrt{k^2-|\alpha-j|^2}\,\mathcal{C}(j) \right\| \to 0 \quad \text{as } |\alpha-\alpha_0| \to 0\,.$$

In [66, Thm. 3], it has been shown that the operator on the left-hand side of (3.12) for all $\alpha \in \Lambda$ is boundedly invertible. We know that $\mathbf{J}(\alpha_0)$ is a finite set and for small enough ρ, the operator $\mathcal{A}(\alpha)$ for all $\alpha \in B(\alpha_0,\rho)$ is a small perturbation of the operator on the left-hand side of (3.12). Therefore, we can use the perturbation theorem given in [75, Thm. 10.1] and conclude that the operator $\mathcal{A}(\alpha)$ is also boundedly invertible for all $\alpha \in B(\alpha_0,\rho)$.

Setting $\widetilde{\mathcal{C}}(j) = (\mathcal{A}(\alpha))^{-1}\mathcal{C}(j)$, we can write the solution w as the Neumann series

$$w(\alpha) = \sum_{n=0}^{\infty} \left(\sum_{j\in\mathbf{J}(\alpha_0)} \sqrt{k^2-|\alpha-j|^2}\;\widetilde{\mathcal{C}}(j) \right)^n (\mathcal{A}(\alpha))^{-1}\mathcal{G}(\alpha)\,.$$

Let $m := \#\mathbf{J}(\alpha_0)$. Applying the multinomial theorem [93, Sec. 26.4] leads to

$$w(\alpha) = \sum_{n=0}^{\infty} \left(\sum_{\substack{\mathrm{K}_1+\mathrm{K}_2+\cdots+\mathrm{K}_m=n,\\ \mathrm{K}_1,\ldots,\mathrm{K}_m\geq 0}} \frac{n}{\mathrm{K}_1!\mathrm{K}_2!\ldots\mathrm{K}_m!} \prod_{\mu=1}^{m} \left(\sqrt{k^2-|\alpha-j_\mu|^2}\,\widetilde{\mathcal{C}}(j_\mu) \right)^{\mathrm{K}_\mu} \right) (\mathcal{A}(\alpha))^{-1}\mathcal{G}\,.$$

Note that all even powers of the square root functions are analytic. Inserting the representation (3.11) for $\mathcal{G}$ into the above equation and combining all analytic terms appropriately into functions $w_{\mathcal{I}}$ for $\mathcal{I} \subseteq \mathbf{J}(\alpha_0)$, gives that w satisfies (C2) with $V = \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})$. □

Remark 3.7. For the two dimensional case, the representation of the transformed solution $w(\alpha)$ contains only one square root function as shown in [7, Sec. 3.2].

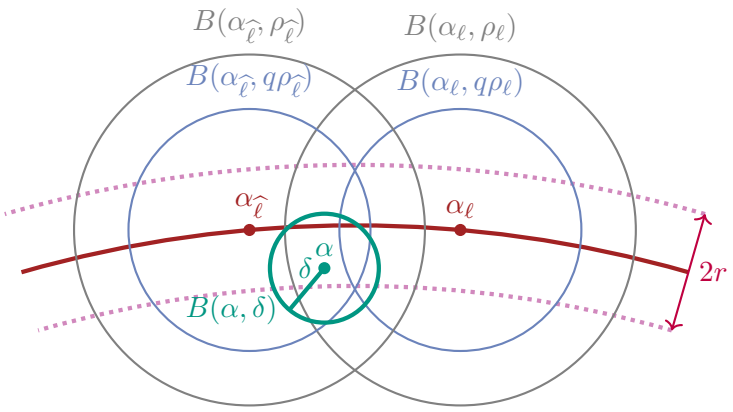


Figure 3.4. Parameters r and δ in the global representation of w

Following up on the previous result, the next theorem guarantees that we can make use of (3.11) for w with the same center of expansion in small balls contained in a neighborhood of Σ.

Theorem 3.8. *There exist open balls $B_\ell = B(\alpha_\ell, \rho_\ell)$ with center points $\alpha_\ell \in \Sigma$ and radii ρ_ℓ, $\ell = 1, \dots, L$, such that $\Sigma \subseteq \bigcup_{\ell=1}^{L} B_\ell$ and the representation (3.11) holds for w on B_ℓ with $\alpha_0 = \alpha_\ell$. Moreover, there exist r, $\delta > 0$ such that*

$$\widetilde{\Sigma} := \{\alpha' \in \Lambda \,:\, \operatorname{dist}(\alpha', \Sigma) < r\} \subseteq \bigcup_{\ell=1}^{L} B_\ell$$

and that for every $\alpha \in \widetilde{\Sigma}$ there exists ℓ with $B(\alpha, \delta) \subseteq B_\ell$.

Proof. For every $\alpha_0 \in \Sigma$, we choose $\rho(\alpha_0) > 0$ such that the representation (3.11) holds for w on $B(\alpha_0, \rho(\alpha_0))$. Then, $\Sigma \subseteq \bigcup_{\alpha_0 \in \Sigma} B(\alpha_0, \rho(\alpha_0))$. Since Σ is a compact set, we can select a finite number of points α_ℓ and radii $\rho_\ell = \rho(\alpha_\ell)$, $\ell = 1, \dots, L$, such that $\Sigma \subseteq \bigcup_{\ell=1}^{L} B(\alpha_\ell, \rho_\ell)$. This yields the first part of the theorem.

Choose $q \in (0, 1)$ such that still $\Sigma \subseteq \bigcup_{\ell=1}^{L} B(\alpha_\ell, q\rho_\ell)$. Choose r such that $\widetilde{\Sigma} \subseteq \bigcup_{\ell=1}^{L} B(\alpha_\ell, q\rho_\ell)$ and set $\delta := (1-q) \min_{\ell=1,\dots,L} \rho_\ell$ (see Figure 3.4). Now, let $\alpha \in \widetilde{\Sigma}$ and $\widehat{\ell}$ such that $|\alpha - \alpha_{\widehat{\ell}}| < q\,\rho_{\widehat{\ell}}$. Then, for any $\alpha' \in B(\alpha, \delta)$, we have

$$\left|\alpha' - \alpha_{\widehat{\ell}}\right| < q\,\rho_{\widehat{\ell}} + \delta = q\,\rho_{\widehat{\ell}} + (1-q) \min_{\ell=1,\dots,L} \rho_\ell \le \rho_{\widehat{\ell}}.$$

This completes the proof. □

The structure of $\widetilde{\Sigma}$ for different values of the wave number k is depicted in Figure 3.5. For any point α in $\widetilde{\Sigma}$, we may use the local representation (3.11) for the transformed field also on a small neighborhood of that point. In our later analysis, we also require a globally valid representation of w, which is provided by the next theorem.

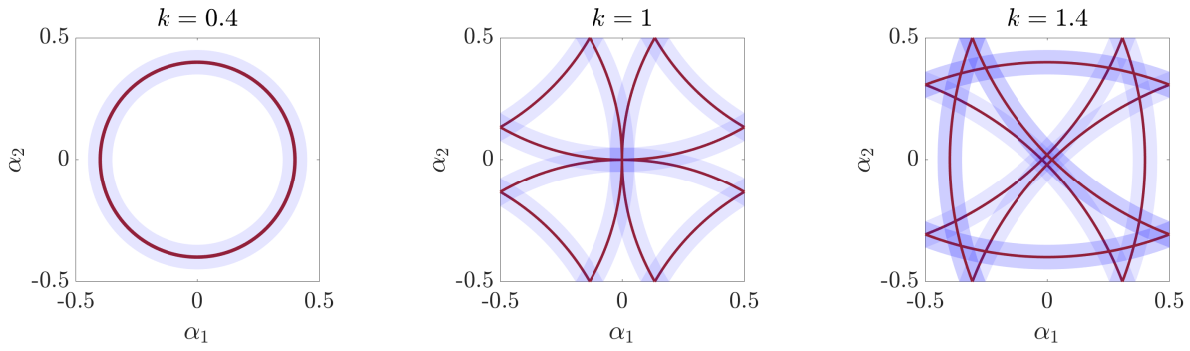


FIGURE 3.5. Structure of $\widetilde{\Sigma}$ for different values of k on Λ.

Theorem 3.9. *Let α_ℓ, $\ell = 1, \dots, L$, denote the points in Theorem 3.8 and set $\mathbf{J} := \bigcup_{\ell=1}^{L} \mathbf{J}(\alpha_\ell)$. Then, there exist $v_{\mathcal{I}} \in C^\infty(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ such that*

$$w(\alpha) = \sum_{\mathcal{I} \subset \mathbf{J}} \left(\prod_{j \in \mathcal{I}} \sqrt{k^2 - |\alpha - j|^2} \right) v_{\mathcal{I}}(\alpha)\,, \qquad \alpha \in \Lambda\,. \tag{3.13}$$

Moreover, for any $\mu \in \mathbb{N}_0$, there exists a constant C_μ such that

$$\left\| \frac{\partial^\mu v_{\mathcal{I}}(\alpha)}{\partial \alpha_\nu^\mu} \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq \frac{C_\mu}{\operatorname{dist}(\alpha, \Sigma)^\mu}\,, \qquad \mathcal{I} \subset \mathbf{J}\,, \quad \nu = 1, 2\,, \quad \alpha \in \Lambda\,. \tag{3.14}$$

Proof. We first recall the covering of Σ by the open balls $B(\alpha_\ell, \delta_\ell)$, for $\ell = 1, \dots, L$, from the proof of Theorem 3.8. Furthermore, let B_0 denote an open subset of $\Lambda \setminus \Sigma$ such that $\Lambda \subseteq B_0 \cup \bigcup_{\ell=1}^{L} B(\alpha_\ell, \delta_\ell)$. Let $\varphi_0, \dots, \varphi_L \in C^\infty(\overline{\Lambda})$ denote a partition of unity subject to this open covering. By Theorem 3.8, in each ball we have

$$w(\alpha) = \sum_{\mathcal{I} \subseteq \mathbf{J}(\alpha_\ell)} \left(\prod_{j \in \mathcal{I}} \sqrt{k^2 - |\alpha - j|^2} \right) w_{\ell,\mathcal{I}}(\alpha)\,, \qquad \alpha \in B(\alpha_\ell, \delta_\ell), \quad \ell = 1, \dots, L$$

with $w_{\ell,\mathcal{I}}$ analytic in $B(\alpha_\ell, \delta_\ell)$. Let $\mathbf{J} = \bigcup_{\ell=1}^{L} \mathbf{J}(\alpha_\ell)$ and define $w_{\ell,\mathcal{I}} = 0$ for $\mathcal{I} \subseteq \mathbf{J}$, but $\mathcal{I} \not\subseteq \mathbf{J}(\alpha_\ell)$, $\ell = 1, \dots, L$. Since the function w on B_0 is itself analytic according to the first part of Theorem 3.6, we set $w_{0,\emptyset} = w$ and $w_{0,\mathcal{I}} = 0$ for all other $\mathcal{I} \subset \mathbf{J}$. Finally, on Λ we define

$$v_{\mathcal{I}} := \sum_{\ell=0}^{L} \varphi_\ell\, w_{\ell,\mathcal{I}}, \qquad \mathcal{I} \subset \mathbf{J}\,, \tag{3.15}$$

where we extend each product on the right-hand side by 0 outside its domain of definition. Then

$$w(\alpha) = \sum_{\mathcal{I} \subset \mathbf{J}} \left(\prod_{j \in \mathcal{I}} \sqrt{k^2 - |\alpha - j|^2} \right) v_{\mathcal{I}}(\alpha)\,, \qquad \alpha \in \Lambda\,.$$

By definition, $v_{\mathcal{I}} \in C^\infty(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$. A standard estimate for analytic functions (see [58,

Thm. 2.2.7]) gives that for some constant C

$$\max_{\alpha\in B(\alpha_\ell,\delta_\ell)}\left\|\frac{\partial^\mu w_{\ell,\mathcal{I}}(\alpha)}{\partial\alpha_\nu^\mu}\right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}\leq\frac{C\,\mu!}{\delta_\ell^\mu}\,,\qquad\nu=1,2\,,\quad\mu\in\mathbb{N}_0\,,\quad\ell=1,\dots,L\,.\tag{3.16}$$

Moreover, we bound each derivative of w on $\overline{B_0}$. For $\alpha\in B_0$, we know that w is analytic in the ball $B(\alpha,\operatorname{dist}(\alpha,\Sigma))$. Hence, again using [58, Thm. 2.2.7] for $\alpha\in B_0$ we end up with

$$\left\|\frac{\partial^\mu w_{\ell,\mathcal{I}}(\alpha)}{\partial\alpha_\nu^\mu}\right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}\leq\frac{\widehat{C}\,\mu!}{\operatorname{dist}(\alpha,\Sigma)^\mu},\quad\nu=1,2\,,\quad\mu\in\mathbb{N}_0\,,\quad\ell=1,\dots,L\,.$$

By considering the definition of $v_{\mathcal{I}}$ given in (3.15), triangle inequality and Leibniz rule, we have for $\nu=1,2$ and $\mu\in\mathbb{N}_0$

$$\begin{aligned}\left\|\frac{\partial^\mu v_{\mathcal{I}}(\alpha)}{\partial\alpha_{\alpha_\nu}^\mu}\right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}&\leq\sum_{\ell=0}^{L}\left\|\frac{\partial^\mu(\phi_\ell w_{\ell,\mathcal{I}})(\alpha)}{\partial\alpha_{\alpha_\nu}^\mu}\right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}\\&\leq\sum_{\ell=0}^{L}\left\|\sum_{n=0}^{\mu}\binom{\mu}{n}\frac{\partial^n\phi_\ell(\alpha)}{\partial\alpha_\nu^n}\frac{\partial^{\mu-n}w_{\ell,\mathcal{I}}(\alpha)}{\partial\alpha_\nu^{\mu-n}}\right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}.\end{aligned}$$

Again using the triangle inequality and applying the bounds on the derivative of the functions φ_ℓ and $w_{\ell,\mathcal{I}}$ given in (3.16), we can write

$$\begin{aligned}\left\|\frac{\partial^\mu v_{\mathcal{I}}(\alpha)}{\partial\alpha_{\alpha_\nu}^\mu}\right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}&\leq\widehat{C}_\mu\sum_{\ell=0}^{L}\sum_{n=0}^{\mu}\left\|\frac{\partial^n\phi_\ell(\alpha)}{\partial\alpha_\nu^n}\frac{\partial^{\mu-n}w_{\ell,\mathcal{I}}(\alpha)}{\partial\alpha_\nu^{\mu-n}}\right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}\\&\leq\widehat{C}_\mu\left(\frac{\widehat{C}\,\mu!}{\operatorname{dist}(\alpha,\Sigma)^\mu}+\sum_{\ell=1}^{L}\sum_{n=0}^{\mu}\frac{(\mu-n)!}{\delta_\ell^{\mu-n}}\right).\end{aligned}$$

Finally, we obtain the assertion as $\delta_\ell\geq\operatorname{dist}(\alpha,\Sigma)$ for each $\ell=1,\dots,L$. □

3.3. A Numerical Inversion of the FB Transform

We hereby propose a high-order numerical scheme to obtain the scattered field. This scheme combines a numerical method, namely the *finite element method*, to compute the transformed field $w(\alpha)$ for fixed α with a *tailor-made quadrature rule* to approximate the inverse FB transform to high order.

The regularity properties of the transformed field reported in the previous section are an essential prerequisite for the derivation of such a rule. According to (2.35), the total field is calculated by means of the inverse FB transform as

$$u(\widetilde{x}+2\pi j,x_3)=\int_\Lambda w(\alpha;x)\,\mathrm{e}^{\mathrm{i}\alpha\cdot(\widetilde{\boldsymbol{x}}+2\pi j)}\;\mathrm{d}\alpha\,,\quad x\in\Omega_H^{2\pi},\;\;j\in\mathbb{Z}^{d-1}\,.\tag{3.17}$$

For an analysis of the approximation of this integral, it obviously suffices to consider the case $j=0$ as the analytic phase factor $\exp(\mathrm{i}\alpha\cdot 2\pi j)$ does not affect the regularity of the integrand.

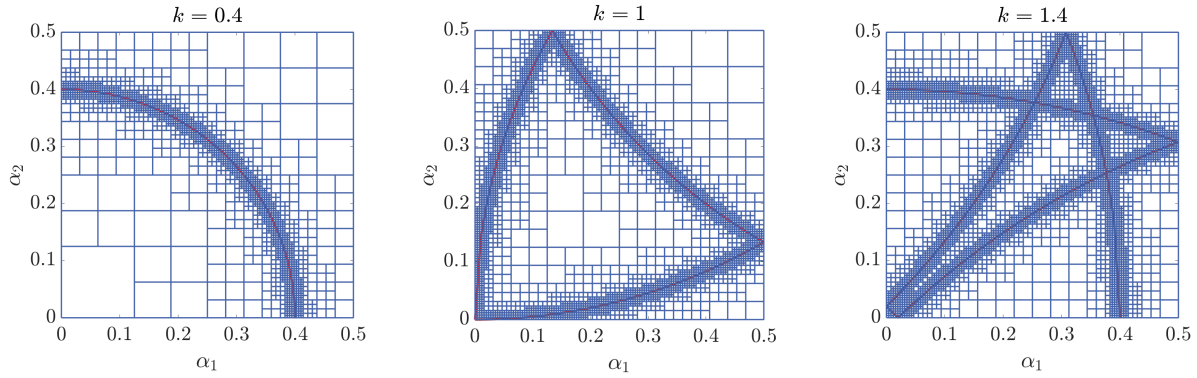


FIGURE 3.6. Generated adapted mesh $\mathcal{G}_6$ for different k by Algorithm 1.

A naive way to approximate the integral in (3.17) is to generate an equidistant uniform square mesh in Λ and then use the set of vertices in this mesh to define a composite trapezoidal rule [73, 74, 84]. However, convergence of such an approach is typically slow: due to the square root singularities present in the representation of $w(\alpha)$ in (3.11), we cannot even attain second order convergence in the mesh width.

We instead propose a specific quadrature rule based on a mesh matching the *a priori* known structure of singularities in w to achieve a high order of convergence. A recursively refined square mesh, dependent only on the wave number, is generated, with elements getting smaller with decreasing distance to the curves of singularities. To approximate the integral in (3.17), a tensor-product trapezoidal rule is applied on the finest squares, whereas a tensor-product Gauss–Legendre rule is used on the remaining squares.

3.3.1. ADAPTIVE MESH GENERATION IN α-SPACE

First, we note that although $\Lambda = [-1/2, 1/2]^2$ it suffices to generate a mesh on $[0, 1/2]^2$ due to the symmetry of the curves of singular points Σ (see Figure 3.3 for an illustration). We start by subdividing $[0, 1/2]^2$ into squares of lateral length $h_0 \coloneqq 1/(2n_0)$ for some $n_0 \in \mathbb{N}_{\geq 2}$. Then N^* refinement steps are performed, further subdividing those squares close to the curves of singular points, which are circular arcs of radius k centred at $j \in \widetilde{\mathbf{J}} \coloneqq \cup_{\alpha \in [0,1/2]^2} \mathbf{J}(\alpha)$. The complete procedure is presented in Algorithm 1, whose output is illustrated in Figure 3.6 for $N^* = 6$ and different values of the wave numbers k.

In Proposition 3.10, we list properties of the adapted mesh $\mathcal{G}_{N^*}$ generated by Algorithm 1. To concisely formulate these results, we introduce the sets of squares of lateral length $h_n \coloneqq h_0/2^n$ in the mesh by

$$\mathcal{M}_n \coloneqq \{Q : Q \in \mathcal{G}_{N^*} \text{ and } Q \text{ has lateral length } h_n\}, \qquad n = 0, \ldots, N^* \tag{3.18}$$

as well as the union of all squares of lateral length h_n,

$$\mathcal{R}_n \coloneqq \bigcup_{Q \in \mathcal{M}_n} \overline{Q}, \qquad n = 0, \ldots, N^*. \tag{3.19}$$

Algorithm 1: generate adapted mesh and tailor-made quadrature rule

Input: k, N^*, n_0, $\widetilde{\mathbf{J}}$

$h_0 \leftarrow 1/(2n_0)$;
$\mathcal{G}_0 \leftarrow \{[\ell_1 h_0, (\ell_1+1)\, h_0] \times [\ell_2 h_0, (\ell_2+1)\, h_0] : \ell_1, \ell_2 = 0, \dots, n_0 - 1\}$;
$\alpha \leftarrow \emptyset$, $\varrho \leftarrow \emptyset$;
for $n = 1, \dots, N^*$ **do**
  $\mathcal{G}_n \leftarrow \emptyset$;
  $h_n \leftarrow h_{n-1}/2$;
  for $Q \in \mathcal{G}_{n-1}$ **do**
    Let C_Q denote the center of Q;
    $\mathrm{dist}(C_Q, \Sigma) \leftarrow \min_{j \in \widetilde{\mathbf{J}}} |k - |C_Q - j||$;
    if $\mathrm{dist}(C_Q, \Sigma) \le 1/2^n$ **then**
      Refine Q into $Q_1, \dots, Q_4$ of lateral length h_n;
      $\mathcal{G}_n \leftarrow \mathcal{G}_n \cup \{Q_1, \dots, Q_4\}$;
    else
      Compute (α_Q, ϱ_Q) corresponding to the Gauss quadrature rule on Q;
      $\alpha \leftarrow \alpha \cup \alpha_Q$;
      $\varrho \leftarrow \varrho \cup \varrho_Q$;
      $\mathcal{G}_n \leftarrow \mathcal{G}_n \cup \{Q\}$;
Compute (α^*, ϱ^*) corresponding to the trapezoidal rule on all squares $Q_{N^*} \setminus Q_{N^*-1}$;
$\alpha \leftarrow \alpha \cup \alpha^*$;
$\varrho \leftarrow \varrho \cup \varrho^*$;
return *Adapted square mesh* $\mathcal{G}_{N^*}$*, quadrature points* α *and weights* ϱ

Proposition 3.10. *Let $Q_n \in \mathcal{M}_n$ (for $n = 0, \dots, N^*$) be squares with centers C_{Q_n}, then*

$$\mathrm{dist}(C_{Q_n}, \Sigma) > \frac{1}{2^{n+1}}, \qquad n = 0, \dots, N^* - 1\,,$$
$$\mathrm{dist}(C_{Q_n}, \Sigma) \le \frac{1}{2^n}\left(1 + \frac{\sqrt{2}}{2} h_0\right), \quad n = 1, \dots, N^*\,.$$

Furthermore,

$$\inf_{x \in \mathcal{R}_n} \mathrm{dist}(x, \Sigma) \ge \frac{1}{2^{n+1}}\left(1 - \sqrt{2} h_0\right) =: d_{\min,n} \quad n = 0, \dots, N^* - 1, \tag{3.20}$$
$$\sup_{x \in \mathcal{R}_n} \mathrm{dist}(x, \Sigma) \le \frac{1}{2^n}\left(1 + \sqrt{2} h_0\right) =: d_{\max,n} \qquad n = 1, \dots, N^*. \tag{3.21}$$

Proof. We consider the square $Q_n \in \mathcal{M}_n$, for $n = 1, \dots, N^*$, with center C_{Q_n}. According to Algorithm 1, Q_n is generated by refining a larger square $Q_{n-1} \in \mathcal{M}_{n-1}$. The center $C_{Q_{n-1}}$ of Q_{n-1} satisfies the condition

$$\mathrm{dist}(C_{Q_{n-1}}, \Sigma) = \left|k - \left|C_{Q_{n-1}} - j\right|\right| \le \frac{1}{2^n} \quad \text{at least for one } j \in \widetilde{\mathbf{J}}\,. \tag{3.22}$$

Based on this refinement, we first conclude that $\operatorname{dist}(C_{Q_n}, C_{Q_{n-1}}) = (\sqrt{2}/4)\, h_{n-1}$ and hence from (3.22) and $h_{n-1} = h_0/2^{n-1}$ that for $n = 1, \dots, N^*$

$$\operatorname{dist}(C_{Q_n}, \Sigma) \le \operatorname{dist}(C_{Q_n}, C_{Q_{n-1}}) + \operatorname{dist}(C_{Q_{n-1}}, \Sigma) \le \frac{1}{2^n}\left(1 + \frac{\sqrt{2}}{2}\, h_0\right). \tag{3.23}$$

A bound for $x \in Q_n$ is obtained by adding half of the diameter of Q_n,

$$\operatorname{dist}(x, \Sigma) \le \frac{\sqrt{2}}{2}\, h_n + \frac{1}{2^n}\left(1 + \frac{\sqrt{2}}{2}\, h_0\right) = \frac{1}{2^n}\left(1 + \sqrt{2}\, h_0\right).$$

As the right-hand side is independent of Q_n, it actually holds for all $x \in \mathcal{R}_n$.

On the other hand, any $Q_n \in \mathcal{M}_n$, $n = 0, \dots, N^* - 1$, that was not subject to the refinement in the $(n+1)$-th refinement step, it implies

$$\operatorname{dist}(C_{Q_n}, \Sigma) > \frac{1}{2^{n+1}}, \qquad n = 0, \dots, N^* - 1\,. \tag{3.24}$$

Hence, for any $x \in Q_n$, we have

$$\operatorname{dist}(x, \Sigma) \ge \operatorname{dist}(C_{Q_n}, \Sigma) - \operatorname{diam}(Q_n)/2 > \frac{1}{2^{n+1}} - \frac{\sqrt{2}}{2}\, h_n = \frac{1}{2^{n+1}}\left(1 - \sqrt{2}\, h_0\right).$$

As the right-hand side is independent of Q_n, this estimate holds for any $x \in \mathcal{R}_n$. □

Remark 3.11. Proposition 3.10 shows that every set $\mathcal{R}_n$ is covered by annuli for which we have explicit bounds for inner and outer radii. As each $\mathcal{R}_n$ is the union of the equally sized squares in $\mathcal{M}_n$, we may estimate the number of squares in $\mathcal{M}_n$. For $n = N^*$, we have

$$|\mathcal{R}_{N^*}| \le \pi\,(k + d_{\max,N^*})^2 - \pi\,(k - d_{\max,N^*})^2 = 4\pi k\, d_{\max,N^*} = \frac{4\pi k}{2^{N^*}}\left(1 + \sqrt{2}\, h_0\right),$$

and hence

$$\#\mathcal{M}_{N^*} = \frac{|\mathcal{R}_{N^*}|}{h_{N^*}^2} \le \frac{4\pi k}{h_0}\left(\sqrt{2} + \frac{1}{h_0}\right) 2^{N^*}.$$

Similarly, for $n = 1, \dots, N^* - 1$, we get

$$|\mathcal{R}_n| \le 4\pi k\,(d_{\max,n} - d_{\min,n}) = \frac{2\pi k}{2^n}\left(1 + 3\sqrt{2}\, h_0\right)$$

and

$$\#\mathcal{M}_n = \frac{|\mathcal{R}_n|}{h_n^2} \le \frac{2\pi k}{h_0}\left(3\sqrt{2} + \frac{1}{h_0}\right) 2^n.$$

3.3.2. Tailor-Made Quadrature Rule and its Convergence Analysis

We will now proceed with defining appropriate quadrature rules on each square in $\mathcal{G}_{N^*}$ and then analyze the corresponding error in computing the integral (3.17). We will strongly rely on the correspondence of the squares in the mesh to representations of the integrand w. In accordance with Theorem 3.8, we may use (3.11) for w on the smallest squares if both $\mathcal{R}_{N^*} \subseteq \widetilde{\Sigma}$

and $h_{N^*} < \sqrt{2}\,\delta$. In the first step, we will use this observation to estimate the error of applying a composite trapezoidal rule on $\mathcal{R}_{N^*}$. Afterwards, we investigate the error of a P-point Gaussian quadrature rule applied on all remaining squares, making use of the representation as derived in Theorem 3.9. Finally, we prove that combining both rules for approximating the inverse FB transform is super-algebraically convergent.

Recall that it suffices to consider the case $j = 0$ when approximating (3.17). Led by the properties of the transformed total field established in Section 2.4, let us first sum up all required assumptions for the integrand. Also recall the definition of the space $\mathcal{X}$ in (3.10).

Assumption 3.12. *We assume that $w \in \mathcal{X}(\widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$ and that r, δ are as in Theorem 3.8. Note that w then will also admit the representation (3.13).*

We first consider a square $Q \in \mathcal{M}_{N^*}$ with center $C_Q = (C_{Q,1}, C_{Q,2})$. The vertices of Q are given by $\alpha_{p,q} = C_Q + (p - \frac{1}{2})\, h_{N^*} \boldsymbol{e}^{(1)} + (q - \frac{1}{2})\, h_{N^*} \boldsymbol{e}^{(2)}$, for p, $q = 0, 1$, where $\boldsymbol{e}^{(j)}$ denotes the j-th coordinate vector. The integral w over Q is approximated by the trapezoidal rule

$$\int_Q w(\alpha)\, \mathrm{d}\alpha = \frac{h^2_{N^*}}{4} \sum_{p,q=0}^{1} w(\alpha_{p,q}) + \mathrm{E}^T_Q w\,,$$

where $E^T_Q w$ denotes the error. To estimate $E^T_Q w$, we require the bilinear interpolation operator P_Q at the points $\alpha_{p,q}$. Well-known estimates for interpolation of any $f \in C^2(\overline{Q})$ give

$$\max_{\alpha \in Q} |f(\alpha) - P_Q f(\alpha)| \leq C\, h^2_{N^*} \max_{\nu=1,2} \left\| \frac{\partial^2 f}{\partial \alpha_\nu^2} \right\|_\infty , \tag{3.25}$$

where $\|\cdot\|_\infty$ denotes the supremum norm. This estimate generalizes to C^2-smooth functions on Q with values in a Sobolev space.

Theorem 3.13. *Let w satisfy Assumption 3.12 and let the parameters h_0, N^* in Algorithm 1 be chosen such that $d_{\max,N^*} < r$ and $h_{N^*} \leq \sqrt{2}\delta$. Then,*

$$\max_{\alpha \in Q} \|w(\alpha) - P_Q w(\alpha)\|_{H^1_{\mathrm{per}}(\Omega^{2\pi}_H)} \leq \mathrm{C}\, 2^{-N^*/2}\,,$$

where the constant C depends on k and the functions $w_{\mathcal{I}}$ appearing in (3.11) for all the centers of the expansion from Theorem 3.8.

Proof. According to Theorems 3.6 and 3.8, there exists $\alpha_0 \in \Sigma$ such that the representation

$$w(\alpha) = \sum_{\mathcal{I} \subseteq \mathbf{J}(\alpha_0)} \left(\prod_{j \in \mathcal{I}} \sqrt{k^2 - |\alpha - j|^2} \right) v_{\mathcal{I}}(\alpha)\,,$$

with analytic functions $v_{\mathcal{I}}$, holds for all $\alpha \in Q$. To establish the assertion, it is necessary to distinguish between curves of singular points close to Q and those at a larger distance. Hence, we define

$$\mathbf{J}_1 \coloneqq \{j \in \mathbf{J}(\alpha_0) : |k - |\alpha - j|| \leq d_{\max,N^*} \text{ for some } \alpha \in Q\}$$

and $\mathbf{J}_2 \coloneqq \mathbf{J}(\alpha_0) \setminus \mathbf{J}_1$. To abbreviate the notation, we set $\gamma_j(\alpha) \coloneqq \sqrt{k^2 - |\alpha - j|^2}$ and introduce

$$w_{\mathcal{I}_1}(\alpha) \coloneqq \begin{cases} v_\emptyset(\alpha) + \displaystyle\sum_{\emptyset \neq \mathcal{I}_2 \subseteq \mathbf{J}_2} v_{\mathcal{I}_2}(\alpha) \prod_{j \in \mathcal{I}_2} \gamma_j(\alpha)\,, & \mathcal{I}_1 = \emptyset\,, \\ \displaystyle\sum_{\mathcal{I}_2 \subseteq \mathbf{J}_2} v_{\mathcal{I}_1 \cup \mathcal{I}_2}(\alpha) \prod_{j \in \mathcal{I}_2} \gamma_j(\alpha)\,, & \mathcal{I}_1 \subseteq \mathbf{J}_1\,,\ \mathcal{I}_1 \neq \emptyset\,. \end{cases}$$

With this notation, the representation of w becomes

$$w(\alpha) = \sum_{\mathcal{I}_1 \subseteq \mathbf{J}_1} w_{\mathcal{I}_1}(\alpha) \prod_{j \in \mathcal{I}_1} \gamma_j(\alpha)\,. \tag{3.26}$$

The goal is thus to establish the asserted estimate for each term in (3.26). This can be done by induction on the cardinality of $\mathcal{I}_1$. Throughout the arguments, we shall make use of a generic C denoting constants, that depend on k, the maximum norms of derivatives of all $v_\mathcal{I}$ up to second order and on maximum norms of all $w_\mathcal{I}$ (but not their derivatives).

We start with $\mathcal{I}_1 = \emptyset$. In this case, the product on the right-hand side of (3.26) is equal to 1. Hence, we only need to prove the assertion for each summand in the definition of $w_\emptyset$. We proceed again by induction on the cardinality of $\mathcal{I}_2$. For $v_\emptyset$, the estimate follows directly from (3.25). For $\mathcal{I}_2 \neq \emptyset$, let $j \in \mathcal{I}_2$ and assume that the estimate has been proven for the bounded continuous function

$$z \coloneqq v_{\mathcal{I}_2}(\alpha) \prod_{j \neq k \in \mathcal{I}_2} \gamma_k(\alpha)\,.$$

That means,

$$\max_{\alpha \in Q} |P_Q(z)(\alpha) - z(\alpha)| \leq C 2^{-N^*}. \tag{3.27}$$

Hence, it remains to estimate

$$\max_{\alpha \in Q} |P_Q(\gamma_j\, z)(\alpha) - \gamma_j(\alpha)\, z(\alpha)| \leq C 2^{-N^*}.$$

Using the triangle inequality and the induction assumption for z given in (3.27), we obtain

$$\begin{aligned} |P_Q(\gamma_j\, z)(\alpha) - \gamma_j(\alpha)\, z(\alpha)| &\leq |P_Q(\gamma_j\, z)(\alpha) - \gamma_j(\alpha)\, P_Q z(\alpha)| + |\gamma_j(\alpha)\, P_Q z(\alpha) - \gamma_j(\alpha) z(\alpha)| \\ &\leq |P_Q(\gamma_j\, z)(\alpha) - \gamma_j(\alpha)\, P_Q z(\alpha)| + C \|\gamma_j\|_{\infty;Q} 2^{-N^*/2}\,. \end{aligned}$$

To estimate the first term, we again use the triangle inequality as follows

$$\begin{aligned} |P_Q(\gamma_j\, z)(\alpha) - \gamma_j(\alpha)\, P_Q z(\alpha)| \leq\ & |P_Q(\gamma_j\, z)(\alpha) - P_Q(\gamma_j\, P_Q z)(\alpha)| \\ & + |P_Q(\gamma_j\, P_Q z)(\alpha) - \gamma_j(\alpha)\, P_Q z(\alpha)|\,. \end{aligned} \tag{3.28}$$

By considering the induction assumption (3.27) and using the properties of the bilinear interpolation P_Q, we get

$$\max_{\alpha \in Q} |P_Q(\gamma_j\, z)(\alpha) - P_Q(\gamma_j\, P_Q z)(\alpha)| \leq C \max_{\alpha \in Q} |\gamma_j(\alpha)\, z(\alpha) - \gamma_j(\alpha)\, P_Q z(\alpha)| \leq C \|\gamma_j\|_{\infty;Q} 2^{-N^*/2}\,.$$

Before estimating the second term of (3.28), from Lemma A.2 and the definition of $\mathbf{J}_2$, we know that

$$\left|\frac{\partial^2\gamma_j(\alpha)}{\partial\alpha_\nu^2}\right| \le C\,\frac{(k+|\alpha-j|)^{1/2}}{|k-|\alpha-j||^{3/2}} \le \frac{C}{d_{\max,N^*}^{3/2}} \le C\,2^{3N^*/2}\,, \quad \alpha\in Q\,, \quad \nu=1,2\,, \tag{3.29}$$

where $d_{\max,N^*}$ is defined as in (3.21). Using (3.25) and the bilinearity of $P_Q z$, the second term of (3.28) can be estimated by

$$\begin{aligned}
|P_Q(\gamma_j\,P_Q z)(\alpha) - \gamma_j(\alpha)\,P_Q z(\alpha)| &\le Ch_{N^*}^2 \max_{\nu=1,2}\left\|\frac{\partial^2}{\partial\alpha_\nu^2}(\gamma_j\,P_Q z)\right\|_{\infty;Q}\\
&\le Ch_{N^*}^2 \max_{\nu=1,2}\left\|\sum_{\ell=1}^{2}\frac{\partial^\ell\gamma_j}{\partial\alpha_\nu^\ell}\frac{\partial^{2-\ell}P_Q z}{\partial\alpha_\nu^{2-\ell}}\right\|_{\infty;Q}\\
&\le Ch_{N^*}^2\,\|z\|_\infty \max_{\nu=1,2}\left\|\frac{\partial^2\gamma_j}{\partial\alpha_\nu^2}\right\|_{\infty;Q} \le C\,2^{-N^*/2}\,,
\end{aligned}$$

where the second last inequality is due to $\|\partial_{\alpha_\nu}\gamma_j\partial_{\alpha_\nu}P_Q z\|_{\infty;Q} \le C\|\partial_{\alpha_\nu}^2\gamma_j P_Q z\|_{\infty;Q}$. By summing up all terms, we obtain the asserted estimate for $w_\emptyset$.

Next, we establish the estimate for $\mathcal{I}_1\neq\emptyset$. For $j\in\mathcal{I}_1$, we consider the bounded continuous function

$$z \coloneqq w_{\mathcal{I}_1}\prod_{j\neq k\in\mathcal{I}_1}\gamma_k$$

for which the asserted estimate is valid. Similarly as before, we estimate

$$\begin{aligned}
|P_Q(\gamma_j\,z)(\alpha) - \gamma_j(\alpha)\,z(\alpha)| &\le |P_Q(\gamma_j\,z)(\alpha) - \gamma_j(\alpha)\,P_Q z(\alpha)| + C\,\|\gamma_j\|_{\infty;Q}\,2^{-N^*/2}\\
&\le |P_Q(\gamma_j\,z)(\alpha)| + |\gamma_j(\alpha)\,P_Q z(\alpha)| + C\,\|\gamma_j\|_{\infty;Q}\,2^{-N^*/2}\\
&\le C\,\left(1+2^{-N^*/2}\right)\,\|\gamma_j\|_{\infty;\mathcal{R}_{N^*}}\,.
\end{aligned}$$

From the definition of $\mathbf{J}_1$, it follows that

$$\|\gamma_j\|_{\infty;\mathcal{R}_{N^*}} \le C\,|k-|\alpha-j|| \le C\,(d_{\max,N^*}+\operatorname{diam}(Q)) \le C\,\left(2^{-N^*}+\sqrt{2}\,h_{N^*}\right) \le C\,2^{-N^*}\,.$$

By induction, the asserted estimate now follows for all terms in (3.26). □

It is now straightforward to obtain a bound for approximating the integral on the union of all $Q\in\mathcal{M}_{N^*}$. The corresponding quadrature operator will be denoted by

$$I_{N^*}^T w \coloneqq \sum_{Q\in\mathcal{M}_{N^*}}\int_Q P_Q w(\alpha)\,\mathrm{d}\alpha\,.$$

Theorem 3.14. *Let w satisfy Assumption 3.12 and let the parameter N^* in Algorithm 1 be chosen such that $d_{\max,N^*} < r$ and $h_{N^*}\le\sqrt{2}\delta$. Then, the error of the trapezoidal rule over $\mathcal{R}_N$ is*

bounded by

$$\left\| \int_{\mathcal{R}_{N^*}} w(\alpha)\, \mathrm{d}\alpha - I^T_{N^*} w \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq C\, 2^{-3N^*/2}.$$

Proof. Using the triangle inequality and Theorem 3.13, we have

$$\begin{aligned}\left\| \int_{\mathcal{R}_{N^*}} w(\alpha)\, \mathrm{d}\alpha - I^T_{N^*} w \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} &\leq \sum_{Q\in\mathcal{M}_{N^*}} \int_Q \|(w - P_Q w)(\alpha)\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}\, \mathrm{d}\alpha \\ &\leq \mathrm{C}(k)\, (\#\mathcal{M}_{N^*})\, h^2_{N^*}\, 2^{-N^*/2}\,.\end{aligned}$$

By using Remark 3.11, which establishes $\#\mathcal{M}_{N^*} \sim 2^{N^*}$, and since by construction $h_{N^*} \sim 2^{-N^*}$, the assertion follows. □

On all squares $Q \in \mathcal{M}_n$ for $n = 1, \ldots, N^* - 1$, we will use a P-point Gauss-Legendre quadrature rule in each coordinate direction to approximate the inverse FB transform. We denote this rule applied to a function f by $I^G_{P,Q} f$ and set $I^G_{P,\mathcal{R}_n} f \coloneqq \sum_{Q\in\mathcal{R}_n} I^G_{P,Q} f$.

In equation (A.2), we recall the classic error estimate of the Gaussian quadrature formula in the two-dimensional case according to [76, Thm. 9.20]. In what follows, we present the general well-known error estimate for applying such a rule in Theorem 3.15.

Theorem 3.15. *Let $f \in C^{2P}(\overline{\mathcal{R}_n}; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$. Then, there is a constant C such that*

$$\left\| \int_{\mathcal{R}_n} f(\alpha)\, \mathrm{d}\alpha - I^G_{P,\mathcal{R}_n} f \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq C \left(\frac{h_0}{2}\right)^{2P} \frac{2^{-(2P+1)n}}{(2P+1)!} \max_{\alpha\in\mathcal{R}_n} \left(\sum_{\nu=1}^{2} \left\| \frac{\partial^{2P} f(\alpha)}{\partial \alpha_\nu^{2P}} \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \right).$$

Proof. From equation (A.2) with $P = n + 1$ and our setting of functions mapping to a Sobolev space, we can estimate the error of the integration over each square $Q \in \mathcal{M}_n$ (for $n = 0, \ldots, N^* - 1$) as follows

$$\left\| \int_{Q} f(\alpha)\, \mathrm{d}\alpha - I^G_{P,Q} f \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq \frac{4}{(2P+1)!} \left(\frac{h_n}{2}\right)^{2P+2} \max_{\alpha\in Q} \left(\sum_{\nu=1}^{2} \left\| \frac{\partial^{2P} f(\alpha)}{\partial \alpha_\nu^{2P}} \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \right).$$

Using the estimates given in Remark 3.11, we obtain the asserted error bound. □

Based on Theorem 3.15, the error of the Gauss-Legendre rule for computing the integral of w over $\mathcal{R}_n$ depends on the $2P$-th partial derivatives of w with respect to either α_1 or α_2. Recalling the representation (3.13), it suffices to estimate the $2P$-th partial derivatives of $\prod_{j\in\mathcal{I}} \sqrt{k^2 - |\alpha - j|^2}\, v_{\mathcal{I}}(\alpha)$ with respect to each coordinate. We do so in the next lemma using some standard estimates for square root functions and their derivatives presented in the appendix.

Lemma 3.16. *For any fixed $\ell \in \mathbb{N}$, there is a constant C such that*

$$\max_{\alpha\in\mathcal{R}_n} \left| \frac{\partial^\ell \sqrt{k^2 - |\alpha - j|^2}}{\partial \alpha_\nu^\ell} \right| \leq \frac{C\, \ell!\, (d_{\max,n})^{1/2}}{(d_{\min,n})^\ell}, \quad n = 1, \ldots, N^* - 1, \quad \nu = 1, 2\,, \tag{3.30}$$

where $d_{\min,n}$ and $d_{\max,n}$ are defined by (3.20) and (3.21), respectively.

Proof. According to Lemma A.2, for all $\alpha \in \mathcal{R}_n, (n = 1, \dots, N^* - 1)$, there is a constant $\widetilde{C}$ such that

$$\left| \frac{\partial^\ell \sqrt{k^2 - |\alpha - j|^2}}{\partial \alpha_\nu^\ell} \right| \leq \frac{\widetilde{C}\, \ell!\, |k + |\alpha - j||^{1/2}}{|k - |\alpha - j||^{\ell - 1/2}}\,.$$

Hence, using (3.20) and (3.21), i.e., $d_{\min,n} \leq |k - |\alpha - j|| \leq d_{\max,n}$, leads to

$$\max_{\alpha \in \mathcal{R}_n} \left| \frac{\partial^\ell \sqrt{k^2 - |\alpha - j|^2}}{\partial \alpha_\nu^\ell} \right| \leq \frac{C\, \ell!\, (d_{\max,n})^{1/2}}{(d_{\min,n})^\ell}\,.$$

This completes the proof. □

Theorem 3.17. *Let $\mathcal{I} \subset \mathbf{J}$ and denote by $\gamma_{\mathcal{I}}(\alpha) \coloneqq \prod_{j \in \mathcal{I}} \sqrt{k^2 - |\alpha - j|^2}\, v_{\mathcal{I}}(\alpha)$ one of the terms in (3.13). Let $m \coloneqq \#\mathcal{I}$. Then, for every $\ell \in \mathbb{N}_0$, there exists $C_\ell > 0$ such that*

$$\max_{\alpha \in \mathcal{R}_n} \left\| \frac{\partial^\ell \gamma_{\mathcal{I}}}{\partial \alpha_\nu^\ell} \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq \frac{\mathrm{C}_\ell\, (d_{\max,n})^{m/2}}{(d_{\min,n})^\ell}, \qquad \nu = 1, 2. \tag{3.31}$$

Proof. From the generalized Leibniz formula, we obtain

$$\frac{\partial^\ell \gamma_{\mathcal{I}}(\alpha)}{\partial \alpha_\nu^\ell} = \sum_{\mathrm{K}_0 + \dots + \mathrm{K}_m = \ell} \frac{\ell!}{\mathrm{K}_0! \cdots \mathrm{K}_m!} \frac{\partial^{\mathrm{K}_0} v_{\mathcal{I}}(\alpha)}{\partial \alpha_\nu^{\mathrm{K}_0}} \prod_{\mu=1}^{m} \frac{\partial^{\mathrm{K}_\mu}}{\partial \alpha_\nu^{\mathrm{K}_\mu}} \sqrt{k^2 - |\alpha - j_\mu|^2}\,.$$

Using (3.14) and Lemma 3.16, we have for $\alpha \in \mathcal{R}_n$

$$\left\| \frac{\partial^\ell \gamma_{\mathcal{I}}(\alpha)}{\partial \alpha_\nu^\ell} \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq C \sum_{\mathrm{K}_0 + \dots + \mathrm{K}_m = \ell} \frac{\ell!}{\mathrm{K}_0! \cdots \mathrm{K}_m!} \frac{C_{\mathrm{K}_0}}{(d_{\min,n})^{\mathrm{K}_0}} \prod_{\mu=1}^{m} \frac{\mathrm{K}_\mu!\, (d_{\max,n})^{1/2}}{(d_{\min,n})^{\mathrm{K}_\mu}}\,.$$

Combining all constants gives the assertion. □

Theorem 3.18. *Let w satisfy Assumption 3.12. Then, for every $P \in \mathbb{N}$, there exists a constant C_P such that*

$$\left\| \sum_{n=1}^{N^*-1} \int_{\mathcal{R}_n} w(\alpha)\, \mathrm{d}\alpha - \sum_{n=1}^{N^*-1} I^G_{P,\mathcal{R}_n} w \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq C_P\, h_0^{2P}\,.$$

Proof. Combining Theorems 3.15 and 3.17, we obtain the estimate

$$\left\| \int_{\mathcal{R}_n} w(\alpha)\, \mathrm{d}\alpha - I^G_{P,\mathcal{R}_n} w \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq C_P \left(\frac{h_0}{2} \right)^{2P} \frac{2^{-(2P+1)n}}{(d_{\min,n})^{2P}}\,,$$

with some constant C_P independent of h_0 and n. From (3.20), we have $d_{\min,n} \geq C\, 2^{-n}$. Hence,

$$\left\| \int_{\mathcal{R}_n} w(\alpha)\, \mathrm{d}\alpha - I^G_{P,\mathcal{R}_n} w \right\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \leq C_P\, \frac{h_0^{2P}}{2^n}\,.$$

Summing over $n = 1, \dots, N^* - 1$ completes the proof. □

Now, we are going to provide the analysis of the total error in numerical inversion of the FB transform. It is straightforward to combine the quadrature rules of the previous two subsections to obtain a super-algebraically convergent approximation to the inverse FB transform of the transformed field.

Corollary 3.19. *Let w satisfy Assumption 3.12 and fix $P \in \mathbb{N}$. Then, there is $C_P > 0$ such that for every h_0 and N^* with $d_{\max,N^*} < r$, $h_{N^*} \leq \sqrt{2}\delta$, there holds*

$$\left\| \int_\Lambda w(\alpha)\,\mathrm{d}\alpha - I^T_{N^*} w - \sum_{n=1}^{N^*-1} I^G_{P,\mathcal{R}_n} w \right\|_{H^1_{\mathrm{per}}(\Omega^{2\pi}_H)} \leq C_P \left(2^{-3N^*/2} + h_0^{2P} \right) ,$$

where $I^T_{N^*}$ and $I^G_{P,\mathcal{R}_n}$ are defined on Pages 47 and 48.

Example 3.20. As examples for the performance achievable with our quadrature rule, we consider functions w that are products of the square root functions occurring in the representation (3.11). In this special case, all $w_{\mathcal{I}}$ are either constant 0 or 1 and thus analytic on Λ. From (3.16) and the estimates in the proof of Theorem 3.17, we expect the constant C_P to be independent of P in this case.

We apply the quadrature rule to the approximation of two integrals,

$$\begin{aligned} \mathbf{I}_1 &= \int_\Lambda \sqrt{k^2 - |\alpha - j|^2}\ \mathrm{d}\alpha && k = 0.4\,,\ j = (0,0)\,, \\ \mathbf{I}_2 &= \int_\Lambda \sqrt{k^2 - |\alpha - j|^2}\,\sqrt{k^2 - |\alpha - l|^2}\ \mathrm{d}\alpha && k = 1.4\,,\ j = (-1,0)\,,\ l = (-1,1)\,. \end{aligned}$$

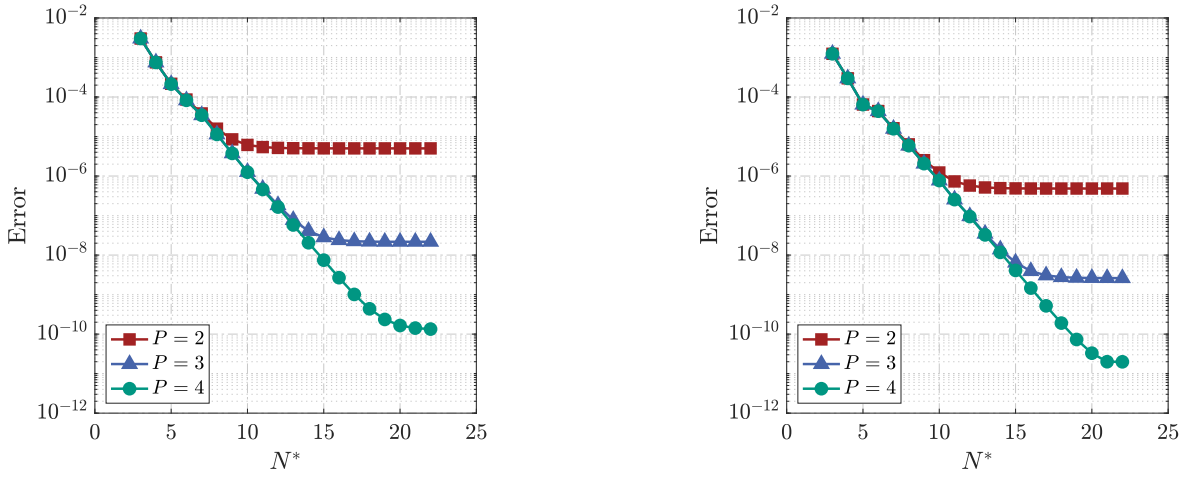
For the first integral, the set Σ is a single circle entirely contained in the set Λ (as depicted in the left image of Figure 3.3). Hence, the exact value of the integral $\mathbf{I}_1$ can be obtained analytically. We have used Maple 2022 to carry out this task and then computed approximations using our quadrature rule for various values of N^* and P.

In the second integral, the integrand is singular along two circular arcs contained in the set Λ. The exact value of this integral is not available. Instead, we have computed a reference value for $N^* = 23$ and $P = 5$ and compared our results against it. The results are presented in Figure 3.7. The theoretically predicted convergence rate from Corollary 3.19 is very well reflected, with exponential convergence with respect to N^*, until the error of the Gauss quadrature rule dominates. The results also nicely illustrate our expectation that C_P is independent of P for these examples.

3.4. Full Discretization of Scattering Problems

To solve (3.3) numerically in $\Lambda \times \Omega^{2\pi}_H$, we use a tetrahedral mesh in $\Omega^{2\pi}_H$ with maximum diameter τ and a special structure on the top surface $\Gamma^{2\pi}_H$. We consider a tensor product of $L_1 + 1$ equidistant nodes in x_1 direction and $L_2 + 1$ in x_2 direction, that is

$$x_\ell := \left(-\pi + \frac{2\pi}{L_1}\ell_1, -\pi + \frac{2\pi}{L_2}\ell_2, H \right)^\top \tag{3.32}$$



(a) $\mathbf{I}_1$ (b) $\mathbf{I}_2$

FIGURE 3.7. Numerical error of the proposed quadrature rule for various N^* and P applied to $\mathbf{I}_j$, $j = 1, 2$.

for $\ell = (\ell_1, \ell_2) \in \{0, \dots, L_1\} \times \{0, \dots, L_2\}$. Moreover, we denote by N_Δ the total number of nodes in the tetrahedral mesh. In the domain $\Lambda = [-1/2, 1/2]^2$, we generate an adapted square mesh using Algorithm 1.

For each quadrature node α in the adapted mesh generated by Algorithm 1, we approximate the solution $w(\alpha)$ of (3.7) by `P1`-conforming piecewise linear finite elements basis functions $\{\phi\}_{n=1}^{N_\Delta}$. Substituting the approximation of the solution into the variational problem (3.7) yields

$$\sum_{n=1}^{N_\Delta} w_n(\alpha_j) a_{\alpha_j}(\phi_n, \phi_m) = \left\langle (\partial_{x_d} - \mathcal{T}_{\alpha_j}^+) \mathcal{J} u^i(\alpha_j), \overline{\phi_m} \right\rangle_{\Gamma_H^{2\pi}} \quad \text{for all } m \in \{1, \dots, N_\Delta\},$$

where the sesquilinear form a_{α_j} is defined as in (3.4). This leads to the following linear system for each α_j

$$\left(\mathbf{D} - 2\mathrm{i}\mathbf{A}_j - (k^2 - |\alpha_j|^2)\mathbf{M} - \mathbf{DtN}_j \right) \mathbf{W}_j = \mathbf{F}_j, \tag{3.33}$$

where $\mathbf{D}$ and $\mathbf{M}$ are the standard diffusion and mass matrices, $(\mathbf{W}_j)_m \coloneqq w(\alpha_j; x_m)$ and

$$\begin{aligned} (\mathbf{A}_j)_{m,n} &\coloneqq \left\langle \alpha_j \cdot \nabla_{\widetilde{x}} \phi_n, \overline{\phi_m} \right\rangle_{\Omega_H^{2\pi}}, \\ (\mathbf{DtN}_j)_{m,n} &\coloneqq \left\langle \mathcal{T}_{\alpha_j}^+ \phi_n, \overline{\phi_m} \right\rangle_{\Gamma_H^{2\pi}}, \\ (\mathbf{F}_j)_m &\coloneqq \left\langle (\partial_{x_d} - \mathcal{T}_{\alpha_j}^+) \mathcal{J} u^i(\alpha_j), \overline{\phi_m} \right\rangle_{\Gamma_H^{2\pi}}, \end{aligned}$$

for $m, n \in \{1, \dots, N_\Delta\}$ and $j \in \{1, \dots, N_\alpha\}$.

To discretize the DtN map, we proceed similarly to the approach in [65, Sec. 3]. By considering the special structure of the generated mesh on $\Gamma_H^{2\pi}$ given in (3.32), we approximate $(\mathcal{T}_{\alpha_j}^+ \phi_n)(x)$

again by the piecewise linear functions as follows

$$(\mathcal{T}^+_{\alpha_j}\phi_n)(x) = \sum_{s=1}^{N_\Delta} (\mathcal{T}^+_{\alpha_j}\phi_n)(x_s)\phi_s(x).$$

In this case, we have

$$\left\langle \mathcal{T}^+_{\alpha_j}\phi_n, \overline{\phi_m} \right\rangle_{\Gamma_H^{2\pi}} = \sum_{s=1}^{N_\Delta} (\mathcal{T}^+_{\alpha_j}\phi_n)(x_s) \left\langle \phi_s, \overline{\phi_m} \right\rangle_{\Gamma_H^{2\pi}}.$$

To compute the matrix corresponding to $(\mathcal{T}^+_{\alpha_j}\phi_n)(x_s)$, we approximate the finite element basis functions ϕ_n by the trigonometric functions ψ_n as follows (see [8, Sec. 3.8])

$$\psi_n(x) \coloneqq \frac{1}{L_1 L_2} \sum_{\ell_1=0}^{L_1} \sum_{\ell_2=0}^{L_2} \mathrm{e}^{\mathrm{i}(\ell_1,\ell_2)\cdot(\widetilde{x}-\widetilde{x}_n)} \quad \text{for } x = (\widetilde{x}, H) \in \Gamma_H^{2\pi}, \tag{3.34}$$

where $\widetilde{x}_n$ are the nodes on the top surface defined in (3.32). Note that we focus only on the nodes on the top surface, since the elements corresponding to the other nodes are zero. From (3.34), we conclude that for $\ell = (\ell_1, \ell_2)$, the ℓ-th Fourier coefficient of ψ_n is $\mathrm{e}^{-\mathrm{i}\ell\cdot\widetilde{x}_n}$. This yields

$$\begin{aligned}(\mathcal{T}^+_{\alpha_j}\psi_n)(x_s) &= \frac{1}{L_1 L_2} \sum_{\ell_1=0}^{L_1} \sum_{\ell_2=0}^{L_2} \sqrt{k^2 - |\alpha_j - \ell|^2}\, \mathrm{e}^{\mathrm{i}\ell\cdot\widetilde{x}_s} \widehat{\psi}_n(\ell) \\ &= \frac{1}{L_1 L_2} \sum_{\ell_1=0}^{L_1} \sum_{\ell_2=0}^{L_2} \sqrt{k^2 - |\alpha_j - \ell|^2}\, \mathrm{e}^{\mathrm{i}\ell\cdot(\widetilde{x}_s-\widetilde{x}_n)}.\end{aligned}$$

Note that the expression on the last line gives us a matrix, which can be decomposed as the product of a diagonal matrix, the two-dimensional discrete Fourier transform and its inverse. Numerically, it is important how to implement this boundary condition because it affects the computational time. We solve the system (3.33) using an iterative method; hence the dense matrix $\mathbf{DtN}_j$ does not need to be assembled, instead it is required to perform the matrix-vector multiplication $\mathbf{DtN}_j\mathbf{W}_j$ for a generic vector $\mathbf{W}_j$. In Algorithm 2, we describe how to perform this matrix-vector multiplication in an efficient way by using the fast Fourier transform.

3.4.1. Error Analysis for Fully-Discrete Scheme

To conclude our analysis, we combine the result of Corollary 3.19 with error bounds for the Galerkin approximation for the solution of the variational problem (3.7).

Theorem 3.21. *Let $(\mathcal{J}u^i)(\alpha) \in H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})$ and $w(\alpha)$ denote the exact solution of the variational formulation of (3.7) and $w_\tau(\alpha)$ its numerical approximation by the finite element method with mesh size τ. For sufficiently small τ,*

$$\|w(\alpha) - w_\tau(\alpha)\|_{H^s_{\mathrm{per}}(\Omega_H^{2\pi})} \le C\,\tau^{2-s} \left\|(\mathcal{J}u^i)(\alpha)\right\|_{H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})} \quad \text{for } s = 0, 1,$$

where the constant C is independent of the Floquet parameter α.

Algorithm 2: full discretization of the scattering problem, given in (3.33)

Input: wave number k and the quadrature rule $\{(\alpha_j, \varrho_j)\}_{j=1}^{N_\alpha}$ from Algorithm 1

Generate a tetrahedral mesh in the cell $\Omega_H^{2\pi}$;
Construct the sparse stiffness matrices $\mathbf{M}$ and $\mathbf{D}$;
for $j = 1, \dots, N_\alpha$ **do in parallel**
  Construct the sparse stiffness matrix $\mathbf{A}_j$;
  Compute the right-hand side $\mathbf{F}_j$ from the given incident field $(\mathcal{J}u^i)(\alpha_j)$;
  Compute the diagonal matrix $\boldsymbol{\beta}_{s,s} \coloneqq \sqrt{k^2 - |\alpha_j - s|^2}$;
  Define the function *LHS*
    Input: the vector $\mathbf{F}_j$
    %Compute $\mathbf{DtN}_j$ by the fast Fourier transform fft and its inverse ifft
    $\mathbf{DtN}_j \leftarrow \mathtt{ifft}\,(\boldsymbol{\beta}\,\mathtt{fft}(\mathbf{W}_j))$;
    %Perform the matrix-vector product in the standard way for $\mathbf{M}, \mathbf{D}, \mathbf{A}_j$
    return $\mathbf{D}\,\mathbf{W}_j - 2\mathrm{i}\mathbf{A}_j\mathbf{W}_j - (k^2 - |\alpha_j|^2)\mathbf{M}\,\mathbf{W}_j - \boldsymbol{DtN}_j$
  Solve (3.33) by GMRES with inputs *LHS*, initial guess zero and tolerance 10^{-5};
  $\mathbf{W}_j \leftarrow \mathbf{W}_j\, \mathrm{e}^{\mathrm{i}\alpha_j \cdot \widetilde{x}}$;
Use the numerical inversion of FB transform to compute the total field u;
return *Numerical total field* u

Proof. The proof is completely analogous to [85, Thm. 14]. □

Combining both error bounds given in Corollary 3.19 and Theorem 3.21, yields the complete estimate for the proposed numerical method. To concisely formulate this result, we introduce operators

$$\Upsilon_\ell \psi(\alpha, x) \coloneqq \psi(\alpha, x)\, \mathrm{e}^{\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi\ell)} \qquad \ell \in \mathbb{Z}\,,$$

$$\mathcal{J}^{-1}_{P,N^*,h_0} \psi(\widetilde{x} + 2\pi\ell, x_3) \coloneqq \left(I^T_{N^*} + \sum_{n=1}^{N^*-1} I^G_{P,\mathcal{R}_n} \right) \Upsilon_\ell \psi(x)\,,$$

where $I^T_{N^*}$ and $I^G_{P,\mathcal{R}_n}$ are defined on Pages 47 and 48.

Theorem 3.22. *Let $u^i \in H^1_r(\Omega_H^{\mathrm{per}})$ for some $|r| < 1$ and additionally $(\mathcal{J}u^i)(\alpha) \in H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})$. Let u denote the total field, i.e., the solution to (3.2), and for any $\alpha \in \Lambda$ by $w_\tau(\alpha)$ the finite element approximation to the solution of (3.7) for sufficiently small mesh size τ. Let h_0 and N^* satisfy $d_{\max,N^*} < r$, $h_{N^*} \le \sqrt{2}\delta$ and fix $P \in \mathbb{N}$. Then, there holds*

$$\left\| u - \mathcal{J}^{-1}_{P,N^*,h_0} w_\tau \right\|_{H^s(\Omega_H^{2\pi})} \le C \left(\tau^{2-s} + 2^{-3N^*/2} + {h_0}^{2P} \right), \qquad s = 0, 1\,,$$

where the constant C depends on the order of Gauss-Legendre rule P and the incident field u^i.

Proof. For any $\alpha \in \Lambda$, denote by $w(\alpha)$ the exact solution to (3.7). By using the inverse FB

transform and then the triangle inequality, we have

$$\begin{aligned}\left\|u-\mathcal{J}_{P,N^*,h_0}^{-1}w_\tau\right\|_{H^s(\Omega_H^{2\pi})} &= \left\|\mathcal{J}^{-1}w-\mathcal{J}_{P,N^*,h_0}^{-1}w_\tau\right\|_{H^s(\Omega_H^{2\pi})}\\ &\le \left\|\left(\mathcal{J}^{-1}-\mathcal{J}_{P,N^*,h_0}^{-1}\right)w\right\|_{H^s(\Omega_H^{2\pi})}\\ &\quad+\left\|\mathcal{J}_{P,N^*,h_0}^{-1}(w-w_\tau)\right\|_{H^s(\Omega_H^{2\pi})}.\end{aligned}\tag{3.35}$$

Note that application of Υ_ℓ is just a multiplication with an analytic function, hence $\Upsilon_\ell w$ satisfies Assumption 3.12. For the first term of (3.35), Corollary 3.19 gives

$$\left\|\left(\mathcal{J}^{-1}-\mathcal{J}_{P,N^*,h_0}^{-1}\right)w\right\|_{H^s(\Omega_H^{2\pi})}\le C_P\left(2^{-3N^*/2}+h_0^{2P}\right).$$

Denote by (α_j,ϱ_j), for $j=1,\dots,N_\alpha$, all the quadrature points and corresponding weights appearing in the rules $I_{N^*}^T$ and $I_{P,\mathcal{R}_n}^G$. It should be noted that all the weights are positive. Accordingly, we may write using Theorem 3.21,

$$\begin{aligned}\left\|\mathcal{J}_{P,N^*,h_0}^{-1}(w-w_\tau)\right\|_{H^s(\Omega_H^{2\pi})} &\le \sum_{j=1}^{N_\alpha}\varrho_j\|w(\alpha_j)-w_\tau(\alpha_j)\|_{H^s_{\mathrm{per}}(\Omega_H^{2\pi})}\\ &\le C\tau^{2-s}\sum_{j=1}^{N_\alpha}\varrho_j\left\|(\mathcal{J}u^i)(\alpha_j)\right\|_{H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})}.\end{aligned}$$

As $\mathcal{J}u^i\in\mathcal{X}(H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi}))$, we may use the same approach as in the proof of Theorem 3.9 to derive an expression analogous to (3.13) for $\mathcal{J}u^i$ and conclude that $\sup_{\alpha\in\Lambda}\|(\mathcal{J}u^i)(\alpha)\|_{H^{1/2}_{\mathrm{per}}(\Gamma_H^{2\pi})}<\infty$. Then, using the fact that $\sum_{j=1}^{N_\alpha}\varrho_j=|\Lambda|=1$, the proof is completed. □

3.5. Numerical Results

In this section, we present numerical examples to illustrate the performance of the proposed method for solving the three-dimensional scattering problems. To have access to an exact solution, we consider the case of a radiation problem: we assume that $\Gamma^{\mathrm{per}}\subseteq\mathbb{R}^3_+$, where $\mathbb{R}^3_+:=\{x\in\mathbb{R}^3_+ : x_3>0\}$ is the upper half-space and that u^i is the Dirichlet Green's function for this upper half-space for some source point y located between Γ^{per} and $x_3=0$,

$$u^i(x)=\mathrm{G}(x,y)=\frac{1}{4\pi}\left(\frac{\exp\left(\mathrm{i}k|x-y|\right)}{|x-y|}-\frac{\exp\left(\mathrm{i}k|x-y'|\right)}{|x-y'|}\right),\quad x\in\mathbb{R}^3_+,\quad x\neq y\,.$$

As indicated above, we assume that the point source $y=(y_1,y_2,y_3)^\top$ satisfies $0<y_3<\zeta^{\mathrm{per}}(y_1,y_2)$, and $y'=(y_1,y_2,-y_3)^\top$ denotes the reflected point source. The reason for using Green's function instead of the standard fundamental solution is its faster decay rate in vertically bounded strips (see Lemma C.2). Moreover, $u^i\in H^1_r(\Omega_H)$ with $r<1$ for the point source below Γ^{per}(see [84, Sec. 7]). As we are considering a radiation problem, the "scattered field" u^s satisfies $u^s=-u^i$ in Ω^{per}. Hence, we are able to compute explicitly the numerical approximation error in the scattered

field obtained by (3.3) for the vanishing total field in the bounded cell $\Omega_H^{2\pi}$.

We assume that the surface Γ^{per} is given by the bi-periodic function

$$\zeta^{\mathrm{per}}(\widetilde{x}) = 0.6 + 0.3\,\sin(x_1)\cos(2x_2) + 0.2\,\sin(2x_1)\sin(3x_2), \quad \text{for } \widetilde{x} = (x_1, x_2) \in \mathbb{R}^2.$$

Moreover, we fix $H = 2$ and consider the source point $y = (0, 0, 0.1)^\top$.

To solve (3.3) in $\Lambda \times \Omega_2^{2\pi}$, we first generate an adapted square mesh in Λ by using Algorithm 1 and tetrahedral meshes in $\Omega_2^{2\pi}$ with $(\mathrm{M}+1)^2 \times (\mathrm{M}/2+1)$ nodes for $\mathrm{M} \in \{16, 32, 64, 128\}$ so that the maximum diameter τ for these four generated meshes is 0.78, 0.41, 0.21 and 0.1, respectively. Note that these values for τ are smaller than the essential limit of one-tenth of the wavelength for each value of k considered below. For each $\alpha \in \Lambda$, we approximate the solution $w(\alpha)$ of (3.7) by `P1`−conforming piecewise linear finite elements (as explained in Section 3.4).

According to Lemma C.3, the FB transform of Green's function for each $\alpha \in \Lambda$ is computed as

$$\mathcal{J}u^i = \mathrm{e}^{-\mathrm{i}\alpha\cdot\widetilde{y}} \sum_{j\in\mathbb{Z}^2} \mathrm{e}^{-\mathrm{i}j\cdot(\widetilde{x}-\widetilde{y})} \begin{cases} \mathrm{e}^{\mathrm{i}\sqrt{k^2-|\alpha-j|^2}\,x_3}\,\mathrm{sinc}\left(\sqrt{k^2-|\alpha-j|^2}\,y_3\right) y_3 & y_3 < x_3\,, \\ \mathrm{e}^{\mathrm{i}\sqrt{k^2-|\alpha-j|^2}\,y_3}\,\mathrm{sinc}\left(\sqrt{k^2-|\alpha-j|^2}\,x_3\right) x_3 & \text{otherwise}\,. \end{cases} \tag{3.36}$$

The formula for $\mathcal{J}u^i$ given above in particular shows that the assumptions of Theorem 3.22 are satisfied. The right-hand side can be evaluated by truncating the infinite series to $|j_1|, |j_2| \le 40$. Eventually, we solve the sparse linear system (3.33) using the iterative solver described in Algorithm 2.

Below, we will demonstrate the dependence of the numerical error on the discretization parameters: the mesh size τ in the spatial space, the number of refinement N^* in α-space and the order P of the Gauss-Legendre rule. Let the relative errors and the computational orders be

$$\mathrm{E}_\tau = \frac{\|u^s - u_\tau^s\|_{L^2(\Omega_H^{2\pi})}}{\|u^s\|_{L^2(\Omega_H^{2\pi})}}, \qquad \mathrm{C}_{\mathrm{order}} = \frac{\log(\mathrm{E}_{\tau_1}/\mathrm{E}_{\tau_2})}{\log(\tau_1/\tau_2)}\,,$$

where u^s is the exact scattered field and u_τ^s denotes its finite element approximation with the mesh size τ. In Table 3.1, we list the relative errors and the computational orders for different values of τ and wave number k. This table indicates that the numerical results are consistent with the analytic results of Theorem 3.22 for each k, since errors converge as the mesh size τ decreases even with a low number of N^* and P.

Note that for large values of the wave number k, the structure of the singular curves becomes more complicated. For example for $k = 3$ there are 20 curves of singular points in the domain Λ. Despite the complicated structure of the singular curves in α-space, the accurate results can still be obtained by using small values of N^* and P, only refining the spatial mesh τ, as reported in Table 3.1.

In Tables 3.2 and 3.3, we report the relative errors with respect to N^* and P for different values of τ. Since the error of the finite element method is dominated in the computational order, we cannot see the exponential convergence of the proposed numerical integration method with respect to N^* and P.

TABLE 3.1. Relative error and computational order with respect to τ by $N^* = 3$, $P = 2$.

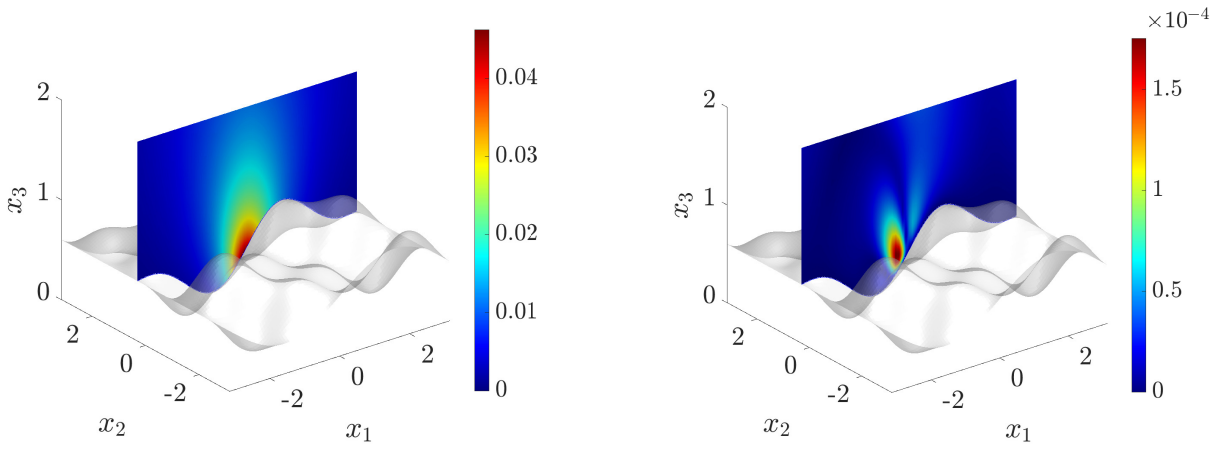
	$k = 0.4$		$k = 1.4$		$k = 3$	
τ	Error	C_{order}	Error	C_{order}	Error	C_{order}
0.78	3.3438×10^{-2}	$--$	3.7156×10^{-2}	$--$	1.9390×10^{-1}	$--$
0.41	1.0870×10^{-2}	1.75	1.0788×10^{-2}	1.92	5.9628×10^{-2}	1.83
0.21	3.0854×10^{-3}	1.88	2.8671×10^{-3}	1.98	1.5824×10^{-2}	1.98
0.10	8.1722×10^{-4}	1.79	7.3826×10^{-4}	1.83	4.0295×10^{-3}	1.84

TABLE 3.2. Relative error with respect to P and N^* for wave number $k = 0.4$.

	$\tau = 0.78$		$\tau = 0.21$	
P	$N^* = 2$	$N^* = 3$	$N^* = 2$	$N^* = 3$
2	3.3658×10^{-2}	3.3438×10^{-2}	3.4548×10^{-3}	3.0854×10^{-3}
3	3.3658×10^{-2}	3.3438×10^{-2}	3.4548×10^{-3}	3.0854×10^{-3}
4	3.3658×10^{-2}	3.3438×10^{-2}	3.4548×10^{-3}	3.0854×10^{-3}

TABLE 3.3. Relative error with respect to N^* and τ for $k = 1$, $P = 2$.

N^*	$\tau = 0.78$	$\tau = 0.41$	$\tau = 0.21$
2	3.4106×10^{-2}	1.1145×10^{-2}	3.5580×10^{-3}
3	3.4054×10^{-2}	1.1137×10^{-2}	3.2018×10^{-3}
4	3.3979×10^{-2}	1.1078×10^{-2}	3.1413×10^{-3}
5	3.3976×10^{-2}	1.1078×10^{-2}	3.1428×10^{-3}

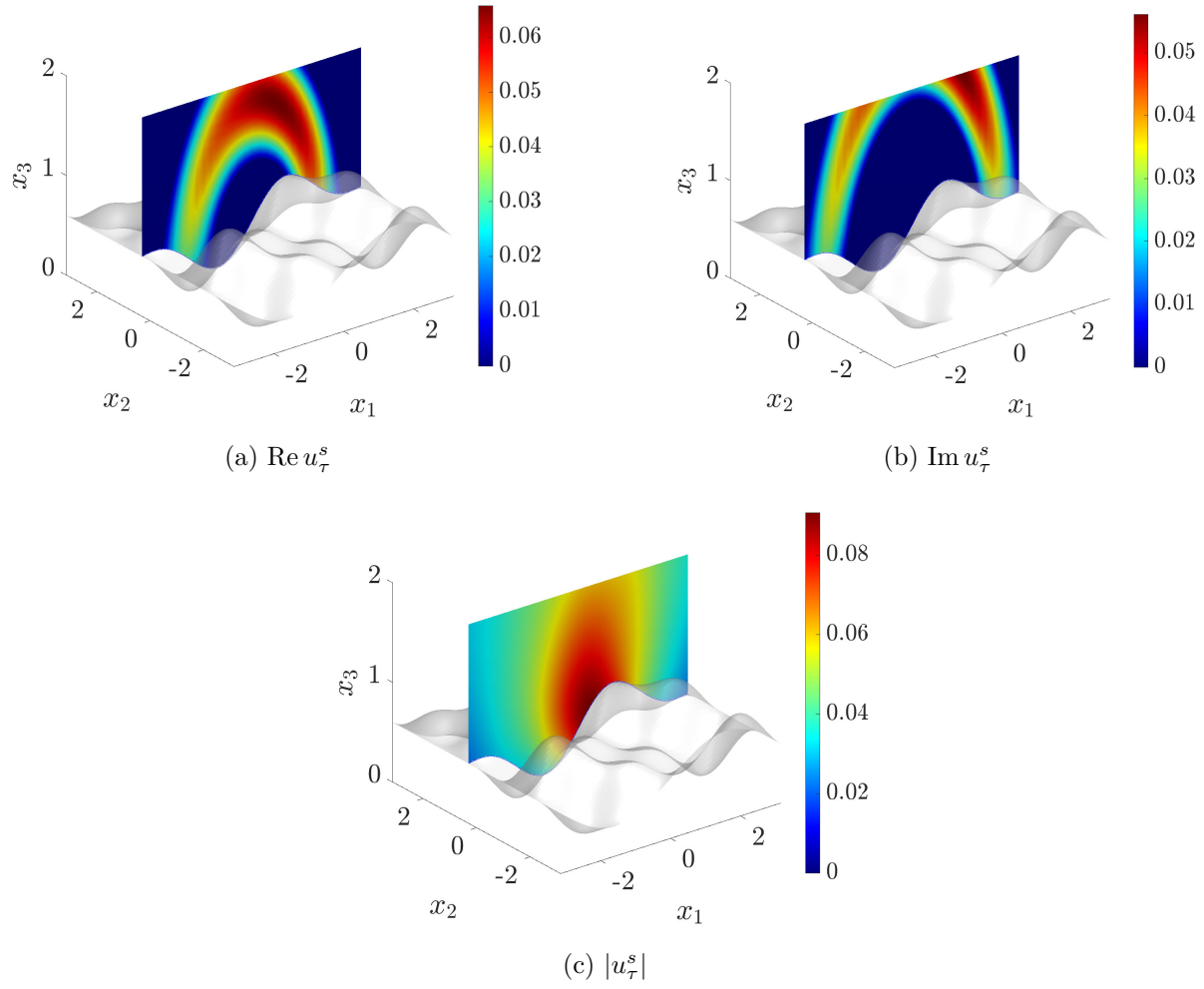


(a) $|u^s_\tau|$ (b) $|u^s - u^s_\tau|$

FIGURE 3.8. Numerical scattered field and its absolute error for $k = 1$ with the point source $y = (0, 0, 0.1)^\top$ located below Γ^{per}.

In Figure 3.8, we show the numerical scattered field u_τ^s and its numerical error in the L^2-norm for $k = 1$ with parameters $\tau = 0.21$, $N^* = 3$ and $P = 2$. As shown in Figure 3.8(b), the maximum value of the numerical error is approximately 10^{-4}, which indicates the accuracy of the proposed method.

Having established the efficiency of the proposed method for a point source below the surface, we now consider the case with the source above the surface. In this case, the exact solution is not available. Therefore, we present only the real, imaginary, and absolute values of the scattered field, to see how the scattered field propagates in $\Omega_2^{2\pi}$. In Figure 3.9, we illustrate the behaviour of the scattered field generated by the point source $y = (-1, 0, 1)^\top$ located inside $\Omega_2^{2\pi}$.



(a) $\operatorname{Re} u_\tau^s$

(b) $\operatorname{Im} u_\tau^s$

(c) $|u_\tau^s|$

FIGURE 3.9. Numerical scattered field for $k = 3$ with the point source located at $y = (-1, 0, 1)^\top$ above Γ^{per}.

In conclusion, our method provides a way to very accurately approximate the inverse FB transform for solutions to a non-periodic scattering problem. Even for very small values of P, the error from this approximation is already dominated by the error from the finite element method. Nevertheless, for larger wave numbers, the structure of the singular curves quickly

becomes quite complicated, making it necessary to use a large number of quadrature points. Thus, the accurate solution of non-periodic scattering problems in periodic domains remains a computational challenge.

CHAPTER 4

Scattering in Unbounded Locally Perturbed Periodic Structures

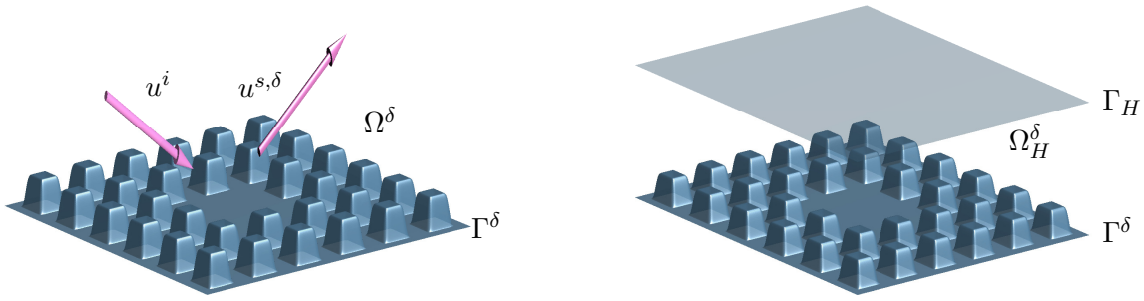
In this chapter, we focus on acoustic scattering from locally perturbed surfaces. We assume that the periodicity of the scatterer Γ^{per} is disrupted by a compactly supported perturbation δ, located in the region $[-\pi,\pi]^{d-1}$ for $d=2,3$. Next, we define a locally perturbed function by $\zeta^{\delta} \coloneqq \zeta^{\text{per}} + \delta$, which generates the locally perturbed scatterer Γ^{δ}. The unbounded domain above Γ^{δ} is denoted by Ω^{δ} (see Figure 4.1(a) for a visualization of these domains).

We begin by employing the DtN map as a truncation method, as introduced in the previous chapter, but now in the context of a locally perturbed case. Next, we apply a diffeomorphism to transform this locally perturbed domain into a periodic one. By subsequently applying the FB transform to the resulting formulation of the scattering problem, we derive a coupled family of periodic problems — indexed by the Floquet parameter — defined in a bounded cell. Furthermore, we approximate the solution of these problems using the PML method and analyze the regularity of the transformed field with respect to the Floquet parameter. The regularity result shows that due to the analyticity of the PML approximation of the DtN map, the resulting operator and the scattered field depend analytically on the Floquet parameter. This allows us to evaluate the inverse FB transform by much fewer values of the Floquet parameter, compared to the method presented in the previous chapter. Furthermore, we prove that the PML approximation of the scattered field converges exponentially to the exact scattered field with respect to the PML parameter in every compact set in two dimensions.

Finally, we propose a fast iterative method to compute the scattered field numerically, which allows us to exploit parallelization despite the problem's coupling. The efficiency of the proposed method is demonstrated through several numerical examples.

4.1. Formulation in a Bounded Cell

We begin by truncating the unbounded domain above the scatterer Γ^{δ} in the vertical direction. To this end, we first introduce some notations. For $H > \max\{\|\zeta^{\text{per}}\|_{\infty}, \|\zeta^{\delta}\|_{\infty}\}$, we define the flat surface $\Gamma_H \coloneqq \mathbb{R}^{d-1} \times \{H\}$. We denote the unbounded domain between the locally perturbed surface Γ^{δ} and the flat surface Γ_H by Ω^{δ}_H (see Figure 4.1(b)).



(a) Locally perturbed domain Ω^δ (b) Truncated domain Ω_H^δ

FIGURE 4.1. Sketch of the unbounded locally perturbed domains.

As explained in Section 2.2, we impose a transparent boundary condition on the artificial boundary Γ_H based on the DtN map. This results in the following boundary value problem posed in Ω_H^δ: For the incident field $u^i \in H_r^1(\Omega_H^\delta)$ with $|r| < 1$, find the scattered field $u^{s,\delta} \in H_r^1(\Omega_H^\delta)$ that satisfies

$$\Delta u^{s,\delta} + k^2 u^{s,\delta} = 0 \qquad \text{in } \Omega_H^\delta\,, \tag{4.1a}$$

$$u^{s,\delta} = -u^i \qquad \text{on } \Gamma^\delta\,, \tag{4.1b}$$

$$\partial_{x_d} u^{s,\delta} = \mathcal{T}^+ u^{s,\delta} \qquad \text{on } \Gamma_H\,, \tag{4.1c}$$

where the DtN map $\mathcal{T}^+$ is defined as in (2.9). Note that (4.1a) is understood in the distributional sense and (4.1b) and (4.1c) in the trace sense.

Considering the total field $u^\delta = u^{s,\delta} + u^i$, we have

$$\Delta u^\delta + k^2 u^\delta = 0 \qquad \text{in } \Omega_H^\delta\,, \tag{4.2a}$$

$$u^\delta = 0 \qquad \text{on } \Gamma^\delta\,, \tag{4.2b}$$

$$(\partial_{x_d} - \mathcal{T}^+) u^\delta = (\partial_{x_d} - \mathcal{T}^+) u^i \qquad \text{on } \Gamma_H\,. \tag{4.2c}$$

The variational form of (4.2) is to find $u^\delta \in \widetilde{H}_r^1(\Omega_H^\delta)$ for $|r| < 1$ such that

$$\begin{aligned} \left\langle \nabla u^\delta, \overline{\nabla v} \right\rangle_{\Omega_H^\delta} - k^2 \left\langle u^\delta, \overline{v} \right\rangle_{\Omega_H^\delta} \\ - \left\langle \mathcal{T}^+ u^\delta, \overline{v} \right\rangle_{\Gamma_H} = \left\langle (\partial_{x_d} - \mathcal{T}^+) u^i, \overline{v} \right\rangle_{\Gamma_H} \quad \text{for all } v \in \widetilde{H}_{-r}^1(\Omega_H^\delta)\,. \end{aligned} \tag{4.3}$$

For all incident fields $u^i \in H_r^1(\Omega_H^\delta)$ for $|r| < 1$, Problem (4.3) is uniquely solvable as proven in Theorem 2.15.

From the numerical point of view, the discretization of the variational problem (4.3) is not possible, since it is posed in a horizontally unbounded domain. To reduce the domain to a bounded cell, one possible choice is the *FB transform* (described in Section 2.4). However, since the domain Ω_H^δ is not periodic, the FB transform cannot be applied directly. A *coordinate mapping* must be defined to transform the perturbed domain into a periodic one. This is achievable because the

periodic and perturbed surfaces are explicitly given. However, this transformation results in a partial differential equation with non-constant coefficients.

Given that $h \in (\|\zeta^{\mathrm{per}}\|_\infty, H]$, we can define the *diffeomorphism* Ψ^δ from the periodic domain Ω_H^{per} to the locally perturbed domain Ω_H^δ as follows

$$\Psi^\delta(x) = (\widetilde{x}, x_d + \beta_h^\delta(x)\delta(\widetilde{x}))\,, \quad x = (\widetilde{x}, x_d) \in \Omega_H^{\mathrm{per}}\,, \tag{4.4}$$

where the non-constant coefficient is defined by

$$\beta_h^\delta(x) \coloneqq \frac{(x_d - h)^3}{(\zeta^{\mathrm{per}}(\widetilde{x}) - h)^3}\chi_h(x_d)\,, \tag{4.5}$$

with $\chi_h(x_d) = 0$ for $x_d \geq h$ and $\chi_h(x_d) = 1$ for $x_d < h$.

Remark 4.1. Based on the parameter h and the definition of β_h^δ, clearly $\beta_h^\delta = 0$ above the surface Γ_h and $\beta_h^\delta = 1$ on the periodic surface Γ^{per}.

Remark 4.2. It is important to mention that the support of $\Psi^\delta - I$ is located in the bounded cell $\Omega_H^{2\pi} \coloneqq ((-\pi,\pi)^{d-1} \times \mathbb{R}) \cap \Omega_H^{\mathrm{per}}$, since we assumed in the beginning of this chapter that the perturbation δ is compactly supported in $[-\pi,\pi]^{d-1}$.

The transformed total field $u_{\mathrm{tra}}^\delta \coloneqq u^\delta \circ \Psi^\delta \in \widetilde{H}_r^1(\Omega_H^{\mathrm{per}})$ with $|r| < 1$ now satisfies the following variational problem posed in the periodic domain Ω_H^{per}

$$\begin{aligned}\Big\langle A^\delta \nabla u_{\mathrm{tra}}^\delta, \overline{\nabla v}\Big\rangle_{\Omega_H^{\mathrm{per}}} - k^2\Big\langle c^\delta u_{\mathrm{tra}}^\delta, \overline{v}\Big\rangle_{\Omega_H^{\mathrm{per}}} \\ - \Big\langle \mathcal{T}^+ u_{\mathrm{tra}}^\delta, \overline{v}\Big\rangle_{\Gamma_H} = \Big\langle (\partial_{x_d} - \mathcal{T}^+)u^i, \overline{v}\Big\rangle_{\Gamma_H} \quad \text{for all } v \in \widetilde{H}_{-r}^1(\Omega_H^{\mathrm{per}})\,,\end{aligned} \tag{4.6}$$

where

$$\begin{aligned}A^\delta(x) &\coloneqq |\det \nabla\Psi^\delta(x)|\,(\nabla\Psi^\delta(x))^{-1}(\nabla\Psi^\delta(x))^{-\top} \in L^\infty(\Omega_H^{\mathrm{per}}, \mathbb{R}^{d\times d})\,,\\ c^\delta(x) &\coloneqq |\det \nabla\Psi^\delta(x)| \in L^\infty(\Omega_H^{\mathrm{per}})\,.\end{aligned} \tag{4.7}$$

Adding and subtracting $k^2\Big\langle u_{\mathrm{tra}}^\delta, \overline{v}\Big\rangle_{\Omega_H^{\mathrm{per}}}$ and $\Big\langle \nabla u_{\mathrm{tra}}^\delta, \overline{\nabla v}\Big\rangle_{\Omega_H^{\mathrm{per}}}$ to (4.6) leads to

$$a(u_{\mathrm{tra}}^\delta, v) + b^\delta(u_{\mathrm{tra}}^\delta, v) = \Big\langle (\partial_{x_d} - \mathcal{T}^+)u^i, \overline{v}\Big\rangle_{\Gamma_H} \quad \text{for all } v \in \widetilde{H}_{-r}^1(\Omega_H^{\mathrm{per}})\,, \tag{4.8}$$

where the sesquilinear forms $a, b^\delta \colon \widetilde{H}_r^1(\Omega_H^{\mathrm{per}}) \times \widetilde{H}_{-r}^1(\Omega_H^{\mathrm{per}}) \to \mathbb{C}$ are defined by

$$\begin{aligned}a(\phi,\psi) &\coloneqq \Big\langle \nabla\phi, \overline{\nabla\psi}\Big\rangle_{\Omega_H^{\mathrm{per}}} - k^2\Big\langle \phi, \overline{\psi}\Big\rangle_{\Omega_H^{\mathrm{per}}} - \Big\langle \mathcal{T}^+\phi, \overline{\psi}\Big\rangle_{\Gamma_H}\,,\\ b^\delta(\phi,\psi) &\coloneqq \Big\langle (A^\delta - I)\nabla\phi, \overline{\nabla\psi}\Big\rangle_{\Omega_H^{\mathrm{per}}} - k^2\Big\langle (c^\delta - 1)\phi, \overline{\psi}\Big\rangle_{\Omega_H^{\mathrm{per}}}\,.\end{aligned} \tag{4.9}$$

Note that the sesquilinear form a is exactly the same as in (3.2) for the periodic case. However, the term b^δ depends on the perturbation δ through the diffeomorphism Ψ^δ.

As the variational formulation (4.8) is now posed in the periodic domain Ω_H^{per}, we can apply the FB transform with respect to the first $d-1$ variables (see Definition 2.24). This leads to a

family of periodic problems, indexed by the Floquet parameter $\alpha \in \Lambda = [-1/2, 1/2]^{d-1}$, posed in the bounded cell $\Omega_H^{2\pi}$.

Since $\nabla\Psi^\delta = I$ outside the bounded cell $\Omega_H^{2\pi}$, we conclude that $A^\delta - I$ and $c^\delta - 1$ are both compactly supported in this cell. Therefore, by using the definition of the FB transform, denoted by $\mathcal{J}$, and some straightforward computations, we have

$$\mathcal{J}((A^\delta - I)\nabla u_{\mathrm{tra}}^\delta) = (A^\delta - I)(\nabla u_{\mathrm{tra}}^\delta)\mathrm{e}^{-\mathrm{i}\alpha\cdot\widetilde{x}},$$
$$\mathcal{J}((c^\delta - 1)u_{\mathrm{tra}}^\delta) = (c^\delta - 1)u_{\mathrm{tra}}^\delta\mathrm{e}^{-\mathrm{i}\alpha\cdot\widetilde{x}}.$$

Considering this fact and using the Plancherel formula (2.36) and Theorem 2.31, we obtain that for all incident fields $u^i \in H_r^1(\Omega_H^\delta)$ with $r \in [0,1)$, the FB transform of the total field, denoted by $w^\delta := \mathcal{J}u_{\mathrm{tra}}^\delta \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, satisfies

$$\int_\Lambda \Big(a_\alpha(w^\delta(\alpha), z(\alpha)) + b_\alpha^\delta(w^\delta, z(\alpha))\Big)\,\mathrm{d}\alpha = \int_\Lambda \Big\langle (\partial_{x_d} - \mathcal{T}_\alpha^+)\mathcal{J}u^i(\alpha), \overline{z(\alpha)}\Big\rangle_{\Gamma_H^{2\pi}}\,\mathrm{d}\alpha \tag{4.10}$$

for all $z \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, where the FB transform of the DtN map $\mathcal{T}_\alpha^+$ is given by (3.5) and the sesquilinear forms $a_\alpha\colon \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}) \times \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}) \to \mathbb{C}$ and $b_\alpha^\delta\colon L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \times \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}) \to \mathbb{C}$ are defined by

$$\begin{aligned} a_\alpha(\phi,\psi) := {} & \Big\langle \nabla_x\phi, \overline{\nabla_x\psi}\Big\rangle_{\Omega_H^{2\pi}} - 2\mathrm{i}\Big\langle \alpha\cdot\nabla_{\widetilde{x}}\phi, \overline{\psi}\Big\rangle_{\Omega_H^{2\pi}} \\ & - (k^2 - |\alpha|^2)\Big\langle \phi, \overline{\psi}\Big\rangle_{\Omega_H^{2\pi}} - \Big\langle \mathcal{T}_\alpha^+\phi, \overline{\psi}\Big\rangle_{\Gamma_H^{2\pi}} \end{aligned} \tag{4.11}$$

and

$$b_\alpha^\delta(\phi,\psi) := \Big\langle (A^\delta - I)\nabla_x\mathcal{J}^{-1}\phi, \overline{\nabla_x(\psi\mathrm{e}^{\mathrm{i}\alpha\cdot\widetilde{x}})}\Big\rangle_{\Omega_H^{2\pi}} - k^2\Big\langle (c^\delta - 1)\mathcal{J}^{-1}\phi, \overline{\psi\mathrm{e}^{\mathrm{i}\alpha\cdot\widetilde{x}}}\Big\rangle_{\Omega_H^{2\pi}}. \tag{4.12}$$

The following theorem states that problems (4.8) and (4.10) are equivalent. Afterwards, the unique solvability of (4.10) can be shown.

Theorem 4.3. *Let the incident field $u^i \in H_r^1(\Omega_H^\delta)$ for $r \in [0,1)$. Then, $u_{\mathrm{tra}}^\delta \in \widetilde{H}_r^1(\Omega_H^{\mathrm{per}})$ is the solution of (4.8) if and only if $w^\delta = \mathcal{J}u_{\mathrm{tra}}^\delta \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ satisfies (4.10).*

Proof. See [83, Thm. 4.1]. □

Theorem 4.4. *Let Γ^δ be the graph of the Lipschitz continuous function ζ^δ. Then, the variational problem (4.10) has a unique solution in $L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ for all incident fields $u^i \in H_r^1(\Omega_H^\delta)$ for $r \in [0,1)$.*

Proof. See [83, Thm. 4.2]. □

The next theorem provides an auxiliary result regarding the regularity of the solution.

Theorem 4.5. *Assume that Γ^δ is the graph of a C^2-function. If the incident field $u^i \in H_r^2(\Omega_H^\delta)$ for $r \in [0,1)$, then $w^\delta(\alpha) \in \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$ for almost all $\alpha \in \Lambda$ and $u_{\mathrm{tra}}^\delta = \mathcal{J}^{-1}w^\delta \in \widetilde{H}^2(\Omega_H^{\mathrm{per}})$.*

Proof. See [83, Thm. 4.3]. □

We can also give an alternative formulation of (4.10) for almost every $\alpha \in \Lambda$.

Theorem 4.6. *Let Γ^δ as in Theorem 4.4 and $u^i \in H^1_r(\Omega^\delta_H)$ for $r \in [0,1)$. Then, the variational formulation (4.10) is equivalent to find $w^\delta \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$, which satisfies*

$$a_\alpha(w^\delta(\alpha), z) + b^\delta_\alpha(w^\delta, z) = \left\langle (\partial_{x_d} - \mathcal{T}^+_\alpha)\mathcal{J}u^i(\alpha), \overline{z} \right\rangle_{\Gamma^{2\pi}_H} \tag{4.13}$$

for almost all $\alpha \in \Lambda$ and all $z \in \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H)$. Additionally, in the two-dimensional case, if $r \in (1/2, 1)$, Problem (4.10) is equivalent to find $w^\delta \in C(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$ such that (4.13) holds for all $\alpha \in \Lambda$.

Proof. See [83, Thm. 4.4]. □

Note that when $A^\delta = I$ and $c^\delta = 1$ in (4.12), we have $b^\delta_\alpha(w^\delta, z) = 0$. Problem (4.13) hence reduces to the periodic problem described in (3.7). In this case, the advantage of applying the FB transform is that one obtains a decoupled family of periodic problems indexed by the Floquet parameter α. Therefore, their solutions can be computed in parallel.

On the other hand, in the general (perturbed) case, where the term $b^\delta_\alpha(w^\delta, z)$ is non-zero, we obtain a family of periodic problems that are fully *coupled* through this additional term. From the definition of the inverse FB transform, it turns out that solving problem (4.13) for each α requires the contribution of the transformed fields, evaluated in all α. Therefore, a naive discretization will lead to a very large linear system, demanding a prohibitive computational cost.

From Chapter 3, we know that the transformed solution is not analytic with respect to α due to the singularity of the DtN operator $\mathcal{T}^+_\alpha$. Approaches similar to those in Chapter 3, which are directly based on the regularity of the DtN operator $\mathcal{T}^+_\alpha$ and use tailor-made inversion formulas are possible, but require substantial computational effort, particularly in the three-dimensional case. A simpler way to obtain a numerical approximation of $\mathcal{T}^+_\alpha$ is provided by the PML, which will be the focus of the following section.

4.2. The PML Approximation of the Solution

Here we use the PML method to approximate the scattered field $u^{s,\delta}$ satisfying (4.1). We first denote by $u^{s,\delta}_\sigma$ the PML approximation of the scattered field and recall from Section 2.3 that the parameter $\sigma \in \mathbb{C}$ controls the absorbing effect of the PML.

As described in Section 2.3, the PML approximation of the scattered field satisfies (4.1), however with a modified boundary condition on Γ_H. More precisely, for any incident field $u^i \in H^1_r(\Omega^\delta_H)$ with $r \in [0,1)$, the PML approximation $u^{s,\delta}_\sigma \in H^1(\Omega^\delta_H)$ satisfies

$$\Delta u^{s,\delta}_\sigma + k^2 u^{s,\delta}_\sigma = 0 \qquad \text{in } \Omega^\delta_H\,, \tag{4.14a}$$

$$u^{s,\delta}_\sigma = -u^i \qquad \text{on } \Gamma^\delta\,, \tag{4.14b}$$

$$\partial_{x_d} u^{s,\delta}_\sigma = \mathcal{T}^+_\sigma u^{s,\delta}_\sigma \quad \text{on } \Gamma_H\,, \tag{4.14c}$$

where $\mathcal{T}^+_\sigma$ is the PML approximation of the DtN map defined in (2.29).

According to Theorem 2.21, problem (4.14) is uniquely solvable, when the PML parameter σ is sufficiently large.

Proceeding in the same way as before, using the diffeomorphism Ψ^δ, we see that the total field

$$u^\delta_{\mathrm{tra},\sigma} := (u^{s,\delta}_\sigma + u^i) \circ \Psi^\delta \in \widetilde{H}^1(\Omega^{\mathrm{per}}_H)$$

satisfies

$$a_\sigma(u^\delta_{\mathrm{tra},\sigma}, v) + b^\delta(u^\delta_{\mathrm{tra},\sigma}, v) = \left\langle (\partial_{x_d} - \mathcal{T}^+_\sigma) u^i, \overline{v} \right\rangle_{\Gamma_H} \quad \text{for all } v \in \widetilde{H}^1(\Omega^{\mathrm{per}}_H)\,, \tag{4.15}$$

where the sesquilinear form $a_\sigma \colon \widetilde{H}^1(\Omega^{\mathrm{per}}_H) \times \widetilde{H}^1(\Omega^{\mathrm{per}}_H) \to \mathbb{C}$ is given by

$$a_\sigma(\phi, \psi) := \left\langle \nabla\phi, \overline{\nabla\psi} \right\rangle_{\Omega^{\mathrm{per}}_H} - k^2 \left\langle \phi, \overline{\psi} \right\rangle_{\Omega^{\mathrm{per}}_H} - \left\langle \mathcal{T}^+_\sigma \phi, \overline{\psi} \right\rangle_{\Gamma_H},$$

and b^δ is defined as in (4.9).

Applying the FB transform to the total field, we obtain that for the incident field $u^i \in H^1_r(\Omega^\delta_H)$ with $r \in [0,1)$, the transformed solution $w^\delta_\sigma := \mathcal{J} u^\delta_{\mathrm{tra},\sigma} \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$ satisfies

$$\int_\Lambda \left(a_{\alpha,\sigma}(w^\delta_\sigma(\alpha), z(\alpha)) + b^\delta_\alpha(w^\delta_\sigma, z(\alpha)) \right) \mathrm{d}\alpha = \int_\Lambda \left\langle (\partial_{x_d} - \mathcal{T}^+_{\alpha,\sigma}) \mathcal{J} u^i(\alpha), \overline{z(\alpha)} \right\rangle_{\Gamma^{2\pi}_H} \mathrm{d}\alpha \tag{4.16}$$

for all $z \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$, where the sesquilinear form b^δ_α is given in (4.12), and $a_{\alpha,\sigma}$ is defined as (4.11) with $\mathcal{T}^+_\alpha$ replaced by the PML approximation $\mathcal{T}^+_{\alpha,\sigma}$. This approximation is given by

$$(\mathcal{T}^+_{\alpha,\sigma}\phi)(\widetilde{x}, H) := \mathrm{i} \sum_{j \in \mathbb{Z}^{d-1}} \sqrt{k^2 - |\alpha - j|^2}\, \coth\left(-\mathrm{i}\sigma \sqrt{k^2 - |\alpha - j|^2} \right) \widehat{\phi}(j) \mathrm{e}^{\mathrm{i} j \cdot \widetilde{x}}\,, \tag{4.17}$$

where $\widehat{\phi}(j)$ denotes the j-th Fourier coefficient of ϕ (see [105, Eq. (16)]).

Remark 4.7. For $u^i \in H^1_r(\Omega^\delta_H)$ with $r \in [0,1)$, similar to Theorem 4.6, the variational formulation (4.16) is equivalent to find $w^\delta_\sigma \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$ such that

$$a_{\alpha,\sigma}(w^\delta_\sigma(\alpha), z) + b^\delta_\alpha(w^\delta_\sigma, z) = \left\langle (\partial_{x_d} - \mathcal{T}^+_{\alpha,\sigma}) \mathcal{J} u^i(\alpha), \overline{z} \right\rangle_{\Gamma^{2\pi}_H} \tag{4.18}$$

for almost all $\alpha \in \Lambda$ and all $z \in \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H)$. Additionally, in the two-dimensional case, if $r \in (1/2, 1)$, Problem (4.16) is equivalent to find $w^\delta_\sigma \in C(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$ such that (4.18) holds for all $\alpha \in \Lambda$.

The following theorem states that (4.16) is uniquely solvable for sufficiently large σ.

Theorem 4.8. *Assume that Γ^δ is the graph of the Lipschitz continuous function ζ^δ with a sufficiently small perturbation δ. The variational problem (4.16) for sufficiently large σ has a unique solution in $L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega^{2\pi}_H))$ for the incident fields $u^i \in H^1_r(\Omega^\delta_H)$ with $r \in [0,1)$.*

Proof. Let $\mathcal{A}_{\alpha,\sigma}\colon \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}) \to (\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))^*$ and $\mathcal{B}^\delta_\alpha\colon L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \to (\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))^*$ be induced by the sesquilinear forms $a_{\alpha,\sigma}$ and b^δ_α such that

$$\left\langle \mathcal{A}_{\alpha,\sigma}\phi, \overline{\psi}\right\rangle_{\Omega_H^{2\pi}} \coloneqq a_{\alpha,\sigma}(\phi,\psi) \quad \text{and} \quad \left\langle \mathcal{B}^\delta_\alpha\phi, \overline{\psi}\right\rangle_{\Omega_H^{2\pi}} \coloneqq b^\delta_\alpha(\phi,\psi)\,. \tag{4.19}$$

Moreover, let the antilinear form $\mathcal{G}_{\alpha,\sigma}$ be defined by

$$\left\langle \mathcal{G}_{\alpha,\sigma}, \overline{\psi}\right\rangle_{\Gamma_H^{2\pi}} \coloneqq \left\langle (\partial_{x_d} - \mathcal{T}^+_{\alpha,\sigma})\mathcal{J}u^i(\alpha), \overline{\psi}\right\rangle_{\Gamma_H^{2\pi}}. \tag{4.20}$$

Now, the operator form of (4.18) can be written as

$$\mathcal{A}_{\alpha,\sigma}w^\delta_\sigma(\alpha) + \mathcal{B}^\delta_\alpha w^\delta_\sigma = \mathcal{G}_{\alpha,\sigma} \quad \text{for almost all } \alpha\in\Lambda\,. \tag{4.21}$$

As mentioned in Remark 2.23, we know that $\|\mathcal{A}_{\alpha,\sigma} - \mathcal{A}_\alpha\| \to 0$ as $|\sigma|\to\infty$, where $\mathcal{A}_\alpha$ is induced by the sesquilinear form (4.11) with the DtN operator $\mathcal{T}^+_\alpha$. Since the operator $\mathcal{A}_\alpha$ is boundedly invertible based on Theorem 4.4 for almost all $\alpha\in\Lambda$, we conclude that the operator $\mathcal{A}_{\alpha,\sigma}$ is also boundedly invertible for sufficiently large σ. Therefore, we can define the operator $\mathcal{D}_\sigma\colon L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \to L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ by

$$(\mathcal{D}_\sigma w^\delta_\sigma)(\alpha) \coloneqq \mathcal{A}^{-1}_{\alpha,\sigma}\left(\mathcal{G}_{\alpha,\sigma} - \mathcal{B}^\delta_\alpha w^\delta_\sigma\right). \tag{4.22}$$

By this definition, we can reformulate equation (4.21) as the *fixed-point problem*

$$\mathcal{D}_\sigma w^\delta_\sigma = w^\delta_\sigma\,. \tag{4.23}$$

To apply the Banach fixed point theorem (see [9, Thm. 4.1.3]), it is sufficient to show that the operator $\mathcal{D}_\sigma$ is a contraction, i.e., for some $q<1$ and all $w^\delta_\sigma, \widetilde{w}^\delta_\sigma \in L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, there holds

$$\left\|\mathcal{D}_\sigma w^\delta_\sigma - \mathcal{D}_\sigma \widetilde{w}^\delta_\sigma\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \le q\left\|w^\delta_\sigma - \widetilde{w}^\delta_\sigma\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}.$$

From the definition of the operator $\mathcal{D}_\sigma$, we have

$$\left\|\mathcal{D}_\sigma w^\delta_\sigma - \mathcal{D}_\sigma \widetilde{w}^\delta_\sigma\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \le \left\|\mathcal{A}^{-1}_{\alpha,\sigma}\mathcal{B}^\delta_\alpha\right\|\left\|w^\delta_\sigma - \widetilde{w}^\delta_\sigma\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}.$$

Due to the definition of the operator $\mathcal{B}^\delta_\alpha$ in (4.19), we know that $\left\|\mathcal{B}^\delta_\alpha\right\| \to 0$ as $\|\delta\|_{1,\infty}\to 0$. As a consequence, $q \coloneqq \left\|\mathcal{A}^{-1}_{\alpha,\sigma}\mathcal{B}^\delta_\alpha\right\| < 1$ for sufficiently small perturbations δ. □

A question that naturally arises here is whether we still need to choose a tailor-made quadrature rule to compute the inverse FB transform. To be able to provide an answer, it is necessary to analyze the regularity of the PML approximation of the transformed field with respect to α.

4.3. REGULARITY OF THE PML APPROXIMATION OF THE TRANSFORMED SOLUTION

In this section, we aim to obtain a representation for w_σ^δ and thus analyze its regularity with respect to α.

From Section 3.2, we recall the fact that the transformed field inherits the regularity of the DtN operator with respect to α. Therefore, the transformed field w^δ, satisfying (4.10), is not analytic due to the (square root) singularities of the DtN operator $\mathcal{T}_\alpha^+$. From (4.16), it follows that the regularity of the PML approximation w_σ^δ depends on the PML approximation of the DtN operator, i.e., $\mathcal{T}_{\alpha,\sigma}^+$. The following theorem states that w_σ^δ is analytic with respect to $\alpha \in \Lambda$. This analyticity is a key advantage of the PML approach.

Theorem 4.9. *Let $u^i \in H^1(\Omega_H^\delta)$ and $\mathcal{J}u^i$ be analytic with respect to α. Then, for sufficiently large σ, the PML approximation w_σ^δ that solves (4.16) is analytic with respect to the Floquet parameter α.*

Proof. Let the operators $\mathcal{A}_{\alpha,\sigma}$, $\mathcal{B}_\alpha^\delta$ and the antilinear operator $\mathcal{G}_{\alpha,\sigma}$ be given as in the proof of Theorem 4.8. From Theorem 3.6, we recall that the transformed field w^δ satisfying (4.10) has singularities in each $\alpha \in \Sigma$, defined in (3.8). In two dimensions, this set has at most two singular points, while in three dimensions, it consists of the union of all arcs centered at points in $\mathbf{J}(\alpha)$ (defined in (3.9)). We focus here on the three-dimensional case. However, the result also holds in two dimensions, where $\mathbf{J}(\alpha)$ contains at most two elements.

Now, we show that w_σ^δ is analytic everywhere, including on the set Σ. For $\alpha_0 \in \Lambda$, let $B(\alpha_0, \rho)$ denote an open ball centred at α_0, with radius ρ. Using the definition of $\mathcal{T}_{\alpha,\sigma}^+$ given in (4.17), we can write

$$\mathcal{T}_{\alpha,\sigma}^+ = \mathcal{T}_{\alpha,\sigma}^{+,0} + \sum_{j \in \mathbf{J}(\alpha_0)} K_{\alpha,\sigma}(j)\mathcal{C}(j)\,, \tag{4.24}$$

where $\mathcal{T}_{\alpha,\sigma}^{+,0}, \mathcal{C}(j)\colon H_{\mathrm{per}}^{1/2}(\Gamma_H^{2\pi}) \to H_{\mathrm{per}}^{-1/2}(\Gamma_H^{2\pi})$ are defined by

$$\mathcal{T}_{\alpha,\sigma}^{+,0} \coloneqq \sum_{j \notin \mathbf{J}(\alpha_0)} K_{\alpha,\sigma}(j)\mathcal{C}(j) \quad \text{and} \quad \mathcal{C}(j)\phi \coloneqq \widehat{\phi}(j)\,\mathrm{e}^{\mathrm{i}j\cdot\widetilde{x}}\,,$$

with

$$K_{\alpha,\sigma}(j) \coloneqq \mathrm{i}\sqrt{k^2 - |\alpha - j|^2}\coth\left(-\mathrm{i}\sigma\sqrt{k^2 - |\alpha - j|^2}\right).$$

Now, we can decompose the operator $\mathcal{A}_{\alpha,\sigma}$ given in (4.19) as

$$\mathcal{A}_{\alpha,\sigma} = \mathcal{A}_\alpha^0 - \sum_{j \in \mathbf{J}(\alpha_0)} K_{\alpha,\sigma}(j)\,\mathcal{C}(j)\,, \tag{4.25}$$

where $\mathcal{A}_\alpha^0\colon \widetilde{H}_{\mathrm{per}}^1(\Omega_H^{2\pi}) \to (\widetilde{H}_{\mathrm{per}}^1(\Omega_H^{2\pi}))^*$ is defined by

$$\begin{aligned}\left\langle \mathcal{A}_\alpha^0\phi, \overline{\psi}\right\rangle_{\Omega_H^{2\pi}} \coloneqq \left\langle \nabla_x\phi, \overline{\nabla_x\psi}\right\rangle_{\Omega_H^{2\pi}} - 2\mathrm{i}\left\langle \alpha\cdot\nabla_{\widetilde{x}}\,\phi, \overline{\psi}\right\rangle_{\Omega_H^{2\pi}} \\ - (k^2 - |\alpha|^2)\left\langle \phi, \overline{\psi}\right\rangle_{\Omega_H^{2\pi}} - \left\langle \mathcal{T}_{\alpha,\sigma}^{+,0}\phi, \overline{\psi}\right\rangle_{\Gamma_H^{2\pi}}.\end{aligned}$$

Substituting (4.25) into Problem (4.21), we see that w_σ^δ satisfies

$$\left(\mathcal{A}_\alpha^0 - \sum_{j\in\mathbf{J}(\alpha_0)} K_{\alpha,\sigma}(j)\,\mathcal{C}(j)\right) w_\sigma^\delta(\alpha) = -\mathcal{B}_\alpha^\delta w_\sigma^\delta + \mathcal{G}_{\alpha,\sigma} \quad \text{for almost all } \alpha\in\Lambda\,. \tag{4.26}$$

According to Theorem 4.8, the operator on the left-hand side of (4.26) is boundedly invertible for sufficiently large σ. To obtain a representation for $w_\sigma^\delta(\alpha)$ allowing us to analyze the regularity with respect to α, we show that the operator $\mathcal{A}_\alpha^0$ is boundedly invertible. Using the perturbation theorem given in [75, Thm. 10.1], it is sufficient to show that $\mathcal{A}_\alpha^0$ is a small perturbation of the left-hand side of (4.26). To this end, we first recall the Laurent expansion of coth, i.e.,

$$\coth(z) = \frac{1}{z} + \frac{z}{3} - \frac{z^3}{45} + \cdots = \sum_{n=0}^{\infty} \frac{2^{2n}B_{2n}}{(2n)!} z^{2n-1} \quad \text{for } 0<|z|<\pi\,,$$

where B_{2n} denotes the Bernoulli numbers. Similar to [105, Lem. 10], from the definition of $K_{\alpha,\sigma}$ and a straightforward computation, we conclude

$$\begin{aligned} K_{\alpha,\sigma}(j) &= \mathrm{i}\sqrt{k^2-|\alpha-j|^2}\sum_{n=0}^{\infty}\frac{2^{2n}B_{2n}}{(2n)!}\left(-\mathrm{i}\sigma\sqrt{k^2-|\alpha-j|^2}\right)^{2n-1} \\ &= -\frac{1}{\sigma}\sum_{n=0}^{\infty}\frac{2^{2n}B_{2n}(-\mathrm{i}\sigma)^{2n}}{(2n)!}\left(k^2-|\alpha-j|^2\right)^n. \end{aligned} \tag{4.27}$$

For each $j\in\mathbf{J}(\alpha_0)$ (i.e., $k=|\alpha_0-j|$) and every $\alpha\in B(\alpha_0,\rho)$, we obtain

$$\begin{aligned} |K_{\alpha,\sigma}(j)| &\le \frac{1}{|\sigma|}\sum_{n=0}^{\infty}\frac{2^{2n}B_{2n}|-\mathrm{i}\sigma|^{2n}}{(2n)!}\left||\alpha_0-j|^2-|\alpha-j|^2\right|^n \\ &\le \sum_{n=0}^{\infty}\frac{2^{2n}B_{2n}|\sigma|^{2n-1}}{(2n)!}\left(|\alpha_0-\alpha|\cdot|\alpha_0+\alpha+2j|\right)^n \\ &\le \sum_{n=0}^{\infty} C_n|\alpha-\alpha_0|^n\,, \end{aligned}$$

where the constants C_n depend on σ. For sufficiently large σ, it is clear that $|K_{\alpha,\sigma}(j)|\to 0$ as $|\alpha-\alpha_0|\to 0$. This shows that $\mathcal{A}_\alpha^0$ is boundedly invertible for all $\alpha\in B(\alpha_0,\rho)$.

Setting $\widetilde{\mathcal{C}}(j)=(\mathcal{A}_\alpha^0)^{-1}\mathcal{C}(j)$, we can write the solution w_σ^δ of (4.26) by the Neumann expansion

$$w_\sigma^\delta(\alpha) = \sum_{n=0}^{\infty}\left(\sum_{j\in\mathbf{J}(\alpha_0)} K_{\alpha,\sigma}(j)\widetilde{\mathcal{C}}(j)\right)^n (\mathcal{A}_\alpha^0)^{-1}\left(-\mathcal{B}_\alpha^\delta w_\sigma^\delta + \mathcal{G}_{\alpha,\sigma}\right). \tag{4.28}$$

We now want to analyze the regularity of the transformed field $w_\sigma^\delta(\alpha)$ with respect to α using the same technique as Theorem 3.6. Note that the operators $\widetilde{\mathcal{C}}(j)$, $\mathcal{A}_\alpha^0$ and the function $\mathcal{B}_\alpha^\delta w_\sigma^\delta$ depend analytically on α. Therefore, it remains to analyze the regularity of $K_{\alpha,\sigma}(j)$ and $\mathcal{G}_{\alpha,\sigma}$. Based on the definition of $\mathcal{G}_{\alpha,\sigma}$ given in (4.20) and using (4.24), we see that the regularity of $\mathcal{G}_{\alpha,\sigma}$ is only dependent on $K_{\alpha,\sigma}(j)$ for $j\in\mathbf{J}(\alpha_0)$.

Next, we distinguish two cases. If $\alpha_0 \notin \Sigma$, then $\mathbf{J}(\alpha_0) = \emptyset$ and $w^\delta \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ depends analytically on α. On the other hand, let $\alpha_0 \in \Sigma$. From (4.27), we see that $K_{\alpha,\sigma}(j)$ is analytic in a neighborhood of α_0. Hence by using (4.28), we find that the function w^δ_σ is also analytic in a neighborhood of α_0. Since this fact holds for every $\alpha_0 \in \Lambda$, w^δ_σ is globally analytic with respect to α. □

4.4. Convergence of the PML Approximation in Two Dimensions

Recently, there has been a number of results published on the convergence of the PML approximation of the solution with respect to the damping parameter σ (see, e.g., [26, 105, 106]). Instead of a scattering problem, in these works *source problems* are considered. Here, we aim to extend these results to scattering problems in two dimensions and show the exponential convergence of the PML approximation on every compact subset of the unbounded periodic and locally perturbed domains.

Let the transformed total field w^δ and its PML approximation w^δ_σ satisfy (4.10) and (4.16), respectively. Since (4.10) depends on the exact DtN map, we know from Section 3.2 that w^δ is not analytic with respect to $\alpha \in \Lambda$. Before proceeding to the convergence analysis, we introduce some preliminaries to modify the integration path in the definition of the inverse FB transform (see (2.35)). Along this path, the function w^δ remains analytic (see [105, Sec. 3]).

Afterwards, we focus on source problems in the periodic structure Ω_H^{per} and outline the convergence results given in [105] for the periodic domain. Finally, we extend these results to scattering problems and show in Sections 4.4.2 and 4.4.3 that the PML approximation of the scattered field is exponentially convergent in every compact subset of the periodic and locally perturbed domains.

4.4.1. Analytic Extension of the Transformed Solution

Recall the regularity results of the transformed solution $w^\delta(\alpha)$ and its PML approximation $w^\delta_\sigma(\alpha)$ from Sections 3.2 and 4.3 and introduce $\gamma_j(\alpha) := \sqrt{k^2 - |\alpha - j|^2}$ for $\alpha \in \Lambda = [-1/2, 1/2]$.

Definition 4.10. Any point $\alpha \in \Lambda$ satisfying $|\alpha - j| = k$ for some $j \in \mathbb{Z}$ is called a *cutoff value*.

When the wave number k is a half-integer, then for a cutoff value $\alpha \in \Lambda$, there exist two integers $j_1, j_2 \in \mathbb{Z}$ such that $|\alpha - j_1| = |\alpha - j_2| = k$. This situation is more involved and will not be addressed in the analysis. Therefore, we impose the following assumption.

Assumption 4.11. *The wave number k satisfies $k \neq \dfrac{m}{2}$ for all $m \in \mathbb{N}$.*

Under Assumption 4.11, there exists a non-negative integer $\widehat{j}$ and a number $\kappa \in (-1/2, 1/2)\backslash\{0\}$ such that $k = \widehat{j} + \kappa$. It follows that $\gamma_{-\widehat{j}}(\kappa) = \gamma_{+\widehat{j}}(-\kappa) = 0$. The numbers $\pm\kappa$ are the only roots of $\gamma_{\mp\widehat{j}}$ in Λ and are cutoff values.

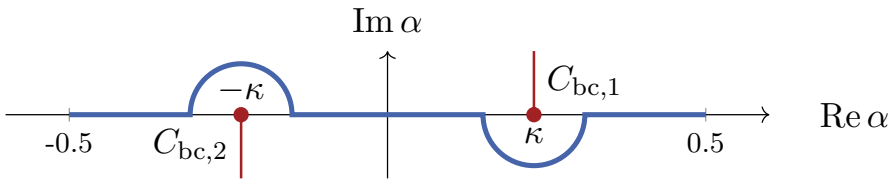


FIGURE 4.2. Sketch of the branch cuts (in red) and the integration path $\mathcal{E}$ (in blue) for $\kappa > 0$.

Assumption 4.12. *A function $f\colon \Lambda \to H^1_{\mathrm{per}}(\Omega_H^{2\pi})$ satisfies this assumption if on $\Lambda \setminus \{\pm\kappa\}$, f depends analytically on α, and there exist open neighborhoods $U_\pm$ of $\pm\kappa$ and analytic functions $f_{\pm,1}$, $f_{\pm,2}\colon U_\pm \to H^1_{\mathrm{per}}(\Omega_H^{2\pi})$, such that*

$$f(\alpha) = f_{\pm,1}(\alpha) + \gamma_{\mp\hat{j}}(\alpha)\, f_{\pm,2}(\alpha)\,, \quad \alpha \in U_\pm\,.$$

According to Theorem 3.6, we see that if $\mathcal{J}u^i \in L^2(\Lambda; H^{1/2}(\Gamma_H^{2\pi}))$ satisfies Assumption 4.12, then also the solution w^δ of (4.10) satisfies Assumption 4.12. This allows us to analytically extend w^δ into (parts of) the complex plane. For this purpose, we shift the branch cut of the function $r(z) = z^{1/2}$, $z \in \mathbb{C} \setminus \{0\}$, from the negative real axis to the curve $C_{\mathrm{bc},0} = \{t^2 - 2\mathrm{i}kt : t > 0\}$. Then, the function $\gamma_{-\hat{j}}$ can be analytically extended to $(\Lambda + \mathrm{i}\mathbb{R}) \setminus C_{\mathrm{bc},1}$ and $\gamma_{+\hat{j}}$ to $(\Lambda + \mathrm{i}\mathbb{R}) \setminus C_{\mathrm{bc},2}$ with the branch cuts

$$C_{\mathrm{bc},1} = \kappa + \mathrm{i}\mathbb{R}_{>0} \qquad \text{and} \qquad C_{\mathrm{bc},2} = -\kappa - \mathrm{i}\mathbb{R}_{>0}\,,$$

respectively. A sketch of the branch cuts $C_{\mathrm{bc},1}$ and $C_{\mathrm{bc},2}$ has been plotted (in red) in Figure 4.2 for $\kappa > 0$. As a consequence, the DtN operator $\mathcal{T}_\alpha^+$ can also be analytically extended to $(\Lambda + \mathrm{i}\mathbb{R}) \setminus (C_{\mathrm{bc},1} \cup C_{\mathrm{bc},2})$. If Assumption 4.12 is satisfied by $\mathcal{J}u^i$, the same analytic extension is valid for w^δ. Moreover, from Theorem 4.9 we know that the PML approximation w_σ^δ is analytic for every $\alpha \in \Lambda$. Hence, it can also be analytically extended to $(\Lambda + \mathrm{i}\mathbb{R}) \setminus (C_{\mathrm{bc},1} \cup C_{\mathrm{bc},2})$.

In order to avoid the cutoff values $\kappa \in (-1/2, 1/2) \setminus \{0\}$, we are going to modify the integration path in the definition of the inverse FB transform. For sufficiently small $\varepsilon > 0$, the integration path $\mathcal{E}$ is defined by

$$\mathcal{E} := \Lambda \setminus [(\kappa - \varepsilon, \kappa + \varepsilon) \cup (-\kappa - \varepsilon, -\kappa + \varepsilon)] \cup \mathcal{E}_+ \cup \mathcal{E}_-\,, \tag{4.29}$$

where $\mathcal{E}_\pm$ denote the semi-circles around the cutoff values $\pm\kappa$ as

$$\mathcal{E}_\pm := \left\{ \pm\kappa \mp \varepsilon\, \mathrm{e}^{\mathrm{i}\vartheta} : \ \vartheta \in (0, \pi) \right\}.$$

The integration path $\mathcal{E}$ has been illustrated (in blue) in Figure 4.2. Therefore, instead of using (2.35), we may compute the inverse FB transform of the transformed field w^δ and its PML approximation w_σ^δ by

$$u^\delta_{\mathrm{tra}} = \mathcal{J}^{-1} w^\delta(x) = \int_{\mathcal{E}} w^\delta(\alpha; x)\, \mathrm{e}^{\mathrm{i}\alpha x_1}\, \mathrm{d}\alpha\,, \quad x = (x_1, x_2) \in \Omega_H^{\mathrm{per}} \tag{4.30}$$

and

$$u^{\delta}_{\mathrm{tra},\sigma} = \mathcal{J}^{-1} w^{\delta}_{\sigma}(x) = \int_{\mathcal{E}} w^{\delta}_{\sigma}(\alpha; x)\, \mathrm{e}^{\mathrm{i}\alpha x_1}\, \mathrm{d}\alpha\,, \quad x \in \Omega^{\mathrm{per}}_H\,. \tag{4.31}$$

Theorem 4.13. *Let k be chosen as in Assumption 4.11, $\kappa \in (-1/2, 1/2) \setminus \{0\}$ such that w^{δ} satisfy Assumption 4.12 for $\alpha \in \Lambda$ and $\mathcal{E}$ be the integration path defined in (4.29). Moreover, assume that the total field $u^{\delta}_{\mathrm{tra}}$ and its PML approximation $u^{\delta}_{\mathrm{tra},\sigma}$ are the solutions of (4.8) and (4.15), respectively. Then, for any compact subset $K \subseteq \overline{\Omega^{\mathrm{per}}_H}$ and sufficiently large σ, there holds*

$$\left\| u^{\delta}_{\mathrm{tra}} - u^{\delta}_{\mathrm{tra},\sigma} \right\|_{H^1(K)} \leq C \left\| w^{\delta} - w^{\delta}_{\sigma} \right\|_{C(\mathcal{E}; H^1_{\mathrm{per}}(\Omega^{2\pi}_H))}\,,$$

where the constant C depends on $\max_{x\in K}|x_1|$.

Proof. From the definition of $u^{\delta}_{\mathrm{tra}}$ and $u^{\delta}_{\mathrm{tra},\sigma}$ given in (4.30) and (4.31), we obtain

$$\begin{aligned}
\left\| u^{\delta}_{\mathrm{tra}} - u^{\delta}_{\mathrm{tra},\sigma} \right\|_{H^1(K)} &\leq \int_{\mathcal{E}} \left\| (w^{\delta} - w^{\delta}_{\sigma})(\alpha) \mathrm{e}^{\mathrm{i}\alpha x_1} \right\|_{H^1(K)} \mathrm{d}\alpha \\
&\leq C \int_{\mathcal{E}} \left\| (w^{\delta} - w^{\delta}_{\sigma})(\alpha) \right\|_{H^1(K)} \left\| \mathrm{e}^{\mathrm{i}\alpha x_1} \right\|_{H^1(K)} \mathrm{d}\alpha \\
&\leq C \max_{\alpha\in\mathcal{E}} \left(\left\| (w^{\delta} - w^{\delta}_{\sigma})(\alpha) \right\|_{H^1(K)} \left\| \mathrm{e}^{\mathrm{i}\alpha x_1} \right\|_{H^1(K)} \right)\,,
\end{aligned}$$

where C denotes a generic constant depending on the length of $\mathcal{E}$. Based on the integration path $\mathcal{E}$, we know that $\mathrm{Im}\,(\alpha) \in [-\varepsilon, \varepsilon]$. Furthermore, $w^{\delta}(\alpha)$ and $w^{\delta}_{\sigma}(\alpha)$ are periodic and the set K is bounded. Therefore, this yields

$$\begin{aligned}
\left\| u^{\delta}_{\mathrm{tra}} - u^{\delta}_{\mathrm{tra},\sigma} \right\|_{H^1(K)} &\leq C \max_{x\in K} \mathrm{e}^{\varepsilon|x_1|} \max_{\alpha\in\mathcal{E}} \left\| (w^{\delta} - w^{\delta}_{\sigma})(\alpha) \right\|_{H^1(K)} \\
&\leq C \max_{\alpha\in\mathcal{E}} \left\| (w^{\delta} - w^{\delta}_{\sigma})(\alpha) \right\|_{H^1(\Omega^{2\pi}_H)}\,,
\end{aligned}$$

where C denotes a generic constant depending on the length of $\mathcal{E}$, measure of K, the radius ε in the integration path $\mathcal{E}$ and $\max_{x\in K}|x_1|$. This completes the proof. □

Since the transformed total field w^{δ} and its PML approximation w^{δ}_{σ} are analytic for each $\alpha \in \mathcal{E}$, we can study the convergence of the PML approximation with respect to the PML parameter σ in the periodic setting and afterwards we generalize it to the locally periodic case.

4.4.2. The Periodic Case

Let $g \in L^2(\Omega^{\mathrm{per}}_H)$ be a compactly supported source in the periodic domain Ω^{per}_H. The aim is to find $v \in \widetilde{H}^1(\Omega^{\mathrm{per}}_H)$ such that

$$\begin{aligned}
\Delta v + k^2 v &= g \qquad \text{in } \Omega^{\mathrm{per}}_H\,, \\
v &= 0 \qquad \text{on } \Gamma^{\mathrm{per}}\,, \\
\partial_{x_2} v &= \mathcal{T}^{+} v \quad \text{on } \Gamma_H\,.
\end{aligned} \tag{4.32}$$

By applying the FB transform to (4.32), we obtain an equivalent formulation for the function $w_g \coloneqq \mathcal{J}v \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ satisfying

$$\int_\Lambda a_\alpha(w_g(\alpha), \varphi(\alpha))\, \mathrm{d}\alpha = \int_\Lambda \left\langle \mathcal{J}g(\alpha), \overline{\varphi(\alpha)} \right\rangle_{\Omega_H^{\mathrm{per}}} \mathrm{d}\alpha\,, \tag{4.33}$$

for all $\varphi \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, where the sesquilinear form a_α is defined as in (4.11). We can simply formulate the problem corresponding to the PML approximation of v by replacing the boundary condition on Γ_H by a condition with the PML approximation of the DtN operator $\mathcal{T}^+_{\alpha,\sigma}$. That means, the PML approximation of v, denoted by $v_\sigma \in \widetilde{H}^1(\Omega_H^{\mathrm{per}})$, satisfies

$$\begin{aligned} \Delta v_\sigma + k^2 v_\sigma &= g && \text{in } \Omega_H^{\mathrm{per}}\,, \\ v_\sigma &= 0 && \text{on } \Gamma^{\mathrm{per}}\,, \\ \partial_{x_2} v_\sigma &= \mathcal{T}^+_\sigma v_\sigma && \text{on } \Gamma_H\,. \end{aligned} \tag{4.34}$$

The corresponding variational formulation for $w_{g,\sigma} \coloneqq \mathcal{J}v_\sigma \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ is

$$\int_\Lambda a_{\alpha,\sigma}(w_{g,\sigma}(\alpha), \varphi(\alpha))\, \mathrm{d}\alpha = \int_\Lambda \left\langle \mathcal{J}g(\alpha), \overline{\varphi(\alpha)} \right\rangle_{\Omega_H^{2\pi}} \mathrm{d}\alpha\,, \tag{4.35}$$

for all $\varphi \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, where the sesquilinear form $a_{\alpha,\sigma}$ is obtained from a_α in (4.11) by replacing $\mathcal{T}^+_\alpha$ with $\mathcal{T}^+_{\alpha,\sigma}$ given by (4.17).

It has been shown in [26] (outlined in Remark 2.23) that the PML approximation v_σ to the exact solution v of such a source problem cannot be expected to converge exponentially on the unbounded domain Ω_H^{per}. However, in [105] it was established that, on compact subsets of the periodic domain Ω_H^{per}, the exponential convergence of the PML is achievable for $g \in L^2(\Omega_H^{\mathrm{per}})$.

Let us first recall some important results from [105] for the source problem (4.32), presented in the following two lemmas.

Lemma 4.14. *Let the bounded linear operators $\mathcal{A}_\alpha$, $\mathcal{A}_{\alpha,\sigma} \colon \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}) \to (\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))^*$ be induced by the sesquilinear forms a_α and $a_{\alpha,\sigma}$, respectively. Then, for any sufficiently large σ, there exist constants $c, C > 0$ independent of α and σ such that*

$$\|\mathcal{A}_\alpha - \mathcal{A}_{\alpha,\sigma}\| \leq C\mathrm{e}^{-c|\sigma|} \quad \text{for all } \alpha \in \mathcal{E}\,,$$

where $\mathcal{E}$ is defined in (4.29).

Proof. See [105, Thm. 9]. □

From this lemma, it is concluded that for any sufficiently large σ there exist well-defined bounded solution operators $\mathcal{R}$ and $\mathcal{R}_\sigma \colon (\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))^* \to C(\mathcal{E}, \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ corresponding to (4.33) and (4.35) such that

$$w_g = \mathcal{R}g\,, \quad w_{g,\sigma} = \mathcal{R}_\sigma g \quad \text{and} \quad \|\mathcal{R} - \mathcal{R}_\sigma\| \leq C\,\mathrm{e}^{-c|\sigma|}\,. \tag{4.36}$$

Lemma 4.15. *Let v and v_σ be the solution of (4.32) and (4.34), respectively. Then, for every compact subset $K \subseteq \overline{\Omega_H^{\mathrm{per}}}$ and sufficiently large σ, there holds*

$$\|v - v_\sigma\|_{H^1(K)} \leq C\mathrm{e}^{-c|\sigma|}\,,$$

where c, C are constants independent of σ.

Proof. See [105, Thm. 11]. □

In the next theorem, we extend these results to cover the approximation of the solution of (4.1) by the solution of (4.14) in the periodic domain Ω_H^{per} (i.e., $\delta = 0$ in these problems). For simplicity, we omit δ in the periodic setting, rather than indicating it with a superscript zero.

Theorem 4.16. *Consider a point source $y \in \Omega_H^{\mathrm{per}}$, with y' its reflection with respect to $\mathbb{R} \times \{0\}$. Let $u^s \in H_r^1(\Omega_H^{\mathrm{per}})$ denote a weak solution to (4.1) for the incident field*

$$u^i(x) = \Phi(x, y) - \Phi(x, y')\,, \qquad x \in \Omega_H^{\mathrm{per}}, \quad x \neq y\,.$$

Moreover, let u_σ^s denote the weak solution of the PML problem (4.14) in the periodic domain Ω_H^{per}. Then, for every compact subset $K \subseteq \overline{\Omega_H^{\mathrm{per}}}$ and sufficiently large σ, there exist constants $c, C > 0$ such that

$$\|u^s - u_\sigma^s\|_{H^1(K)} \leq C\,\mathrm{e}^{-c|\sigma|}\,.$$

Proof. We begin the proof by considering scattering problems, where waves are scattered by the flat surface $\Gamma_0 \coloneqq \mathbb{R} \times \{0\}$. Let Ω_0^H be the unbounded domain between Γ_0 and Γ_H. According to [23, Eq. (2.8)], there is a compactly supported function $g \in L^2(\Omega_H^{\mathrm{per}})$, such that the weak solution $v^i \in H_r^1(\Omega_0^H)$ to

$$\begin{aligned} \Delta v^i + k^2 v^i &= g && \text{in } \Omega_0^H\,, \\ v^i &= 0 && \text{on } \Gamma_0\,, \\ \partial_{x_2} v^i &= \mathcal{T}^+ v^i && \text{on } \Gamma_H\,, \end{aligned} \tag{4.37}$$

is equal to u^i in $\Omega_H^{\mathrm{per}} \setminus \mathrm{supp}\,(g)$. Therefore, in Problem (4.1) corresponding to the periodic case, we may replace u^i by v^i in the boundary condition on Γ^{per}.

On the other hand, we denote by v_σ^i the PML approximation of v^i, i.e., the solution to (4.37) with $\mathcal{T}^+$ replaced by $\mathcal{T}_\sigma^+$. Let v_σ^s denote the solution to (4.14) corresponding to the periodic case with u^i replaced by v_σ^i. Now, we can estimate

$$\begin{aligned} \|u^s - u_\sigma^s\|_{H^1(K)} &= \left\|u^s + v^i - v^i + v_\sigma^i - v_\sigma^i + v_\sigma^s - v_\sigma^s - u_\sigma^s\right\|_{H^1(K)} \\ &\leq \left\|u^s + v^i - (v_\sigma^s + v_\sigma^i)\right\|_{H^1(K)} + \left\|v_\sigma^i - v^i\right\|_{H^1(K)} + \|v_\sigma^s - u_\sigma^s\|_{H^1(K)}\,. \end{aligned} \tag{4.38}$$

We start by estimating the first term of expression (4.38). The function $u^s + v^i$ is the solution of the source problem (4.32), while $v_\sigma^s + v_\sigma^i$ is its PML approximation, i.e., the solution to (4.34). Hence, due to Lemma 4.15, there exist two constants c, C such that

$$\left\|u^s + v^i - (v_\sigma^s + v_\sigma^i)\right\|_{H^1(K)} \leq C\,\mathrm{e}^{-c|\sigma|}\,,$$

for every compact subset $K \subseteq \overline{\Omega_H^{\mathrm{per}}}$.

The function $q := v_\sigma^i - v^i$ in the second term of (4.38) satisfies the Helmholtz equation in Ω_H^{per} with the homogeneous boundary condition on Γ^{per} and

$$\partial_{x_2} q = \mathcal{T}_\sigma^+ q + (\mathcal{T}^+ - \mathcal{T}_\sigma^+) v^i \quad \text{on } \Gamma_H \,.$$

By using the FB transform, we obtain that q satisfies (4.35), but with the different right-hand side. Therefore, similar to the first part, we can write

$$\|q\|_{H^1(K)} \le C \|\mathcal{R} - \mathcal{R}_\sigma\| \left\| v^i \right\|_{H^1(K)} \le C \mathrm{e}^{-c|\sigma|} \,,$$

where the last inequality follows from (4.36). Now, it remains only to obtain an error bound for the last term of (4.38), i.e., $v_\sigma^s - u_\sigma^s$. Setting $z := v_\sigma^s - u_\sigma^s$, we see that this function is the weak solution to

$$\begin{aligned} \Delta z + k^2 z &= 0 && \text{in } \Omega_H^{\mathrm{per}} \,, \\ z &= u^i - v_\sigma^i && \text{on } \Gamma^{\mathrm{per}} \,, \\ \partial_{x_2} z &= \mathcal{T}_\sigma^+ z && \text{on } \Gamma_H \,. \end{aligned}$$

To analyze $\|z\|_{H^1(K)}$, we apply the FB transform to the problem above. Using Theorem 4.13 and the continuity of the solution with respect to the boundary data leads to

$$\|z\|_{H^1(K)} \le C_1 \|\mathcal{J} z\|_{C(\mathcal{E}; H^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \le C_2 \left\| \mathcal{J}(u^i - v_\sigma^i) \right\|_{C(\mathcal{E}; H^{1/2}_{\mathrm{per}}(\Gamma^{2\pi}))} \,,$$

for some constants C_1 and C_2. From Lemma C.3, it is known that

$$\mathcal{J} u^i(\alpha; x) = \begin{cases} \dfrac{1}{2\pi} \displaystyle\sum_{j \in \mathbb{Z}} \mathrm{e}^{\mathrm{i} j (x_1 - y_1) + \mathrm{i} \gamma_j(\alpha) y_2} \operatorname{sinc}(\gamma_j(\alpha)\, x_2)\, x_2 \,, & 0 < x_2 \le y_2 \,, \\ \dfrac{1}{2\pi} \displaystyle\sum_{j \in \mathbb{Z}} \mathrm{e}^{\mathrm{i} j (x_1 - y_1) + \mathrm{i} \gamma_j(\alpha) x_2} \operatorname{sinc}(\gamma_j(\alpha)\, y_2)\, y_2 \,, & x_2 > y_2 \,. \end{cases}$$

Likewise, applying the FB transform to [26, Eq. (28)] yields

$$\mathcal{J} v_\sigma^i(\alpha; x) = \begin{cases} \dfrac{1}{2\pi} \displaystyle\sum_{j \in \mathbb{Z}} \mathrm{e}^{\mathrm{i} j (x_1 - y_1)} \dfrac{\sin(\gamma_j(\alpha)\,(\sigma + H - y_2))}{\sin(\gamma_j(\alpha)\,(\sigma + H))} \operatorname{sinc}(\gamma_j(\alpha)\, x_2)\, x_2 \,, & 0 < x_2 \le y_2 \,, \\ \dfrac{1}{2\pi} \displaystyle\sum_{j \in \mathbb{Z}} \mathrm{e}^{\mathrm{i} j (x_1 - y_1)} \dfrac{\sin(\gamma_j(\alpha)\,(\sigma + H - x_2))}{\sin(\gamma_j(\alpha)\,(\sigma + H))} \operatorname{sinc}(\gamma_j(\alpha)\, y_2)\, y_2 \,, & x_2 > y_2 \,. \end{cases}$$

We consider the case $0 < x_2 \le y_2$. Clearly, we can write

$$\mathcal{J}(u^i - v_\sigma^i)(\alpha; x) = \frac{1}{2\pi} \sum_{j \in \mathbb{Z}} \mathrm{e}^{\mathrm{i} j (x_1 - y_1)} \left[\mathrm{e}^{\mathrm{i} \gamma_j(\alpha) y_2} - \frac{\sin\left(\gamma_j(\alpha)(\sigma + H - y_2)\right)}{\sin\left(\gamma_j(\alpha)(\sigma + H)\right)} \right] \operatorname{sinc}(\gamma_j(\alpha) x_2) x_2 \,.$$

Using Euler's formula, it is straightforward to derive the general identity

$$\mathrm{e}^{\mathrm{i} A} - \frac{\sin(B - A)}{\sin(B)} = \mathrm{i} \sin(A)\,(1 - \coth(-\mathrm{i} B)) \,, \quad A \in \mathbb{C}, B \in \mathbb{C} \setminus \pi\mathbb{Z} \,.$$

Applying the identity with $A = \gamma_j(\alpha)\, y_2$, $B = \gamma_j(\alpha)\,(\sigma + H)$, we obtain

$$\mathcal{J}(u^i - v^i_\sigma)(\alpha; x) = \frac{\mathrm{i}}{2\pi} \sum_{j\in\mathbb{Z}} \big(\gamma_j(\alpha)\,[1 - \coth(-\mathrm{i}\gamma_j(\alpha)(\sigma + H))] \\ \times \mathrm{e}^{\mathrm{i}j(x_1-y_1)}\,\mathrm{sinc}(\gamma_j(\alpha)\,x_2)\,\mathrm{sinc}(\gamma_j(\alpha)\,y_2)\,x_2\,y_2\big)\,.$$

As this expression is symmetric with respect to x_2 and y_2, it also holds in the case of $x_2 > y_2$.

Note that the sinc functions depend analytically on α. As the Floquet parameter α takes values in a bounded domain, the term $\mathrm{e}^{\mathrm{i}j(x_1-y_1)}\,\mathrm{sinc}(\gamma_j(\alpha)\,x_2)\,\mathrm{sinc}(\gamma_j(\alpha)\,y_2)\,x_2\,y_2$ is uniformly bounded with respect to α. We hence need to analyze the term containing the singularities with respect to α, that is,

$$\gamma_j(\alpha)\;[1 - \coth(-\mathrm{i}\gamma_j(\alpha)(\sigma + H))]\,.$$

As established in [105, Lem. 8, 71, Lem. 18], we know that there exist constants $\widetilde{c}, C$ such that

$$|1 - \coth(-\mathrm{i}\gamma_j(\alpha)(\sigma + H))| = \frac{2}{|\exp(-2\mathrm{i}\gamma_j(\alpha)\sigma) - 1|} \le C \exp\left(-\widetilde{c}|\sigma|\sqrt{|\mathrm{Re}\,(\alpha) + j + k|}\right).$$

Therefore, setting $c \coloneqq \widetilde{c}\sqrt{k}$, we conclude for sufficiently large σ

$$\left\|\mathcal{J}(u^i - v^i_\sigma)(\alpha)\right\|_{H^{1/2}_{\mathrm{per}}(\Gamma^{2\pi})} \le C\,\mathrm{e}^{-c\,|\sigma|} \quad \text{for all } \alpha \in \mathcal{E}\,. \tag{4.39}$$

Summing up the estimates for all three terms gives the asserted result. □

4.4.3. The Perturbed Case

Now, we turn to the general locally perturbed periodic case. Again, we first consider the corresponding source problem: For a given compactly supported source $g \in L^2(\Omega^\delta_H)$, find the weak solution $v^\delta \in \widetilde{H}^1(\Omega^\delta_H)$ of

$$\begin{aligned} \Delta v^\delta + k^2 v^\delta &= g && \text{in } \Omega^\delta_H\,, \\ v^\delta &= 0 && \text{on } \Gamma^\delta\,, \\ \partial_{x_2} v^\delta &= \mathcal{T}^+ v^\delta && \text{on } \Gamma_H\,. \end{aligned} \tag{4.40}$$

The PML approximation v^δ_σ is obtained by solving the same problem, but replacing the DtN map $\mathcal{T}^+$ with its PML approximation $\mathcal{T}^+_\sigma$ in the boundary condition on Γ_H, i.e.,

$$\begin{aligned} \Delta v^\delta_\sigma + k^2 v^\delta_\sigma &= g && \text{in } \Omega^\delta_H\,, \\ v^\delta_\sigma &= 0 && \text{on } \Gamma^\delta\,, \\ \partial_{x_2} v^\delta_\sigma &= \mathcal{T}^+_\sigma v^\delta_\sigma && \text{on } \Gamma_H\,. \end{aligned} \tag{4.41}$$

As these problems are a special case of the rough scattering problem, we know from [23, Thm. 4.1, 26, Sec. 3] that both problems are uniquely solvable and that $\|v^\delta - v^\delta_\sigma\|_{H^1(\Omega^\delta_H)} \to 0$ as $|\sigma| \to \infty$. Furthermore, it has been established that the convergence cannot be expected to be faster than linear with respect to σ in the unbounded domain Ω^δ_H.

As described in Section 4.1, we use the diffeomorphism Ψ^δ to reformulate these problems in the periodic domain Ω_H^{per}. This allows us to apply the FB transform and obtain that v^δ is a solution of (4.40) if and only if $w_g^\delta \coloneqq \mathcal{J}(v^\delta \circ \Psi^\delta) \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ satisfies

$$\int_\Lambda \Big(a_\alpha(w_g^\delta(\alpha), \varphi(\alpha)) + b_\alpha^\delta(w_g^\delta, \varphi(\alpha)) \Big) \,\mathrm{d}\alpha = \int_\Lambda \Big\langle \mathcal{J}(g \circ \Psi^\delta)(\alpha), \overline{\varphi(\alpha)} \Big\rangle_{\Omega_H^{2\pi}} \,\mathrm{d}\alpha \,,$$

for all $\varphi \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$. Similarly, the PML approximation v_σ^δ is a solution of (4.41) if and only if $w_{g,\sigma}^\delta \coloneqq \mathcal{J}(v_\sigma^\delta \circ \Psi^\delta) \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ satisfies

$$\int_\Lambda \Big(a_{\alpha,\sigma}(w_{g,\sigma}^\delta(\alpha), \varphi(\alpha)) + b_\alpha^\delta(w_{g,\sigma}^\delta, \varphi(\alpha)) \Big) \,\mathrm{d}\alpha = \int_\Lambda \Big\langle \mathcal{J}(g \circ \Psi^\delta)(\alpha), \overline{\varphi(\alpha)} \Big\rangle_{\Omega_H^{2\pi}} \,\mathrm{d}\alpha \,,$$

for all $\varphi \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$.

As explained in Section 4.4.1, we know that w_g^δ and $w_{g,\sigma}^\delta$ analytically depend on α except for possible branch cuts. We may hence change the integration over Λ to the integration over $\mathcal{E}$. Note that the bounded linear operator $\mathcal{B}_\alpha^\delta \colon L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \to (\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))^*$ induced by the sesquilinear form b_α^δ also depends analytically on α. Therefore, it is also well defined for $\alpha \in \mathcal{E}$.

Now using the solution operators for the periodic domain, we may write

$$w_g^\delta + \mathcal{R}\,\mathcal{B}_\alpha^\delta w_g^\delta = \mathcal{R} g \,, \quad \text{and} \quad w_{g,\sigma}^\delta + \mathcal{R}_\sigma\,\mathcal{B}_\alpha^\delta w_{g,\sigma}^\delta = \mathcal{R}_\sigma g \,, \tag{4.42}$$

where $\mathcal{R}$ and $\mathcal{R}_\sigma$ are given in (4.36). As both problems are equivalent to the corresponding source problems, we know that the operators on the left-hand side are boundedly invertible.

Theorem 4.17. *Let v^δ and v_σ^δ be the solutions of (4.40) and (4.41) for $g \in L^2(\Omega_H^\delta)$. Then, for every compact subset $K \subseteq \overline{\Omega_H^\delta}$ and sufficiently large σ, there exist some constants c, C such that*

$$\left\| v^\delta - v_\sigma^\delta \right\|_{H^1(K)} \leq C \mathrm{e}^{-c|\sigma|} \,.$$

Proof. We recall the diffeomorphism Ψ^δ, which maps Ω_H^{per} to Ω_H^δ. Therefore, every compact subset $K \subseteq \overline{\Omega_H^\delta}$ can be transformed into the compact subset $K_{\mathrm{tra}} \coloneqq (\Psi^\delta)^{-1}(K) \subseteq \overline{\Omega_H^{\mathrm{per}}}$. By using Theorem 4.13, for some generic constant C we conclude

$$\left\| v^\delta - v_\sigma^\delta \right\|_{H^1(K)} \leq C \left\| v^\delta \circ \Psi^\delta - v_\sigma^\delta \circ \Psi^\delta \right\|_{H^1(K_{\mathrm{tra}})} \leq C \left\| w_g^\delta - w_{g,\sigma}^\delta \right\|_{C(\mathcal{E}, \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} .$$

Using the perturbation theorem (see [75, Thm. 10.1]) for the problems in (4.42), we see that for any sufficiently large σ

$$\begin{aligned} \left\| w_g^\delta - w_{g,\sigma}^\delta \right\|_{C(\mathcal{E}, \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} &\leq C \left(\left\| \mathcal{R}\mathcal{B}_\alpha^\delta - \mathcal{R}_\sigma \mathcal{B}_\alpha^\delta \right\| \left\| w_g^\delta \right\|_{C(\mathcal{E}, \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} + \|\mathcal{R} - \mathcal{R}_\sigma\| \, \|g\|_{L^2(\Omega_H^{2\pi})} \right) \\ &\leq C \|\mathcal{R} - \mathcal{R}_\sigma\| \left(\left\| w_g^\delta \right\|_{C(\mathcal{E}, \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} + \|g\|_{L^2(\Omega_H^{2\pi})} \right) . \end{aligned}$$

From (4.42), we get $w_g^\delta = (I + \mathcal{R}\mathcal{B}_\alpha^\delta)^{-1}\mathcal{R}g$. Hence, we obtain

$$\left\| w_g^\delta - w_{g,\sigma}^\delta \right\|_{C(\mathcal{E},\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \leq C\|\mathcal{R} - \mathcal{R}_\sigma\| \|g\|_{L^2(\Omega_H^{2\pi})} .$$

Using (4.36), we finally conclude that for any sufficiently large σ

$$\left\| w_g^\delta - w_{g,\sigma}^\delta \right\|_{C(\mathcal{E},\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \leq C\mathrm{e}^{-c|\sigma|} ,$$

where C depends on the L^2-norm of g and $\|\mathcal{B}_\alpha^\delta\|$. This completes the proof. □

In the next theorem, we extend these results to the non-periodic scattering problems.

Theorem 4.18. *Consider a point source $y \in \Omega_H^{\mathrm{per}}$, with y' its reflection with respect to $\mathbb{R} \times \{0\}$. Let $u^{s,\delta} \in H^1_r(\Omega_H^\delta)$ denote a weak solution to (4.1) for*

$$u^i(x) = \Phi(x,y) - \Phi(x,y') , \quad x \in \Omega_H^\delta, \quad x \neq y .$$

Moreover, let $u_\sigma^{s,\delta}$ denote the weak solution of the PML approximation (4.14). Then, for every compact subset $K \subseteq \overline{\Omega_H^\delta}$ and sufficiently large σ, there exist C, $c > 0$ such that

$$\left\| u^{s,\delta} - u_\sigma^{s,\delta} \right\|_{H^1(K)} \leq C\,\mathrm{e}^{-c|\sigma|} .$$

Proof. Let $g \in L^2(\Omega_H^\delta)$ be compactly supported, v^i be the solution of (4.37) with the right-hand side g and v_σ^i its PML approximation. Further, we denote by $v_\sigma^{s,\delta}$ the solution of the scattering problem (4.14) with u^i replaced by v_σ^i. Adding and subtracting these functions yields

$$\left\| u^{s,\delta} - u_\sigma^{s,\delta} \right\|_{H^1(K)} \leq \left\| u^{s,\delta} + v^i - (v_\sigma^{s,\delta} + v_\sigma^i) \right\|_{H^1(K)} + \left\| v^i - v_\sigma^i \right\|_{H^1(K)} + \left\| v_\sigma^{s,\delta} - u_\sigma^{s,\delta} \right\|_{H^1(K)} .$$

Note that $u^{s,\delta} + v^i$ satisfies the source problem (4.40) and $v_\sigma^{s,\delta} + v_\sigma^i$ is its PML approximation, i.e., the solution of (4.41). Therefore, from Theorem 4.17, it follows that there exist constants c, C such that

$$\left\| u^{s,\delta} + v^i - (v_\sigma^{s,\delta} + v_\sigma^i) \right\|_{H^1(K)} \leq C\mathrm{e}^{-c|\sigma|} .$$

By the same arguments as in Theorem 4.16, the corresponding estimate holds for $v^i - v_\sigma^i$. Now, it only remains to estimate $\|v_\sigma^{s,\delta} - u_\sigma^{s,\delta}\|_{H^1(K)}$. As in the proof of Theorem 4.16, we consider $z \coloneqq v_\sigma^{s,\delta} - u_\sigma^{s,\delta}$, which is a weak solution to

$$\begin{aligned} \Delta z + k^2 z &= 0 && \text{in} \quad \Omega_H^\delta , \\ z &= u^i - v_\sigma^i && \text{on} \quad \Gamma^\delta , \\ \partial_{x_2} z &= \mathcal{T}_\sigma^+ z && \text{on} \quad \Gamma_H . \end{aligned}$$

Recalling the diffeomorphism Ψ^δ from (4.4) and the compact subset $K_{\mathrm{tra}} \subseteq \overline{\Omega_H^{\mathrm{per}}}$ (defined as in

the proof of Theorem 4.17) and using Theorem 4.13, we can write

$$\begin{aligned}\|z\|_{H^1(K)} \le C\left\|z\circ\Psi^\delta\right\|_{H^1(K_{\mathrm{tra}})} &\le C\left\|\mathcal{J}(z\circ\Psi^\delta)\right\|_{C(\mathcal{E};\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\\ &\le C\left\|\mathcal{J}(u^i\circ\Psi^\delta - v^i_\sigma\circ\Psi^\delta)\right\|_{C(\mathcal{E};H^{1/2}_{\mathrm{per}}(\Gamma^{2\pi}))}.\end{aligned}$$

Note that $\Psi^\delta = I$ in $\Omega_H^{\mathrm{per}}\setminus\Omega_H^{2\pi}$, thus for every $f\in H^{1/2}_{\mathrm{per}}(\Gamma^{\mathrm{per}}\cup\Gamma^\delta)$ we have

$$\mathcal{J}(f\circ\Psi^\delta)(\alpha;x) = \mathcal{J}f(\alpha;x) + \mathrm{e}^{-\mathrm{i}\alpha x_1}((f\circ\Psi^\delta)(x) - f(x))\,,\quad \alpha\in\mathcal{E}\,,\ x\in\Gamma^{2\pi}\,.$$

Considering $f := u^i - v^i_\sigma$ in the relation above and using the estimate (4.39), we obtain

$$\begin{aligned}\left\|\mathcal{J}(u^i\circ\Psi^\delta - v^i_\sigma\circ\Psi^\delta)\right\|_{C(\mathcal{E};H^{1/2}_{\mathrm{per}}(\Gamma^{2\pi}))} &\le C\mathrm{e}^{-c|\sigma|}\\ &\quad + \left\|(u^i - v^i_\sigma)\circ\Psi^\delta\right\|_{H^{1/2}_{\mathrm{per}}(\Gamma^{2\pi})} + \left\|u^i - v^i_\sigma\right\|_{H^{1/2}_{\mathrm{per}}(\Gamma^{2\pi})}.\end{aligned}$$

The remaining two terms are estimated as in the proof of Theorem 4.16, which gives the asserted result. □

4.5. Full Discretization of the PML Problem

To represent the connection between the total field $u^\delta_{\mathrm{tra},\sigma}$ and the transformed total field w^δ_σ, we need to discretize the inverse FB transform (2.35). Since the transformed field w^δ_σ is analytic with respect to the Floquet parameter α when $\mathcal{J}u^i$ is analytic with respect to α (see Theorem 4.9), we can use a Gauss quadrature formula: For N_α quadrature points $\alpha_j\in\Lambda$ and weights μ_j, $j\in\{1,\dots,N_\alpha\}$, we have the approximation

$$u^\delta_{\mathrm{tra},\sigma}(x+2\pi\ell) = \int_\Lambda w^\delta_\sigma(\alpha;x)\mathrm{e}^{\mathrm{i}\alpha\cdot(\widetilde{x}+2\pi\ell)}\,\mathrm{d}\alpha \approx \sum_{j=1}^{N_\alpha}\mu_j w^\delta_\sigma(\alpha_j;x)\mathrm{e}^{\mathrm{i}\alpha_j\cdot(\widetilde{x}+2\pi\ell)}\,, \tag{4.43}$$

for any $\ell\in\mathbb{Z}^{d-1}$ and $x\in\Omega_H^{2\pi}$.

To approximate the total field $u^\delta_{\mathrm{tra},\sigma}$, it is required to solve (4.18) only for the quadrature points α_j. Substituting the above representation for $\ell=0$ into (4.18), we end up with a system involving the quantities $w^\delta_\sigma(\alpha_j;x)$, which we need to discretize with respect to the spatial variable x. We therefore use the finite element method (FEM) and for ease of presentation, we restrict ourselves to the finite element functions of piecewise linear polynomials. We generate a triangular mesh on the domain $\Omega_H^{2\pi}$ supporting a family $\{\phi_n\}_{n=1}^{N_\Delta}$ of $N_\Delta\in\mathbb{N}$ such basis functions and then approximate the transformed total field

$$w^\delta_\sigma(\alpha_j)\approx\sum_{n=1}^{N_\Delta}W_{j,n}\,\phi_n\,,$$

for each quadrature point α_j as well as the total field

$$u^{\delta}_{\mathrm{tra},\sigma} \approx \sum_{n=1}^{N_\Delta} U_n \phi_n \,.$$

The relation between U_n and $W_{j,n}$ is obtained by (4.43), which yields

$$U_n = \sum_{j=1}^{N_\alpha} \mu_j \, \mathrm{e}^{\mathrm{i}\alpha_j \cdot \widetilde{x}_n} W_{j,n} \quad \text{for all } n \in \{1,\dots,N_\Delta\}\,. \tag{4.44}$$

Now, we can formulate the Galerkin approximation of (4.18) for each quadrature point α_j, for $j \in \{1,\dots,N_\alpha\}$, as

$$\begin{aligned} &\sum_{n=1}^{N_\Delta} W_{j,n} a_{\alpha_j}(\phi_n,\phi_m) + \sum_{n=1}^{N_\Delta} U_n \, b^{\delta}_{\alpha_j}(\phi_n,\phi_m) \\ &\qquad = \left\langle (\partial_{x_d} - \mathcal{T}^{+}_{\alpha,\sigma}) \mathcal{J} u^i(\alpha_j), \overline{\phi_m} \right\rangle_{\Gamma^{2\pi}_H} \quad \text{for all } m \in \{1,\dots,N_\Delta\}\,. \end{aligned} \tag{4.45}$$

Defining the vectors of unknowns

$$\mathbf{W}_j \coloneqq \begin{bmatrix} W_{j,1} & \cdots & W_{j,N_\Delta} \end{bmatrix}^\top \quad \text{and} \quad \mathbf{U} \coloneqq \begin{bmatrix} U_1 & \cdots & U_{N_\Delta} \end{bmatrix}^\top,$$

we can rewrite system (4.44)-(4.45) in a block vector-matrix form as

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{N_\alpha} & \mathbf{B}_{N_\alpha} \\ \mathbf{C}_1 & \cdots & \cdots & \mathbf{C}_{N_\alpha} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \vdots \\ \mathbf{W}_{N_\alpha} \\ \mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_1 \\ \vdots \\ \vdots \\ \mathbf{F}_{N_\alpha} \\ \mathbf{0} \end{bmatrix} \in \mathbb{C}^{(N_\alpha+1)N_\Delta}\,, \tag{4.46}$$

where the block matrices $\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j \in \mathbb{C}^{N_\Delta \times N_\Delta}$ and the vector $\mathbf{F}_j \in \mathbb{C}^{N_\Delta}$ are defined by

$$\begin{aligned} (\mathbf{A}_j)_{m,n} &\coloneqq a_{\alpha_j}(\phi_n,\phi_m), \\ (\mathbf{B}_j)_{m,n} &\coloneqq -b^{\delta}_{\alpha_j}(\phi_n,\phi_m), \\ (\mathbf{C}_j)_{m,n} &\coloneqq \mu_j \mathrm{e}^{\mathrm{i}\alpha_j \cdot \widetilde{x}_m} \delta_{m,n}, \\ (\mathbf{F}_j)_m &\coloneqq \left\langle (\partial_{x_d} - \mathcal{T}^{+}_{\alpha,\sigma}) \mathcal{J} u^i(\alpha_j; x_m), \overline{\phi_m} \right\rangle_{\Gamma^{2\pi}_H} \end{aligned}$$

for all $m,n \in \{1,\dots,N_\Delta\}$ and $j \in \{1,\dots,N_\alpha\}$.

The coefficient matrix in (4.46) is known as a permuted square arrowhead matrix, which frequently arises in applications. These include modeling of wireless communication systems [78] and radiationless transitions in isolated moleculs [14, 47]. One of the main challenges in these applications is to solve large linear systems in parallel [34, 36, 48, 51, 52, 103]. As a result, the computation of the inverse of these matrices has attracted considerable attention.

In [53], an adaptive approximate inverse method based on an LU-type factorization procedure is proposed for the explicit computation of the inverse of an arrowhead matrix. Additionally, in [91], a modified Sherman–Morrison inverse matrix method is introduced, while [98] applies the Sherman–Morrison–Woodbury formula to facilitate the inversion of block arrowhead matrix. Nevertheless, assembling and inverting the coefficient matrix of (4.46) is still expensive. In contrast, in our arguments below, we are going to propose an alternative method for solving the linear system (4.46) without inverting the coefficient matrix. In Appendix B, we compare the computational cost of the proposed iterative solver in Algorithm 3 with a direct solver introduced in [98, Sec. 2]. These results show that we have significantly reduced the computational time.

In the following theorem, we utilize the recursive Schur complement to obtain an equivalent form of the system (4.46), which can be parallelized more easily.

Theorem 4.19. *Let* $\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j$, *and* $\mathbf{F}_j$ *for all* $j \in \{1, \dots, N_\alpha\}$ *be defined as above. The linear system* (4.46) *is equivalent to*

$$\left(\mathbf{I} - \sum_{j=1}^{N_\alpha} \mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{B}_j\right) \mathbf{U} = -\sum_{j=1}^{N_\alpha} \mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{F}_j. \tag{4.47}$$

This means that if $[\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_{N_\alpha}, \mathbf{U}]^\top$ *solves* (4.46), *then* $\mathbf{U}$ *solves* (4.47). *If* $\mathbf{U}$ *solves* (4.47), *then* $[\mathbf{A}_1^{-1}(\mathbf{F}_1 - \mathbf{B}_1\mathbf{U}), \dots, \mathbf{A}_{N_\alpha}^{-1}(\mathbf{F}_{N_\alpha} - \mathbf{B}_{N_\alpha}\mathbf{U}), \mathbf{U}]^\top$ *solves* (4.46).

Proof. The proof presents an algorithm to reduce (4.46) to (4.47), by recursively applying a procedure that removes one unknown vector $\mathbf{W}_\ell$ (for $\ell \in \{1, \dots, N_\alpha\}$). The assertions follow by induction on the number of removed unknowns.

For the initial step ($\ell = 1$), we rewrite the $(N_\alpha + 1)N_\Delta$ square system (4.46) as follows

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1^{(\mathrm{rem})} \\ \mathbf{C}_1^{(\mathrm{rem})} & \mathbf{D}_1^{(\mathrm{rem})} \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_1^{(\mathrm{rem})} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_1^{(\mathrm{rem})} \end{bmatrix}, \tag{4.48}$$

where the blocks $\mathbf{B}_1^{(\mathrm{rem})} \in \mathbb{C}^{N_\Delta \times (N_\alpha N_\Delta)}$, $\mathbf{C}_1^{(\mathrm{rem})} \in \mathbb{C}^{(N_\alpha N_\Delta) \times N_\Delta}$ and $\mathbf{W}_1^{(\mathrm{rem})}, \mathbf{F}_1^{(\mathrm{rem})} \in \mathbb{C}^{N_\alpha N_\Delta}$ are defined by

$$\begin{aligned} \mathbf{B}_1^{(\mathrm{rem})} &\coloneqq \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_1 \end{bmatrix}, \quad \mathbf{F}_1^{(\mathrm{rem})} \coloneqq \begin{bmatrix} \mathbf{F}_2 & \cdots & \mathbf{F}_{N_\alpha} & \mathbf{0} \end{bmatrix}^\top, \\ \mathbf{C}_1^{(\mathrm{rem})} &\coloneqq \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_1 \end{bmatrix}^\top, \quad \mathbf{W}_1^{(\mathrm{rem})} \coloneqq \begin{bmatrix} \mathbf{W}_2 & \cdots & \mathbf{W}_{N_\alpha} & \mathbf{U} \end{bmatrix}^\top \end{aligned}$$

and the block $\mathbf{D}_1^{(\mathrm{rem})}$ is given by

$$\mathbf{D}_1^{(\mathrm{rem})} \coloneqq \begin{bmatrix} \mathbf{A}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_2 \\ \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{N_\alpha} & \mathbf{B}_{N_\alpha} \\ \mathbf{C}_2 & \cdots & \cdots & \mathbf{C}_{N_\alpha} & \mathbf{I} \end{bmatrix} \in \mathbb{C}^{N_\alpha N_\Delta \times N_\alpha N_\Delta}.$$

From the first equation in (4.48), we obtain $\mathbf{W}_1 = \mathbf{A}_1^{-1}\left(\mathbf{F}_1 - \mathbf{B}_1^{(\mathrm{rem})}\mathbf{W}_1^{(\mathrm{rem})}\right)$. To remove the first unknown, substituting $\mathbf{W}_1$ into the second equation of (4.48) leads to

$$\mathbf{D}_1^{(\mathrm{rem})}\mathbf{W}_1^{(\mathrm{rem})} = \mathbf{F}_1^{(\mathrm{rem})} - \mathbf{C}_1^{(\mathrm{rem})}\mathbf{W}_1 = \mathbf{F}_1^{(\mathrm{rem})} - \mathbf{C}_1^{(\mathrm{rem})}\mathbf{A}_1^{-1}\left(\mathbf{F}_1 - \mathbf{B}_1^{(\mathrm{rem})}\mathbf{W}_1^{(\mathrm{rem})}\right).$$

By straightforward computations, we obtain

$$\mathbf{C}_1^{(\mathrm{rem})}\mathbf{A}_1^{-1} = \begin{bmatrix}\mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_1\mathbf{A}_1^{-1}\end{bmatrix}^\top \in \mathbb{C}^{N_\alpha N_\Delta \times N_\Delta},$$
$$\mathbf{C}_1^{(\mathrm{rem})}\mathbf{A}_1^{-1}\mathbf{B}_1^{(\mathrm{rem})} = \begin{bmatrix}\mathbf{0}_{(N_\alpha-1)N_\Delta\times(N_\alpha-1)N_\Delta} & \mathbf{0}_{(N_\alpha-1)N_\Delta\times N_\Delta} \\ \mathbf{0}_{N_\Delta\times(N_\alpha-1)N_\Delta} & \mathbf{C}_1\mathbf{A}_1^{-1}\mathbf{B}_1\end{bmatrix} \in \mathbb{C}^{N_\alpha N_\Delta \times N_\alpha N_\Delta}.$$

Then, the remaining system can be written as

$$\left(\mathbf{D}_1^{(\mathrm{rem})} - \mathbf{C}_1^{(\mathrm{rem})}\mathbf{A}_1^{-1}\mathbf{B}_1^{(\mathrm{rem})}\right)\mathbf{W}_1^{(\mathrm{rem})} = \mathbf{F}_1^{(\mathrm{rem})} - \mathbf{C}_1^{(\mathrm{rem})}\mathbf{A}_1^{-1}\mathbf{F}_1, \tag{4.49}$$

where the coefficient matrix is given by

$$\mathbf{D}_1^{(\mathrm{rem})} - \mathbf{C}_1^{(\mathrm{rem})}\mathbf{A}_1^{-1}\mathbf{B}_1^{(\mathrm{rem})} = \begin{bmatrix}\mathbf{A}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_2 \\ \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{N_\alpha} & \mathbf{B}_{N_\alpha} \\ \mathbf{C}_2 & \cdots & \cdots & \mathbf{C}_{N_\alpha} & \mathbf{I} - \mathbf{C}_1\mathbf{A}_1^{-1}\mathbf{B}_1\end{bmatrix} \in \mathbb{C}^{N_\alpha N_\Delta \times N_\alpha N_\Delta}$$

and the right-hand side is determined by

$$\mathbf{F}_1^{(\mathrm{rem})} - \mathbf{C}_1^{(\mathrm{rem})}\mathbf{A}_1^{-1}\mathbf{F}_1 = \begin{bmatrix}\mathbf{F}_2 & \cdots & \mathbf{F}_{N_\alpha} & -\mathbf{C}_1\mathbf{A}_1^{-1}\mathbf{F}_1\end{bmatrix}^\top.$$

So far, the first unknown $\mathbf{W}_1$ has been removed in the initial step. Now, we assume that the theorem holds true for $\ell - 1$. To prove that it also holds for ℓ, we need to solve the following linear system

$$\begin{bmatrix}\mathbf{A}_\ell & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_\ell \\ \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{N_\alpha} & \mathbf{B}_{N_\alpha} \\ \mathbf{C}_\ell & \cdots & \cdots & \mathbf{C}_{N_\alpha} & \mathbf{I} - \sum_{j=1}^{\ell-1}\mathbf{C}_j\mathbf{A}_j^{-1}\mathbf{B}_j\end{bmatrix}\begin{bmatrix}\mathbf{W}_\ell \\ \vdots \\ \vdots \\ \mathbf{W}_{N_\alpha} \\ \mathbf{U}\end{bmatrix} = \begin{bmatrix}\mathbf{F}_\ell \\ \vdots \\ \vdots \\ \mathbf{F}_{N_\alpha} \\ -\sum_{j=1}^{\ell-1}\mathbf{C}_j\mathbf{A}_j^{-1}\mathbf{F}_j\end{bmatrix},$$

which can be written as

$$\begin{bmatrix}\mathbf{A}_\ell & \mathbf{B}_\ell^{(\mathrm{rem})} \\ \mathbf{C}_\ell^{(\mathrm{rem})} & \mathbf{D}_\ell^{(\mathrm{rem})}\end{bmatrix}\begin{bmatrix}\mathbf{W}_\ell \\ \mathbf{W}_\ell^{(\mathrm{rem})}\end{bmatrix} = \begin{bmatrix}\mathbf{F}_\ell \\ \mathbf{F}_\ell^{(\mathrm{rem})}\end{bmatrix} \in \mathbb{C}^{(N_\alpha+1-(\ell-1))N_\Delta}, \tag{4.50}$$

where the block matrices $\mathbf{B}_\ell^{(\mathrm{rem})} \in \mathbb{C}^{N_\Delta \times N_\Delta(N_\alpha+1-\ell)}$, $\mathbf{C}_\ell^{(\mathrm{rem})} \in \mathbb{C}^{N_\Delta(N_\alpha+1-\ell)\times N_\Delta}$ and the vectors

$\mathbf{F}_\ell^{(\mathrm{rem})}, \mathbf{W}_\ell^{(\mathrm{rem})} \in \mathbb{C}^{N_\Delta(N_\alpha+1-\ell)}$ are defined by

$$\mathbf{B}_\ell^{(\mathrm{rem})} \coloneqq \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_\ell \end{bmatrix}, \quad \mathbf{F}_\ell^{(\mathrm{rem})} \coloneqq \begin{bmatrix} \mathbf{F}_{\ell+1} & \cdots & \mathbf{F}_{N_\alpha} & -\sum_{j=1}^{\ell-1} \mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{F}_j \end{bmatrix}^\top,$$
$$\mathbf{C}_\ell^{(\mathrm{rem})} \coloneqq \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_\ell \end{bmatrix}^\top, \quad \mathbf{W}_\ell^{(\mathrm{rem})} \coloneqq \begin{bmatrix} \mathbf{W}_{\ell+1} & \cdots & \mathbf{W}_{N_\alpha} & \mathbf{U} \end{bmatrix}^\top$$

and the block $\mathbf{D}_\ell^{(\mathrm{rem})}$ is given by

$$\mathbf{D}_\ell^{(\mathrm{rem})} = \begin{bmatrix} \mathbf{A}_{\ell+1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_{\ell+1} \\ \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{N_\alpha} & \mathbf{B}_{N_\alpha} \\ \mathbf{C}_\ell & \cdots & \cdots & \mathbf{C}_{N_\alpha} & \mathbf{I} - \sum_{j=1}^{\ell-1} \mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{B}_j \end{bmatrix} \in \mathbb{C}^{(N_\alpha+1-\ell)N_\Delta \times (N_\alpha+1-\ell)N_\Delta}.$$

By solving system (4.50), we obtain

$$\mathbf{W}_\ell = \mathbf{A}_\ell^{-1} \left(\mathbf{F}_\ell - \mathbf{B}_\ell^{(\mathrm{rem})} \mathbf{W}_\ell^{(\mathrm{rem})} \right),$$
$$\mathbf{D}_\ell^{(\mathrm{rem})} \mathbf{W}_\ell^{(\mathrm{rem})} = \mathbf{F}_\ell^{(\mathrm{rem})} - \mathbf{C}_\ell^{(\mathrm{rem})} \mathbf{W}_\ell.$$

Substituting $\mathbf{W}_\ell$ into the second equation gives us

$$\left(\mathbf{D}_\ell^{(\mathrm{rem})} - \mathbf{C}_\ell^{(\mathrm{rem})} \mathbf{A}_\ell^{-1} \mathbf{B}_\ell^{(\mathrm{rem})} \right) \mathbf{W}_\ell^{(\mathrm{rem})} = \mathbf{F}_\ell^{(\mathrm{rem})} - \mathbf{C}_\ell^{(\mathrm{rem})} \mathbf{A}_\ell^{-1} \mathbf{F}_\ell,$$

where the coefficient matrix can be written as

$$\mathbf{D}_\ell^{(\mathrm{rem})} - \mathbf{C}_\ell^{(\mathrm{rem})} \mathbf{A}_\ell^{-1} \mathbf{B}_\ell^{(\mathrm{rem})} = \begin{bmatrix} \mathbf{A}_{\ell+1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_{\ell+1} \\ \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{N_\alpha} & \mathbf{B}_{N_\alpha} \\ \mathbf{C}_{\ell+1} & \cdots & \cdots & \mathbf{C}_{N_\alpha} & \mathbf{I} - \sum_{j=1}^{\ell} \mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{B}_j \end{bmatrix},$$

and the right-hand side is

$$\mathbf{F}_\ell^{(\mathrm{rem})} - \mathbf{C}_\ell^{(\mathrm{rem})} \mathbf{A}_\ell^{-1} \mathbf{F}_\ell = \begin{bmatrix} \mathbf{F}_{\ell+1} & \cdots & \mathbf{F}_{N_\alpha} & -\sum_{j=1}^{\ell} \mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{F}_j \end{bmatrix}^\top,$$

since we have $\mathbf{C}_\ell^{(\mathrm{rem})} \mathbf{A}_\ell^{-1} = \left(\mathbf{0}, \ldots, \mathbf{0}, \mathbf{C}_\ell \mathbf{A}_\ell^{-1} \right)^\top$ and

$$\mathbf{C}_\ell^{(\mathrm{rem})} \mathbf{A}_\ell^{-1} \mathbf{B}_\ell^{(\mathrm{rem})} = \begin{bmatrix} \mathbf{0}_{(N_\alpha-\ell)N_\Delta \times (N_\alpha-\ell)N_\Delta} & \mathbf{0}_{(N_\alpha-\ell)N_\Delta \times N_\Delta} \\ \mathbf{0}_{N_\Delta \times (N_\alpha-\ell)N_\Delta} & \mathbf{C}_\ell \mathbf{A}_\ell^{-1} \mathbf{B}_\ell \end{bmatrix}.$$

At step $\ell = N_\alpha + 1$, the N_α-th unknown has been removed and this completes the proof. □

Equation (4.47) can now be solved by an iterative method as described in Algorithm 3. Note that the summands on the right-hand side, i.e., $\mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{F}_j$ and the matrix vector multiplications

Algorithm 3: iterative method for solving (4.47)

Input: number of quadrature nodes N_α, initial guess $\mathbf{U}_0$

Compute the Gauss quadrature nodes and weights (α_j, μ_j) for $j \in \{1, \dots, N_\alpha\}$;
for $j = 1, \dots, N_\alpha$ **do in parallel**
    Construct the matrices $\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j$ and the vector $\mathbf{F}_j$ using FEM;
    Compute the LU decomposition of $\mathbf{A}_j$;
    Solve the system $\mathbf{A}_j \mathbf{RHS}_j = \mathbf{F}_j$ using the above LU decomposition;
    $\mathbf{RHS}_j \leftarrow \mathbf{C}_j \mathbf{RHS}_j$;
$\mathbf{RHS} \leftarrow \sum_{j=1}^{N_\alpha} \mathbf{RHS}_j$;
%To solve the systems on the left-hand side of (4.47), the following function computes the matrix-vector multiplication for each input.
Define the function *LHS*
    Input: the vector $\mathbf{U}$
    for $j = 1, \dots, N_\alpha$ **do in parallel**
        Solve the system $\mathbf{A}_j \mathbf{X}_j = \mathbf{B}_j \mathbf{U}$ using the precomputed LU decomposition of $\mathbf{A}_j$;
        $\mathbf{X}_j \leftarrow \mathbf{C}_j \mathbf{X}_j$;
    return $\mathbf{U} - \sum_{j=1}^{N_\alpha} \mathbf{X}_j$;
Solve the linear system (4.47) by GMRES with tolerance 10^{-5} and inputs $\mathbf{U}_0$ and *LHS*;
return *Numerical solution of* (4.47)

by the summands on the left-hand side, i.e., $\mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{B}_j \mathbf{U}$ are all independent of each other. Hence they can be carried out in parallel.

4.6. Numerical Results

In this section, we aim to illustrate the efficiency and accuracy of the iterative method described in Algorithm 3 for solving non-periodic scattering problems. Our focus lies on the two-dimensional case. The extension of the proposed method to the three-dimensional case is straightforward. However, it is numerically much more costly.

We again select the non-periodic incident field as the upper half-space Dirichlet Green's function

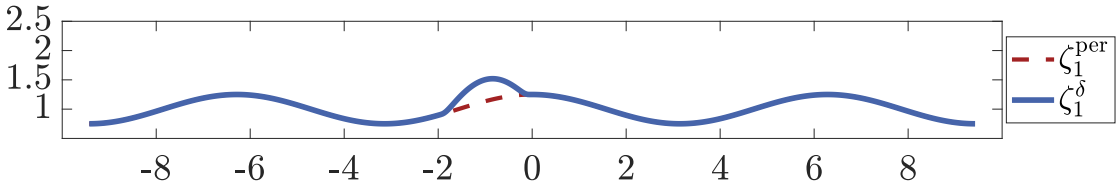
$$u^i(x) = \frac{\mathrm{i}}{4} \left(H_0^{(1)}(k|x - y|) - H_0^{(1)}(k|x - y'|) \right) ,$$

where $y = (y_1, y_2)^\top$ is a fixed point source and $y' = (y_1, -y_2)^\top$ is its reflection with respect to $\{x \in \mathbb{R}^2 : x_2 = 0\}$.

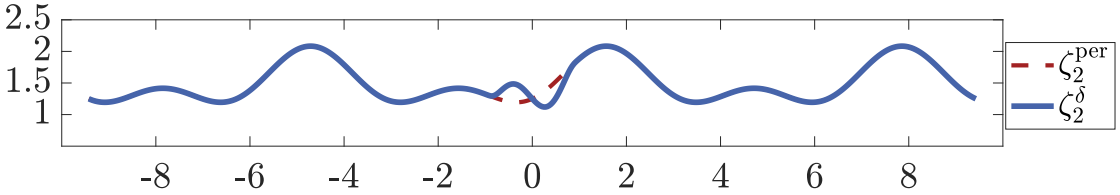
We apply our proposed method to compute the scattered field produced by the locally perturbed scatterers described in the following two examples.

Example 1. We consider the periodic function

$$\zeta_1^{\mathrm{per}}(x) = 1 + \frac{\cos(x)}{4}, \quad x \in \mathbb{R},$$



(a) Surface Γ_1^δ graph of $\zeta_1^\delta := \zeta_1^{\mathrm{per}} + \delta_1$



(b) Surface Γ_2^δ graph of $\zeta_2^\delta := \zeta_2^{\mathrm{per}} + \delta_2$

FIGURE 4.3. Illustration of the locally perturbed curves.

with the perturbation

$$\delta_1(x) = \frac{1}{2}\exp\left(\frac{1}{x(x+2)}\right)\left(\cos\left(\frac{\pi(x+2)}{2}\right)+1\right)\chi_{[-2,0]}(x)\,,$$

where $\chi_{[a,b]}(x) = 1$ for $x \in [a,b]$ and $\chi_{[a,b]}(x) = 0$ for $x \in \mathbb{R}\backslash[a,b]$. The resulting locally perturbed surface $\Gamma_1^\delta = \{(x, \zeta_1^{\mathrm{per}}(x) + \delta_1(x)) : x \in \mathbb{R}\}$ is plotted in Figure 4.3(a).

Example 2. We consider the locally perturbed curve $\Gamma_2^\delta = \{(x, \zeta_2^{\mathrm{per}}(x) + \delta_2(x)) : x \in \mathbb{R}\}$, plotted in Figure 4.3(b), with the periodic function

$$\zeta_2^{\mathrm{per}}(x) = 1.5 + \frac{\sin(x)}{3} - \frac{\cos(2x)}{4}, \quad x \in \mathbb{R}$$

and the perturbation

$$\delta_2(x) = \exp\left(\frac{1}{x^2-1}\right)\sin(\pi(x+1))\chi_{[-1,1]}(x)\,.$$

To calculate the error explicitly, we consider the point source y between the flat surface $\mathbb{R}\times\{0\}$ and the locally perturbed scatterers, since in this case the total field vanishes inside Ω_H^{per}. That means the exact solution is equal to minus the incident field.

In the first example, the point source is located at $y = (-2, 0.2)^\top$ below the surface Γ_1^δ, while in the second, it is positioned at $y = (0, 0.5)^\top$ below the surface Γ_2^δ. In both examples, we choose $H = 2.5$, set the PML thickness to $\lambda = 1.5$ and consider the PML function (2.21) depending only on a positive parameter ρ. It is important to mention that we illustrate the numerical scattered field only in the region below the PML, as the solution inside the PML is not related to the actual scattered field.

To approximate the scattered field in the main bounded cell $\Omega_H^{2\pi} = \{x \in \Omega_H^\delta : x_1 \in (-\pi, \pi)\}$ numerically, we use the iterative solver introduced in Algorithm 3 by setting $N_\alpha = 20$ and $\mathbf{U}_0 = 0$.

Once we have the numerical solution for the main cell $\Omega_H^{2\pi}$, we can extend it to the neighboring cells $\Omega_H^{2\pi+\ell} := \{x \in \Omega_H^\delta : x_1 \in (-\pi, \pi) + 2\pi\ell\}$, for $\ell = \pm 1$. This extension is obtained by using the discrete inverse FB transform defined in (4.43), for $\ell = \pm 1$.

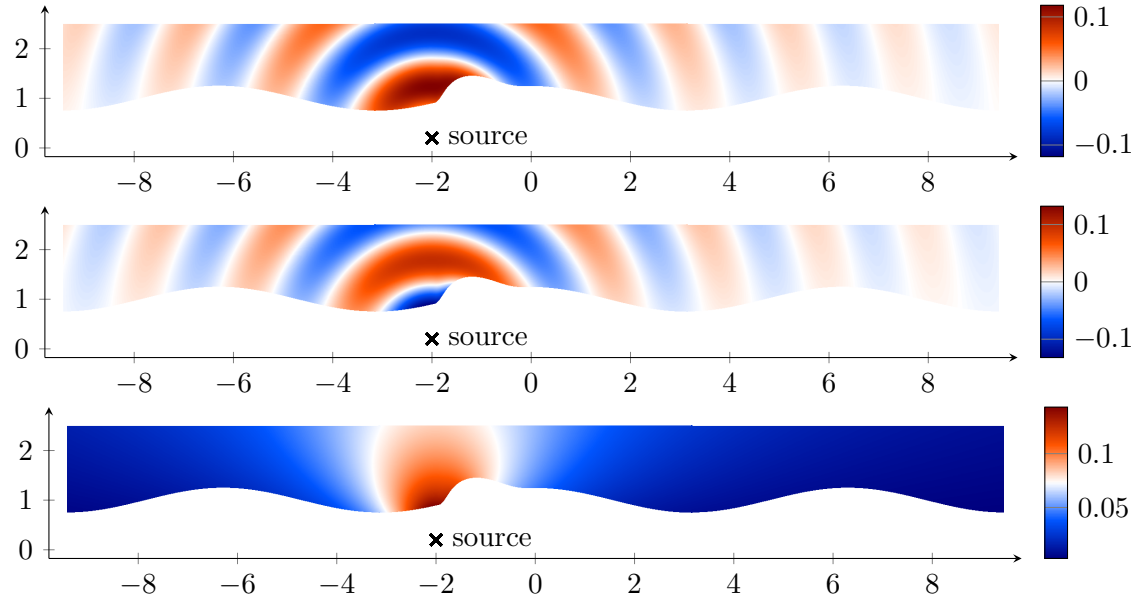


FIGURE 4.4. Numerical scattered field for Example 1 with $k = 3$ and a point source at $y = (-2, 0.2)^\top$ (top: real part, middle: imaginary part, bottom: absolute value).

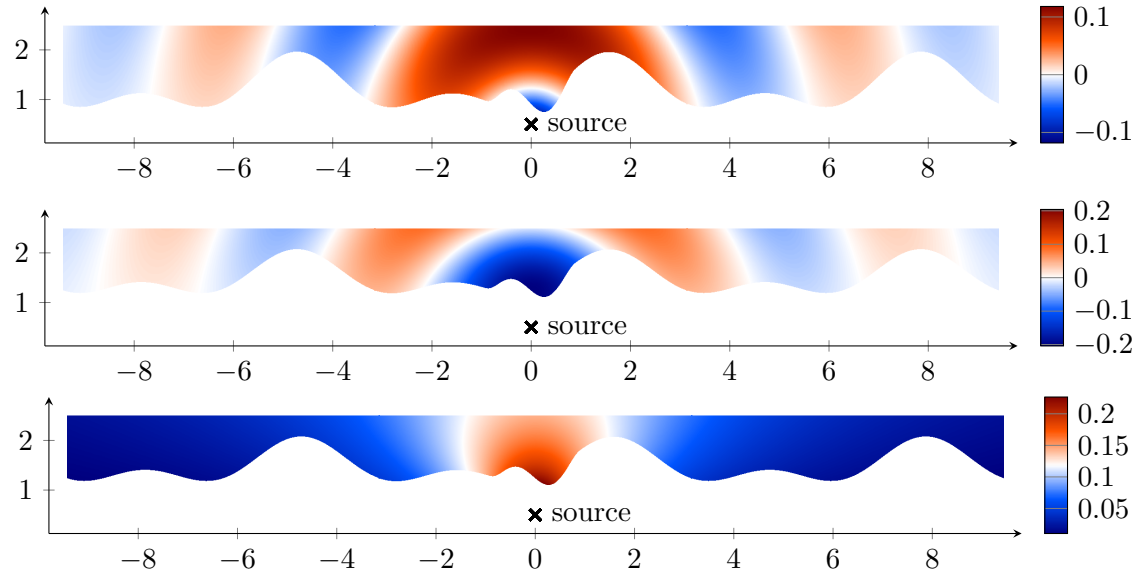


FIGURE 4.5. Numerical scattered field for Example 2 with $k = 1.5$ and a point source at $y = (0, 0.5)^\top$ (top: real part, middle: imaginary part, bottom: absolute value).

The behaviour of the numerical scattered field is illustrated in Figures 4.4 and 4.5 for Example 1 with $k = 3$ and Example 2 with $k = 1.5$, respectively. These results were obtained using the

mesh size of $\tau = 0.01$ and the PML parameter $\rho = 20$. Additionally, the absolute values of the numerical errors are plotted for examples 1 and 2 in Figures 4.6 and 4.7. They demonstrate that the maximum value of the error is less than 2×10^{-5}, which indicates the accuracy of the proposed method for these examples. Moreover, it is evident that the absolute value of the error increases while approaching the PML. This behavior is expected, as the PML introduces a numerical error due to the approximation of the DtN map.

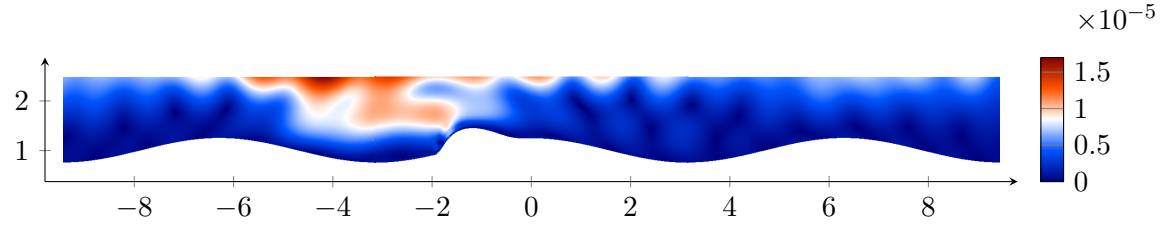


Figure 4.6. Absolute value of the error for Example 1 with $k = 3$.

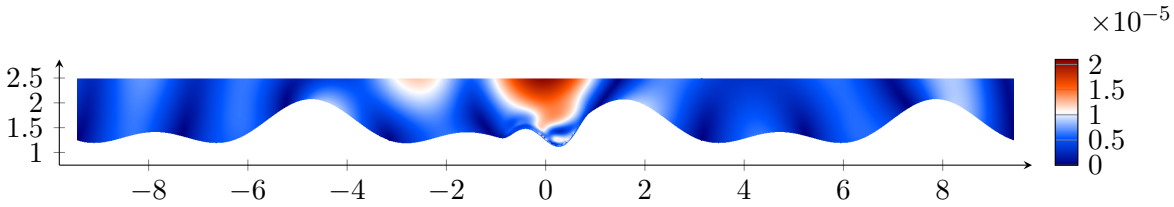


Figure 4.7. Absolute value of the error for Example 2 with $k = 1.5$.

In Table 4.1, we report the required number of iterations and the computational time used in Algorithm 3. This shows that the proposed iterative method is relatively fast and the number of iterations does not depend on the spatial discretization.

		$k = 3$		$k = 5$	
		$\tau = 0.02$	$\tau = 0.01$	$\tau = 0.02$	$\tau = 0.01$
Example 1	# iterations	7	7	8	8
	CPU time (s)	42	220	46	224
Example 2	# iterations	10	10	11	11
	CPU time (s)	40	269	43	280

Table 4.1. Number of iterations and CPU time used by Algorithm 3.

In what follows, we analyze the dependence of the relative L^2-error on the PML parameter ρ for various discretization parameters.

In Tables 4.2 and 4.3, we report the relative L^2-error of the proposed method with respect to the PML parameter ρ and mesh size τ for a fixed wave number k. These results are depicted in Figure 4.8 for both examples. The error decreases exponentially with increasing ρ up to a certain threshold, $\rho = 10$ for Example 1 and $\rho = 6$ for Example 2. Beyond these values, the error is

dominated by the discretization of the FEM. This behavior is evident from the results shown in Tables 4.2 and 4.3. For ρ values exceeding the threshold, where exponential convergence ceases, the method exhibits quadratic convergence with respect to the mesh size.

ρ	$\tau = 0.04$	$\tau = 0.02$	$\tau = 0.01$
2	1.2264×10^{-1}	1.2257×10^{-1}	1.2237×10^{-1}
4	1.5009×10^{-2}	1.4833×10^{-2}	1.4780×10^{-2}
6	1.6885×10^{-3}	1.6770×10^{-3}	1.6879×10^{-3}
8	3.0241×10^{-4}	1.8580×10^{-4}	1.9533×10^{-4}
10	3.2462×10^{-4}	8.5791×10^{-5}	3.8574×10^{-5}
12	3.5687×10^{-4}	9.4007×10^{-5}	3.8132×10^{-5}
14	3.9627×10^{-4}	1.0929×10^{-4}	3.7405×10^{-5}
16	4.2576×10^{-4}	1.2144×10^{-4}	3.3633×10^{-5}
18	4.5329×10^{-4}	1.4066×10^{-4}	4.0047×10^{-5}
20	4.8254×10^{-4}	1.6213×10^{-4}	5.2374×10^{-5}

TABLE 4.2. Relative L^2-error with respect to the PML parameter ρ and mesh size τ for Example 1 with wave number $k = 1.5$.

ρ	$\tau = 0.04$	$\tau = 0.02$	$\tau = 0.01$
2	1.4153×10^{-2}	1.3983×10^{-2}	1.3914×10^{-2}
4	1.3769×10^{-3}	4.2694×10^{-4}	2.3453×10^{-4}
6	1.2833×10^{-3}	3.3116×10^{-4}	8.5045×10^{-5}
8	1.2716×10^{-3}	3.3061×10^{-4}	8.5209×10^{-5}
10	1.2548×10^{-3}	3.2831×10^{-4}	8.4986×10^{-5}
12	1.2430×10^{-3}	3.2554×10^{-4}	8.5183×10^{-5}
14	1.2359×10^{-3}	3.2263×10^{-4}	8.4497×10^{-5}
16	1.2306×10^{-3}	3.2100×10^{-4}	8.3787×10^{-5}
18	1.2267×10^{-3}	3.2149×10^{-4}	8.4301×10^{-5}
20	1.2249×10^{-3}	3.2328×10^{-4}	8.5755×10^{-5}

TABLE 4.3. Relative L^2-error with respect to the PML parameter ρ and mesh size τ for Example 2 with wave number $k = 3$.

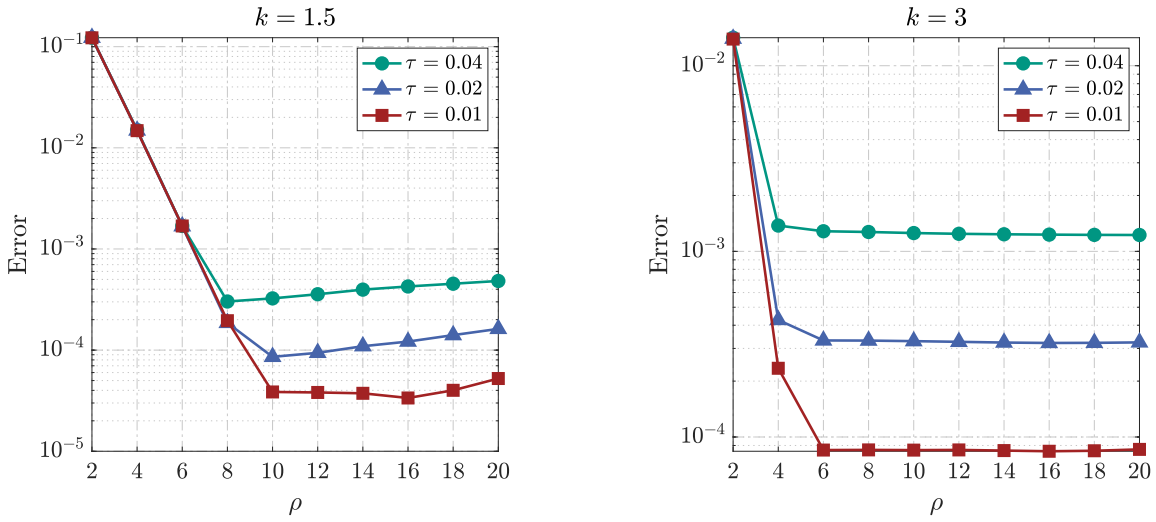


Figure 4.8. Relative L^2-error with respect to the PML parameter ρ for different mesh sizes τ (left: Example 1, right: Example 2).

In Tables 4.4 and 4.5, we report the relative L^2-error with respect to the PML parameter ρ and wave number k for the mesh size $\tau = 0.01$. These results are depicted in Figure 4.9 for both examples. We again observe an exponential rate of convergence for the wave numbers $k = \sqrt{2}, \sqrt{5}$ and $\sqrt{10}$. Furthermore, the graphs indicate that the damping effect of the PML is more pronounced when the value of $k\rho$ is larger. That is, the convergence is faster and is reached at a lower value of ρ when the wave number k is larger. For each fixed ρ, the error is smaller for larger k unless the error of the spatial discretization dominates.

ρ	$k = \sqrt{2}$	$k = \sqrt{5}$	$k = \sqrt{10}$
2	1.3110×10^{-1}	4.4683×10^{-2}	1.2256×10^{-2}
4	1.8509×10^{-2}	1.9087×10^{-3}	1.8733×10^{-4}
6	2.3898×10^{-3}	9.3975×10^{-5}	9.8692×10^{-5}
8	3.0169×10^{-4}	4.2240×10^{-5}	9.8247×10^{-5}
10	4.6151×10^{-5}	4.2936×10^{-5}	9.7619×10^{-5}
12	2.8695×10^{-5}	4.3412×10^{-5}	9.7097×10^{-5}
14	3.2423×10^{-5}	4.4872×10^{-5}	9.6927×10^{-5}
16	3.4231×10^{-5}	4.5801×10^{-5}	9.6704×10^{-5}
18	3.7340×10^{-5}	4.6738×10^{-5}	9.6429×10^{-5}
20	4.3658×10^{-5}	4.8820×10^{-5}	9.6445×10^{-5}

Table 4.4. Relative L^2-error with respect to the PML parameter ρ and wave number k for Example 1.

ρ	$k = \sqrt{2}$	$k = \sqrt{5}$	$k = \sqrt{10}$
2	1.1309×10^{-1}	4.2618×10^{-2}	1.0850×10^{-2}
4	1.5468×10^{-2}	1.7487×10^{-3}	1.6890×10^{-4}
6	1.9835×10^{-3}	6.5705×10^{-5}	9.8360×10^{-5}
8	2.3638×10^{-4}	4.2657×10^{-5}	9.8753×10^{-5}
10	4.9921×10^{-5}	4.2107×10^{-5}	9.8922×10^{-5}
12	4.5381×10^{-5}	4.1870×10^{-5}	9.8961×10^{-5}
14	4.7108×10^{-5}	4.2330×10^{-5}	9.9012×10^{-5}
16	4.9491×10^{-5}	4.2909×10^{-5}	9.9185×10^{-5}
18	5.2962×10^{-5}	4.3539×10^{-5}	9.9429×10^{-5}
20	5.8908×10^{-5}	4.4983×10^{-5}	9.9875×10^{-5}

TABLE 4.5. Relative L^2-error with respect to the PML parameter ρ and wave number k for Example 2.

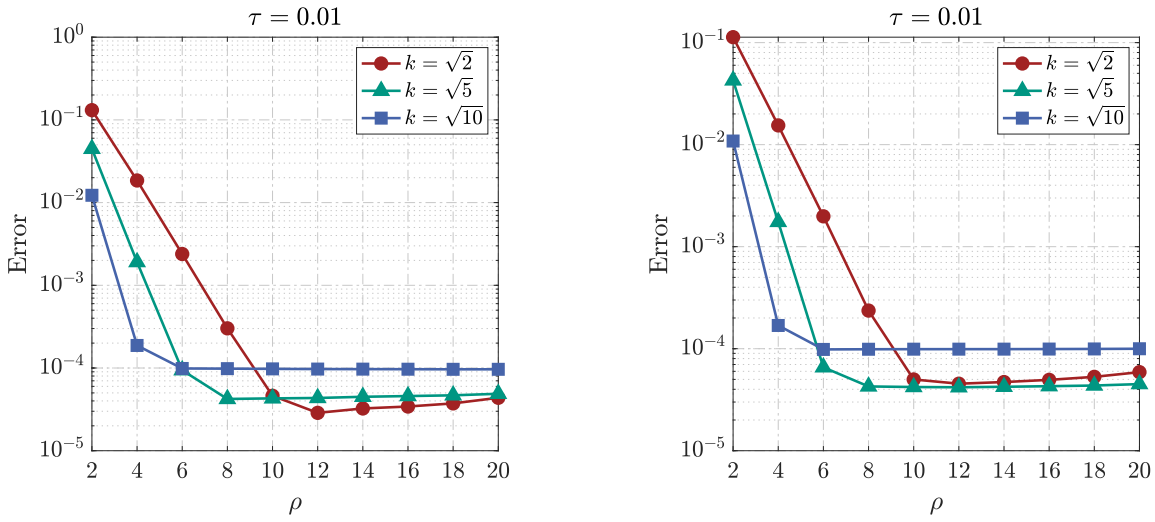


FIGURE 4.9. Relative L^2-error with respect to the PML parameter ρ and wave number k (left: Example 1, right: Example 2).

So far, we have presented numerical results for point sources located below the locally perturbed scatterer. Now, we want to illustrate how the numerical scattered field, generated by the incident field u^i with the point source located above the scatterer, propagates inside the strip between the bottom surface and the PML. In this situation, the exact solution is not available. Hence, we only show the numerical solutions which are obtained by $k = 5$, $\tau = 0.01$ and the PML parameter $\rho = 20$.

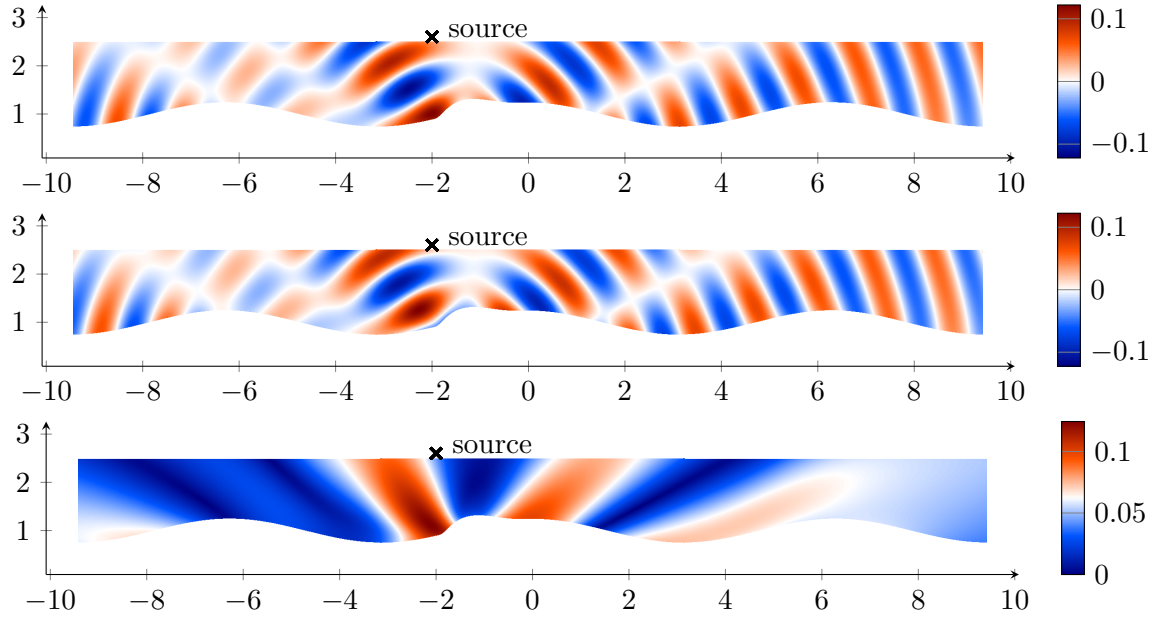


Figure 4.10. Numerical scattered field for Example 1 with a point source at $y = (-2, 2.6)^\top$ (top: real part, middle: imaginary part, bottom: absolute value).

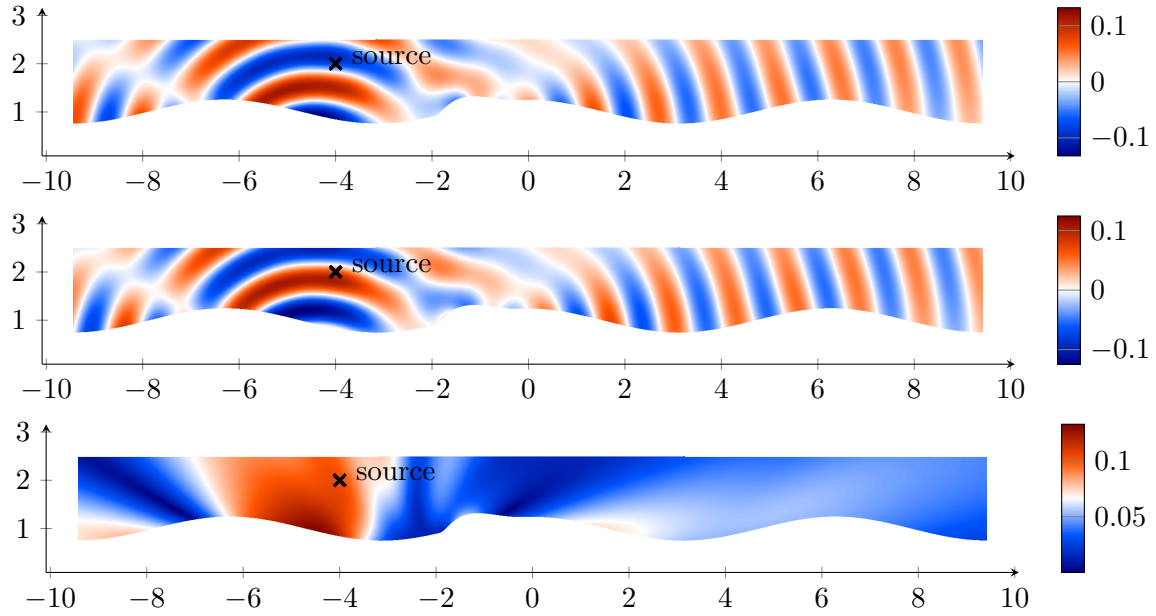


Figure 4.11. Numerical scattered field for Example 1 with a point source at $y = (-4, 2)^\top$ (top: real part, middle: imaginary part, bottom: absolute value).

Figures 4.10 and 4.11 show that the numerical scattered field corresponding to Example 1 for two different locations of the point source. In the former, the point source is located at $y = (-2, 2.6)^\top$ above the perturbation, whereas in the latter it is located at $y = (-4, 2)^\top$ away from the perturbation. In these figures, the overall propagating pattern is similar to Green's

function; however, near the point source, some interference of waves scattered from different points is visible.

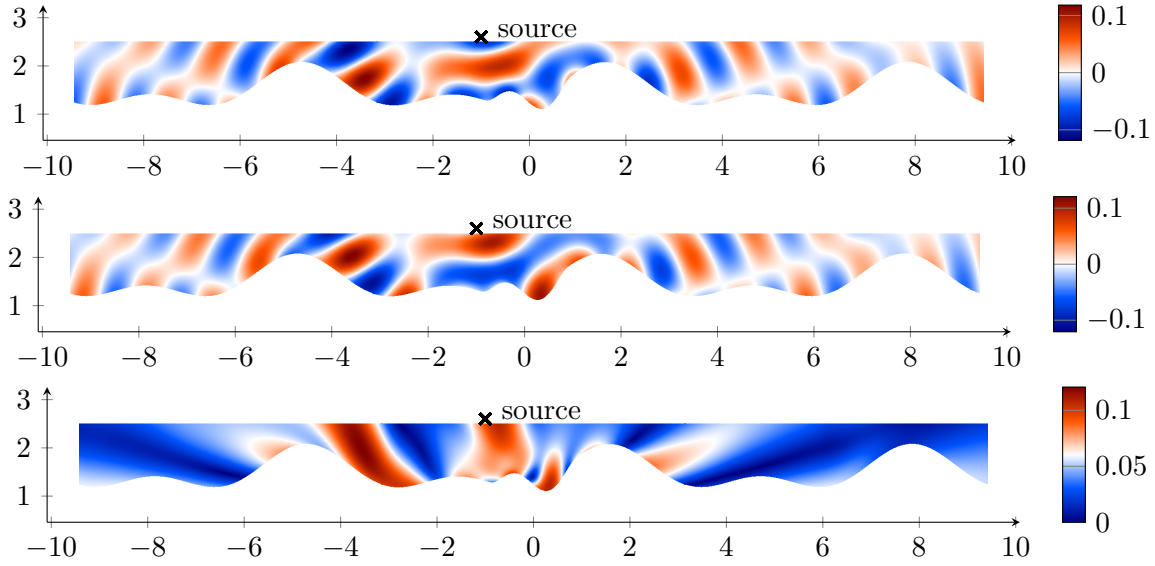


FIGURE 4.12. Numerical scattered field for Example 2 with a point source at $y = (-1, 2.6)^\top$ (top: real part, middle: imaginary part, bottom: absolute value).

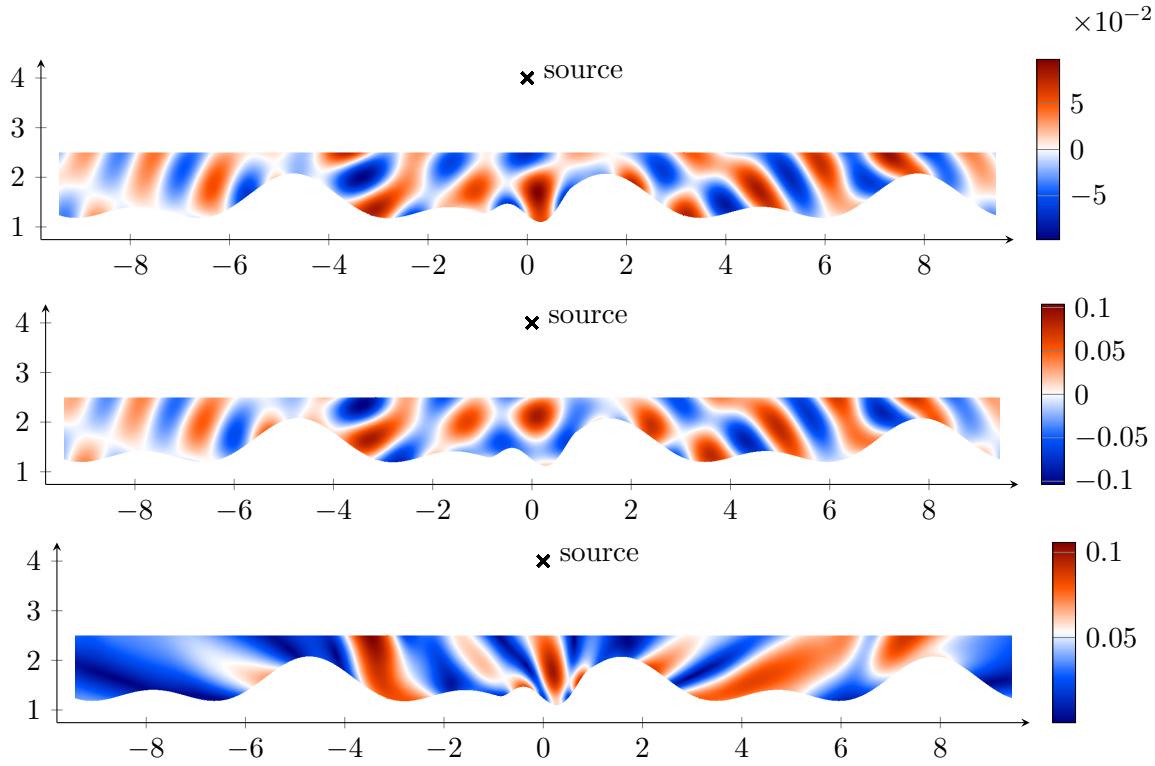


FIGURE 4.13. Numerical scattered field for Example 2 with a point source at $y = (0, 4)^\top$ (top: real part, middle: imaginary part, bottom: absolute value).

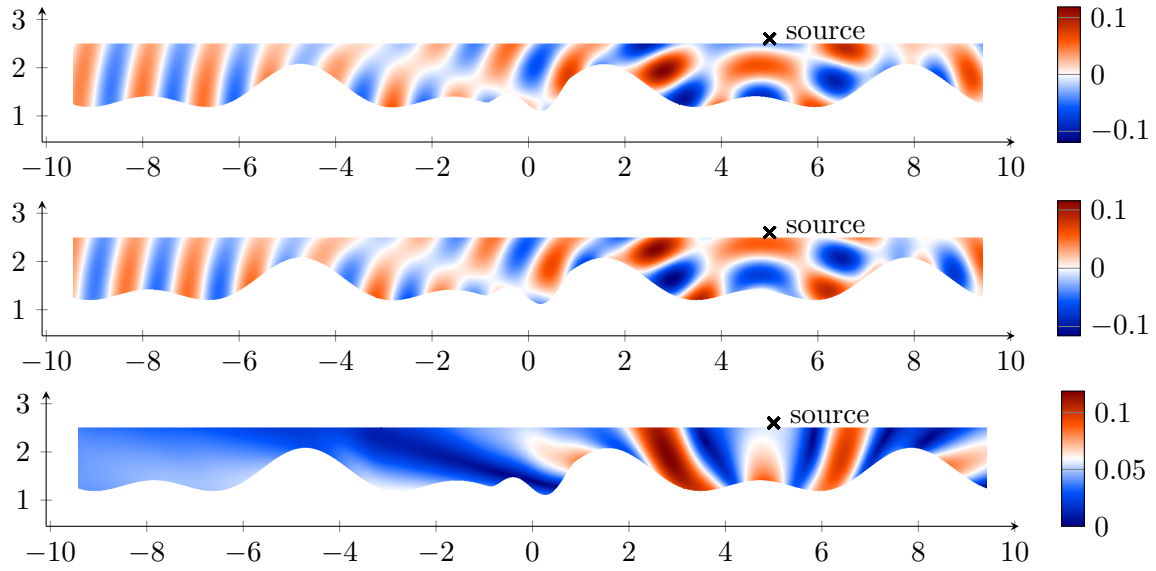


Figure 4.14. Numerical scattered field for Example 2 with a point source at $y = (5, 2.6)^\top$ (top: real part, middle: imaginary part, bottom: absolute value).

For Example 2, we set the point source in three different locations: $y = (0, 4)^\top$ above the scatterer, $y = (-1, 2.6)^\top$ relatively close to the scatterer and $y = (5, 2.6)^\top$ outside the perturbed region. The corresponding numerical scattered fields are plotted in Figures 4.12 to 4.14. Due to the complex structure of the bottom surface, the interference of the scattered waves leads to a more complicated pattern. Furthermore, in Figure 4.12, the decay of the scattered field can be seen in the horizontal direction. Finally, in Figure 4.14, it is visible that far away from the point source, the scattered field behaves like Green's function.

CHAPTER 5

Reconstruction of Local Perturbations

In this chapter, we study inverse scattering problems, where we aim to reconstruct unknown perturbations of an unbounded periodic scatterer, using measured data. This data is obtained by recording the resulting scattered field at various points in a compact set when a non-periodic incident field hits the locally perturbed structure.

We assume an a priori knowledge of the 2π-periodic function ζ^{per} which generates the periodic surface Γ^{per}. Additionally, the non-periodic incident field u^i and the corresponding measured near-field data D are also provided. The objective is to determine the shape of the defect, which generates this near field. The main difficulty here lies in the ill-posedness of such problems.

We restrict ourselves to the two-dimensional case. Without loss of generality, we consider the support of the perturbation to be a subset of $[-\pi, \pi]$ and define the set of admissible perturbations

$$X \coloneqq \left\{ \delta \in C^2(\mathbb{R}) : \operatorname{supp}(\delta) \subset [-\pi, \pi] \right\}.$$

In Chapter 4, to each perturbation δ we associate the bottom surface Γ^δ, generated by $\zeta^\delta \coloneqq \zeta^{\mathrm{per}} + \delta$, and denote by Ω_H^δ the perturbed domain between the surfaces Γ^δ and Γ_H.

In the *direct scattering problem* posed in the domain Ω_H^δ, for a given non-periodic incident field $u^i \in H^2(\Omega_H^\delta)$, we seek the nonlinear scattering operator $\mathcal{S}$

$$\begin{aligned} \mathcal{S} \colon X &\to L^2(\Gamma_H^{2\pi}), \\ \delta &\mapsto u^\delta\big|_{\Gamma_H^{2\pi}}, \end{aligned}$$

which maps a given perturbation δ to the solution of Problem (4.2) restricted on the compact set $\Gamma_H^{2\pi} \coloneqq [-\pi, \pi] \times \{H\}$.

In *inverse scattering problems*, with complete knowledge of the scattering operator $\mathcal{S}$, we aim to determine the unknown perturbation $\delta \in X$ satisfying

$$\mathcal{S}(\delta) = D \tag{5.1}$$

for given near-field data $D \coloneqq u^\delta|_{\Gamma_H^{2\pi}}$. In practice, the near field D would be obtained through measurements that include some level of noise. As a result, instead of D, a noisy right-hand side D_p

is typically considered, where $p > 0$ represents a priori knowledge of the noise level, satisfying $\|D - D_p\|_{L^2(\Gamma_H^{2\pi})} / \|D\|_{L^2(\Gamma_H^{2\pi})} \leq p$. Due to the noise in the measured data, we cannot expect to find a perturbation δ which satisfies (5.1) exactly. We hence reformulate it as a *least squares problem*, i.e., we consider the following nonlinear optimization problem: for given measured data D, find $\delta^* \in X$ such that

$$\delta^* = \arg\min_{\delta \in X} \|\mathcal{S}(\delta) - D\|^2_{L^2(\Gamma_H^{2\pi})}. \tag{5.2}$$

In Section 5.1, we prove that the nonlinear mapping $\mathcal{S}$ is *completely continuous*. This shows that the inverse problem (5.1) and hence (5.2) are ill-posed in the sense of Hadamard [54]. To find a stable approximation of the solution to such ill-posed problems, we aim to employ a regularization method, namely an iterative regularized Newton–type method. To apply this method, we require the Fréchet derivative of the scattering operator $\mathcal{S}$ at δ. In Section 5.2, we prove that this derivative exists and can be represented by the solution of a boundary value problem, which can be solved numerically by the iterative solver proposed in the previous chapter. Moreover, we introduce and discretize the regularized version of inverse problem (5.2) in Section 5.3. Finally, in Section 5.4, we will provide some numerical reconstructions that illustrate the performance of the proposed method.

Before starting, we provide an overview of the scattering problems posed in periodic and locally perturbed domains discussed in Chapters 3 and 4.

We consider as a "reference problem" the variational formulation of the direct scattering problem in the periodic domain Ω_H^{per}, given in (3.3).

Reference Problem: For the incident field $u^i \in H^1_r(\Omega_H^{\mathrm{per}})$ with $r \in [0,1)$, find the total field $u = \mathcal{J}^{-1} w \in \widetilde{H}^1(\Omega_H^{\mathrm{per}})$ such that $w \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ satisfies

$$\int_\Lambda a_\alpha(w(\alpha), z(\alpha))\,\mathrm{d}\alpha = \int_\Lambda \left\langle (\partial_{x_2} - \mathcal{T}_\alpha^+)\mathcal{J}u^i(\alpha), \overline{z(\alpha)} \right\rangle_{\Gamma_H^{2\pi}} \mathrm{d}\alpha$$

for all $z \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, where

$$a_\alpha(\phi, \psi) := \left\langle \nabla\phi, \overline{\nabla\psi} \right\rangle_{\Omega_H^{2\pi}} - 2\mathrm{i}\alpha \left\langle \partial_{x_1}\phi, \overline{\psi} \right\rangle_{\Omega_H^{2\pi}} - (k^2 - |\alpha|^2)\left\langle \phi, \overline{\psi} \right\rangle_{\Omega_H^{2\pi}} - \left\langle \mathcal{T}_\alpha^+ \phi, \overline{\psi} \right\rangle_{\Gamma_H^{2\pi}}$$

and the FB transform of the DtN map, denoted by $\mathcal{T}_\alpha^+$, is defined as in (3.5).

We next recall the variational formulation of the direct scattering problem in the locally perturbed domains Ω_H^δ, which is transformed by the diffeomorphism Ψ^δ defined in (4.4) to an equivalent problem in a periodic domain Ω_H^{per}. This transformed formulation is referred to as the "perturbed problem".

Perturbed Problem: For a given compactly supported perturbation $\delta \in X$ and incident field $u^i \in H^1_r(\Omega_H^\delta)$ with $r \in [0,1)$, find the total field $u^\delta = (\mathcal{J}^{-1} w^\delta) \circ (\Psi^\delta)^{-1} \in \widetilde{H}^1(\Omega_H^\delta)$ such that $w^\delta \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$ satisfies

$$\int_\Lambda a_\alpha^\delta(w^\delta, z(\alpha))\,\mathrm{d}\alpha = \int_\Lambda \left\langle (\partial_{x_2} - \mathcal{T}_\alpha^+)\mathcal{J}u^i(\alpha), \overline{z(\alpha)} \right\rangle_{\Gamma_H^{2\pi}} \mathrm{d}\alpha$$

for all $z \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, where $a^\delta_\alpha(\phi,\psi) \coloneqq a_\alpha(\phi(\alpha),\psi) + b^\delta_\alpha(\phi,\psi)$ with a_α as in the *Reference Problem* and

$$b^\delta_\alpha(\phi,\psi) \coloneqq \left\langle (A^\delta - I)\nabla \mathcal{J}^{-1}\phi, \overline{\nabla(\psi \mathrm{e}^{\mathrm{i}\alpha x_1})} \right\rangle_{\Omega_H^{2\pi}} - k^2 \left\langle (c^\delta - 1)\mathcal{J}^{-1}\phi, \overline{\psi \mathrm{e}^{\mathrm{i}\alpha x_1}} \right\rangle_{\Omega_H^{2\pi}}.$$

The coefficients A^δ and c^δ are given in (4.7).

Note that for $u^i \in H^2_r(\Omega^\delta_H)$ with $r \in [0,1)$, from Theorem 4.5, we know that w^δ belongs to $L^2(\Lambda; \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{\mathrm{per}}))$. Using the inverse FB transform and the inverse of the diffeomorphism Ψ^δ, we have $u^\delta = (\mathcal{J}^{-1} w^\delta) \circ (\Psi^\delta)^{-1} \in \widetilde{H}^2(\Omega^\delta_H)$.

5.1. Continuity and Compactness of the Scattering Operator

To show the continuity of the scattering operator $\mathcal{S}$, it suffices to analyze the dependence of the solution u^δ on the boundary curve ζ^δ. In [66, Thm. 9], it has been proven that solutions of quasi-periodic scattering problems depend continuously and differentiably on the periodic boundary. A straightforward extension of these results to non-periodic scattering problems is not possible, since the reduction of the problem to a bounded cell requires a periodic domain. In this case, we will follow the approach outlined in Section 4.1. To analyze the stability of the direct scattering problem, we use techniques given in [66] and prove that a small perturbation of ζ^{per} leads to small changes in the solution. To this end, we need some preliminary lemmas.

Lemma 5.1. *For $\delta \in X$, let a_α and a^δ_α be defined as in the* Reference Problem *and* Perturbed Problem*, respectively. For every $\phi, \psi \in L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, there exists a constant C such that*

$$\left\| a^\delta_\alpha(\phi,\psi(\alpha)) - a_\alpha(\phi(\alpha),\psi(\alpha)) \right\|_{L^2(\Lambda)} \leq C \|\delta\|_{1,\infty} \|\phi\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \|\psi\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))},$$

where $\|\cdot\|_{1,\infty}$ denotes the norm in $C^1([-\pi,\pi])$.

Proof. Recall that

$$\left| a^\delta_\alpha(\phi,\psi(\alpha)) - a_\alpha(\phi(\alpha),\psi(\alpha)) \right| = \left| b^\delta_\alpha(\phi,\psi(\alpha)) \right|$$

with b^δ_α defined as in the *Perturbed Problem*. From Lemma A.3, we know that

$$c^\delta = 1 + \mathcal{O}(\|\delta\|_{1,\infty}) \quad \text{and} \quad A^\delta = I + \mathcal{O}(\|\delta\|_{1,\infty}) \quad \text{as } \|\delta\|_{1,\infty} \to 0.$$

Applying the mapping property of the FB transform given in Theorem 2.28(b), we obtain

$$\begin{aligned}
\left| b^\delta_\alpha(\phi,\psi(\alpha)) \right| &\leq \left| \left\langle (A^\delta - I)\nabla_x \mathcal{J}^{-1}\phi, \overline{\nabla_x(\psi(\alpha)\mathrm{e}^{\mathrm{i}\alpha x_1})} \right\rangle_{\Omega_H^{2\pi}} \right| + \left| k^2 \left\langle (c^\delta - 1)\mathcal{J}^{-1}\phi, \overline{\psi(\alpha)\mathrm{e}^{\mathrm{i}\alpha x_1}} \right\rangle_{\Omega_H^{2\pi}} \right| \\
&\leq C\|\delta\|_{1,\infty} \left(\left\| \nabla_x \mathcal{J}^{-1}\phi \right\|_{L^2(\Omega_H^{2\pi})} \left\| \nabla_x(\psi(\alpha)\mathrm{e}^{\mathrm{i}\alpha x_1}) \right\|_{L^2(\Omega_H^{2\pi})} \right. \\
&\quad \left. + \left\| \mathcal{J}^{-1}\phi \right\|_{L^2(\Omega_H^{2\pi})} \left\| \psi(\alpha)\mathrm{e}^{\mathrm{i}\alpha x_1} \right\|_{L^2(\Omega_H^{2\pi})} \right) \\
&\leq C\|\delta\|_{1,\infty} \left\| \mathcal{J}^{-1}\phi \right\|_{H^1(\Omega_H^{\mathrm{per}})} \|\psi(\alpha)\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})} \\
&\leq C\|\delta\|_{1,\infty} \|\phi\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \|\psi(\alpha)\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})},
\end{aligned}$$

where the generic constant C depends on the wave number k. Hence, we can write

$$\left\|a_\alpha^\delta(\phi,\psi(\alpha)) - a_\alpha(\phi(\alpha),\psi(\alpha))\right\|_{L^2(\Lambda)}^2 \leq C^2\|\delta\|_{1,\infty}^2\|\phi\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}^2 \int_\Lambda \|\psi(\alpha)\|_{H^1_{\mathrm{per}}(\Omega_H^{2\pi})}^2 \,\mathrm{d}\alpha\,.$$

The assertion follows by the definition of the $L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$-norm. □

In the following lemma, we reformulate [75, Thm. 10.1] in terms of sesquilinear forms.

Lemma 5.2. *Let a_α and a_α^δ be defined as before and $w, w^\delta \in L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \setminus \{0\}$ satisfy the* Reference Problem *and the* Perturbed Problem*, respectively. For every sufficiently small perturbation δ, we can estimate the perturbation of the solution by*

$$\left\|w^\delta - w\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \leq C \sup_{\substack{z\in L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))\\ z\neq 0}} \frac{\left\|a_\alpha^\delta(w,z(\alpha)) - a_\alpha(w(\alpha),z(\alpha))\right\|_{L^2(\Lambda)}}{\|w\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\|z\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}}\,, \tag{5.3}$$

where the constant C depends on k, ζ^{per} and the non-periodic incident field u^i.

Proof. We begin by defining the operators $\mathcal{A}_\alpha, \mathcal{A}_\alpha^\delta \colon L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \to (\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))^*$ such that for all $\psi \in \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})$

$$\left\langle \mathcal{A}_\alpha\phi, \overline{\psi}\right\rangle_{\Omega_H^{2\pi}} \coloneqq a_\alpha(\phi(\alpha),\psi) \quad \text{and} \quad \left\langle \mathcal{A}_\alpha^\delta\phi, \overline{\psi}\right\rangle_{\Omega_H^{2\pi}} \coloneqq a_\alpha^\delta(\phi,\psi)\,.$$

Moreover, we define $\mathcal{A}, \mathcal{A}^\delta \colon L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \to (L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})))^*$ by

$$(\mathcal{A}\phi)(\alpha) \coloneqq \mathcal{A}_\alpha\phi \quad \text{and} \quad (\mathcal{A}^\delta\phi)(\alpha) \coloneqq \mathcal{A}_\alpha^\delta\phi\,. \tag{5.4}$$

To use the perturbation theorem given in [75, Thm. 10.1], it is required that

$$\left\|\mathcal{A}^{-1}(\mathcal{A}^\delta - \mathcal{A})\right\| < 1\,.$$

We now show that this holds for a sufficiently small perturbation δ. From the definition of the operator norm and the dual pairing in $L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))$, we obtain

$$\begin{aligned}\left\|\mathcal{A}^\delta - \mathcal{A}\right\| &= \sup_{\substack{w,z\in L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))\\ w,z\neq 0}} \frac{\left|\left\langle (\mathcal{A}^\delta - \mathcal{A})w, \overline{z}\right\rangle_{\Lambda\times\Omega_H^{2\pi}}\right|}{\|w\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\|z\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}} \\ &= \sup_{\substack{w,z\in L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))\\ w,z\neq 0}} \frac{\left|\int_\Lambda \left\langle ((\mathcal{A}^\delta - \mathcal{A})w)(\alpha), \overline{z(\alpha)}\right\rangle_{\Omega_H^{2\pi}} \mathrm{d}\alpha\right|}{\|w\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\|z\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}}\,.\end{aligned} \tag{5.5}$$

Using the definition of the operators $\mathcal{A}^\delta$ and $\mathcal{A}$ in (5.4), it follows that

$$\begin{aligned}\left|\int_\Lambda \left\langle ((\mathcal{A}^\delta - \mathcal{A})w)(\alpha), \overline{z(\alpha)}\right\rangle_{\Omega_H^{2\pi}} \mathrm{d}\alpha\right| &= \left|\int_\Lambda \left\langle (\mathcal{A}_\alpha^\delta - \mathcal{A}_\alpha)w, \overline{z(\alpha)}\right\rangle_{\Omega_H^{2\pi}} \mathrm{d}\alpha\right| \\ &= \left|\int_\Lambda a_\alpha^\delta(w, z(\alpha)) - a_\alpha(w(\alpha), z(\alpha))\,\mathrm{d}\alpha\right| \\ &\leq \left\|a_\alpha^\delta(w, z(\alpha)) - a_\alpha(w(\alpha), z(\alpha))\right\|_{L^2(\Lambda)}.\end{aligned} \tag{5.6}$$

Combining estimates (5.5), (5.6) and Lemma 5.1, we obtain

$$\left\|\mathcal{A}^\delta - \mathcal{A}\right\| \leq C\|\delta\|_{1,\infty}.$$

Since the operator $\mathcal{A}$ is boundedly invertible, we can see that for sufficiently small δ

$$\left\|\mathcal{A}^{-1}(\mathcal{A}^\delta - \mathcal{A})\right\| \leq \left\|\mathcal{A}^{-1}\right\|\left\|\mathcal{A}^\delta - \mathcal{A}\right\| \leq C_1\|\delta\|_{1,\infty} < 1,$$

where the constant C_1 depends on the wave number k and $\|\mathcal{A}^{-1}\|$.

Now, we can use the perturbation theorem given in [75, Thm. 10.1] and obtain the following estimate

$$\begin{aligned}\left\|w^\delta - w\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} &\leq \frac{\|\mathcal{A}^{-1}\|}{(1 - \|\mathcal{A}^{-1}(\mathcal{A}^\delta - \mathcal{A})\|)}\left\|(\mathcal{A}^\delta - \mathcal{A})w\right\|_{(L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})))^*} \\ &\leq \frac{\|\mathcal{A}^{-1}\|}{(1 - \|\mathcal{A}^{-1}(\mathcal{A}^\delta - \mathcal{A})\|)}\left\|\mathcal{A}^\delta - \mathcal{A}\right\|\|w\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}.\end{aligned} \tag{5.7}$$

From the *Reference Problem*, we have

$$\|w\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} = \left\|\mathcal{A}^{-1}F\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \leq \left\|\mathcal{A}^{-1}\right\|\|F\|_{L^2(\Lambda;\widetilde{H}^{-1/2}(\Gamma_H^{2\pi}))},$$

with $F := (\partial_{x_2} - \mathcal{T}_\alpha^+)\mathcal{J}u^i$. Since the conormal derivative ∂_{x_2} and the operator $\mathcal{T}_\alpha^+$ are continuous, there exists a constant C_2 such that $\|F\|_{L^2(\Lambda;\widetilde{H}^{-1/2}(\Gamma_H^{2\pi}))} \leq C_2$. By substituting this estimate into (5.7) and using (5.5) and (5.6), we obtain

$$\left\|w^\delta - w\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \leq C_3 \sup_{\substack{w,z\in L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})) \\ w,z\neq 0}} \frac{\left\|a_\alpha^\delta(w, z(\alpha)) - a_\alpha(w(\alpha), z(\alpha))\right\|_{L^2(\Lambda)}}{\|w\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\|z\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}}$$

where $C_3 := C_2\|\mathcal{A}^{-1}\|^2(1 - \|\mathcal{A}^{-1}(\mathcal{A}^\delta - \mathcal{A})\|)^{-1}$. □

Theorem 5.3. *Let $\zeta^{\mathrm{per}} \in C^2(\mathbb{R})$ be a 2π-periodic function and $\zeta^\delta := \zeta^{\mathrm{per}} + \delta$ be a locally perturbed function for a sufficiently small perturbation $\delta \in C^2(\mathbb{R})$. Suppose that $u \in \widetilde{H}^1(\Omega_H^{\mathrm{per}})$ is the solution of the* Reference Problem*, whereas $u^\delta \in \widetilde{H}^1(\Omega_H^\delta)$ satisfies the* Perturbed Problem*. Moreover, let K be a compact set such that for every $(x_1, x_2) \in K$, it holds*

$$\max\left\{\|\zeta^{\mathrm{per}}\|_\infty, \left\|\zeta^\delta\right\|_\infty\right\} < x_2 \leq H.$$

Then, there exists a constant $\widehat{C}$ depending on k, ζ^{per} and K such that

$$\left\|u^\delta - u\right\|_{H^1(K)} \leq \widehat{C}\|\delta\|_{1,\infty}\,.$$

Proof. Let $u^\delta_{\mathrm{tra}} \coloneqq u^\delta \circ \Psi^\delta$ with the diffeomorphism Ψ^δ given in (4.4). This diffeomorphism depends on the parameter h, which we select to be $h \coloneqq \min\{x_2 : x \in K\}$. Note that with this choice, we have $\Psi^\delta|_K = I$, which together with the definition of u^δ_{tra} leads to $u^\delta|_K = u^\delta_{\mathrm{tra}}|_K$. Hence, we obtain

$$\left\|u^\delta - u\right\|_{H^1(K)} = \left\|u^\delta_{\mathrm{tra}} - u\right\|_{H^1(K)} \leq \left\|u^\delta_{\mathrm{tra}} - u\right\|_{H^1(\Omega_H^{\mathrm{per}})}\,.$$

From the definition of the inverse FB transform in (2.35) and afterwards using the mapping property of the FB transform given in Theorem 2.28(b), there exists a constant C such that for $j \in \mathbb{Z}$

$$\begin{aligned}\left\|u^\delta - u\right\|_{H^1(K)} &\leq \left\|\int_\Lambda \left(w^\delta(\alpha;x) - w(\alpha;x)\right) \mathrm{e}^{\mathrm{i}\alpha(x_1+2\pi j)}\,\mathrm{d}\alpha\right\|_{H^1(\Omega_H^{\mathrm{per}})} \\ &\leq C\left\|w^\delta - w\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\,.\end{aligned}$$

Combining Lemmas 5.1 and 5.2, we can see that

$$\left\|w^\delta - w\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \leq C \sup_{\substack{z\in L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))\\ z\neq 0}} \frac{\left\|a^\delta_\alpha(w, z(\alpha)) - a_\alpha(w(\alpha), z(\alpha))\right\|_{L^2(\Lambda)}}{\|w\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\|z\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}} \leq \widetilde{C}\|\delta\|_{1,\infty}\,,$$

where the constant $\widetilde{C}$ depends on k and $\mathcal{J}u^i$. □

In the next theorem, we prove that the scattering operator $\mathcal{S}$ corresponding to unbounded scatterers is locally compact.

Theorem 5.4. *The mapping $\mathcal{S}\colon X \to L^2(\Gamma_H^{2\pi})$ is locally compact.*

Proof. Let the compact set $K \coloneqq [-\pi,\pi] \times [h,H]$ for some $h > \max\{\|\zeta^{\mathrm{per}}\|_\infty, \|\zeta^\delta\|_\infty\}$ and $\Gamma_H^{2\pi} \subseteq \partial K$. To show the compactness of the nonlinear operator $\mathcal{S}$, we need to prove that it maps every sufficiently small neighborhood U of $\delta \in X$ into a relatively compact subset of $L^2(\Gamma_H^{2\pi})$.

In Theorem 5.3, we show the continuity of the operator $\mathcal{S}_K\colon X \to H^1(K)$ with respect to the perturbation $\delta \in X$. This means, the operator maps a bounded set $U \subset X$ into a bounded subset of $H^1(K)$. Since the trace operator $\gamma_{\mathrm{D}}\colon H^1(K) \to H^{1/2}(\Gamma_H^{2\pi})$ is continuous, then $\gamma_{\mathrm{D}} \circ \mathcal{S}_K$ is also continuous. From the *compact embedding theorem* in fractional Sobolev spaces from [32, Cor. 7.2], we know that the embedding $\mathscr{I}\colon H^{1/2}(\Gamma_H^{2\pi}) \hookrightarrow L^2(\Gamma_H^{2\pi})$ is compact. That means, every bounded set in $H^{1/2}(\Gamma_H^{2\pi})$ is relatively compact with respect to the $L^2(\Gamma_H^{2\pi})$-norm. Thus, $\mathcal{S} = \mathscr{I} \circ \gamma_{\mathrm{D}} \circ \mathcal{S}_K$ maps every bounded subset of X into a relatively compact subset of $L^2(\Gamma_H^{2\pi})$ and is therefore compact. □

According to [31, Thm. 4.2], the inverse problem (5.1) is ill-posed since the scattering operator $\mathcal{S}$ is continuous and compact. Consequently, the optimization problem (5.2) is also unstable.

Before introducing the regularized version of (5.2), we focus on the Fréchet derivative of the scattering operator with respect to the perturbation. This is a key requirement for applying a Newton–type method.

5.2. Fréchet Differentiability of the Scattering Operator

So far, we have shown that the scattering operator $\mathcal{S}\colon X \to L^2(\Gamma_H^{2\pi})$ is continuous and locally compact. In this section, we are going to derive its Fréchet derivative, which is denoted by $\mathcal{S}'(\delta)\colon X \to L^2(\Gamma_H^{2\pi})$ and satisfies

$$\frac{1}{\|\eta\|_{1,\infty}}\|\mathcal{S}(\delta+\eta)-\mathcal{S}(\delta)-\mathcal{S}'(\delta)\eta\|_{L^2(\Gamma_H^{2\pi})} \to 0 \quad \text{as } \|\eta\|_{1,\infty} \to 0\,. \tag{5.8}$$

Fréchet differentiability of scattering operators with respect to the boundary is studied for bounded obstacles in [68], whereas in [66] the Fréchet differentiability of the quasi-periodic field with respect to the unbounded periodic curve has been shown.

In this section, we aim to establish the Fréchet differentiability of the scattering operator $\mathcal{S}$ for non-periodic incident fields with respect to the perturbation δ imposed on the periodic curve. Since the problem lacks periodicity, we cannot directly exploit the usual reduction to a bounded reference cell and use the result of [66]. The main idea is to use a diffeomorphism to transform the perturbed structure to the periodic one and afterwards use the FB transform (the same technique as in the previous chapter). In this case, for each Floquet parameter α, we can use a similar approach as in [66, Thm. 9] to prove the Fréchet differentiability of the scattering operator $\mathcal{S}$ with respect to δ.

In Theorem 5.5, we prove the Fréchet differentiability of the scattering operator $\mathcal{S}$ at $\delta = 0$ (corresponding to the periodic curve) and compute its Fréchet derivative. Afterwards, in Theorem 5.6, we extend these results to a sufficiently small δ (corresponding to a perturbed curve) by proving that the operator $\mathcal{S}$ is differentiable at δ.

Theorem 5.5. *Let K be as in Theorem 5.3 and $\eta \in X$. Then, the Fréchet derivative $\mathcal{S}'(0)$ of $\mathcal{S}$ at $\delta = 0$ in the direction η exists and is given by $u'|_K \in H^1(\Omega_H^{\mathrm{per}})$, where u' satisfies*

$$\Delta u' + k^2 u' = 0 \qquad \text{in } \Omega_H^{\mathrm{per}}, \tag{5.9a}$$

$$\partial_{x_2} u' = \mathcal{T}^+ u' \qquad \text{on } \Gamma_H, \tag{5.9b}$$

$$u' = -\frac{\eta}{\sqrt{1+(\zeta^{\mathrm{per}})'^2}}\partial_n u = -\eta\partial_{x_2} u \qquad \text{on } \Gamma^{\mathrm{per}}, \tag{5.9c}$$

with $\mathcal{T}^+$ defined as in (2.29) and the total field $u \in \widetilde{H}^2(\Omega_H^{\mathrm{per}})$ satisfying

$$\begin{aligned} \Delta u + k^2 u &= 0 && \text{in } \Omega_H^{\mathrm{per}}\,,\\ u &= 0 && \text{on } \Gamma^{\mathrm{per}}\,,\\ (\partial_{x_2} - \mathcal{T}^+)u &= (\partial_{x_2} - \mathcal{T}^+)u^i && \text{on } \Gamma_H\,, \end{aligned}$$

for a non-periodic incident field $u^i \in H^2(\Omega_H^{\mathrm{per}})$.

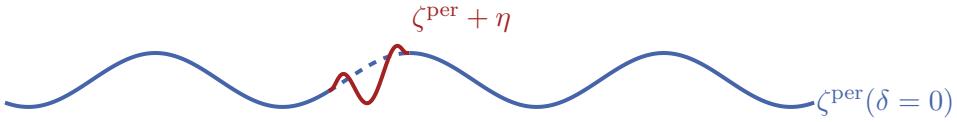


FIGURE 5.1. Periodic function ζ^{per} and locally perturbed function $\zeta^{\mathrm{per}} + \eta$.

Proof. Let u' satisfy problem (5.9) and the perturbation η be such that $\|\zeta^{\mathrm{per}}+\eta\|_\infty < H$ (depicted in Figure 5.1). In this case, we can define the domain

$$\Omega_H^\eta := \{x \in \mathbb{R}^2 : \zeta^{\mathrm{per}} + \eta < x_2 < H\}\,.$$

Moreover, we consider that $u^\eta := \mathcal{J}^{-1} w^\eta \circ (\Psi^\eta)^{-1}$ is a solution of the *Perturbed Problem* for the perturbation η and the diffeomorphism Ψ^η, mapping Ω_H^{per} to Ω_H^η, is defined as in (4.4).

The diffeomorphism Ψ^η depends on the auxiliary function β_h^η given in (4.5) with the parameter $h := \min\{x_2 : x \in K\}$. With this choice of h, we have $\beta_h^\eta|_K = 0$ and $\Psi^\eta|_K = I$.

According to the definition of the Fréchet derivative in (5.8), it suffices to show that

$$\|u^\eta - u - u'\|_{H^1(K)} = \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0\,. \tag{5.10}$$

Considering $u_{\mathrm{tra}}^\eta := u^\eta \circ \Psi^\eta$ and using $\Psi^\eta|_K = I$, we see that

$$\|u^\eta - u - u'\|_{H^1(K)} = \|u_{\mathrm{tra}}^\eta - u - u'\|_{H^1(K)}\,.$$

By defining $v^\eta(x) := \eta(x_1)\beta_h^\eta(x)\partial_{x_2} u(x)$, it is sufficient to prove

$$\|u_{\mathrm{tra}}^\eta - u - (u' + v^\eta)\|_{H^1(\Omega_H^{\mathrm{per}})} = \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0\,,$$

since $v^\eta = 0$ on the compact set K. To this end, by using the definition of the inverse FB transform (2.35) and afterwards applying the mapping property of the FB transform in Theorem 2.28(b), we obtain that there is a constant C such that

$$\begin{aligned}\|u_{\mathrm{tra}}^\eta - u - (u' + v^\eta)\|_{H^1(\Omega_H^{\mathrm{per}})} &= \left\| \int_\Lambda \left(w^\eta - w - \mathcal{J}(u' + v^\eta)\right) \mathrm{e}^{\mathrm{i}\alpha(x_1 + 2\pi j)}\, \mathrm{d}\alpha \right\|_{H^1(\Omega_H^{\mathrm{per}})} \\ &\le C \|w^\eta - w - \mathcal{J}(u' + v^\eta)\|_{L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\,.\end{aligned}$$

Let a_α and a_α^η be defined as in the *Reference Problem* and the *Perturbed Problem*. Using the inf-sup condition (2.15) for each $\alpha \in \Lambda$, we get

$$\|w^\eta - w - \mathcal{J}(u' + v^\eta)\|_{L^2(\Lambda; \widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))} \le \sup_{\substack{z \in \widetilde{H}^2(\Omega_H^{2\pi}) \\ z \neq 0}} \frac{a_\alpha(w^\eta - w - \mathcal{J}(u' + v^\eta), z)}{C_{\mathrm{infsup}} \|z\|_{H^1(\Omega_H^{2\pi})}}\,,$$

where C_{infsup} denotes the inf-sup constant. Now it is enough to compute for almost all $\alpha \in \Lambda$ and all $z \in \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$

$$a_\alpha(w^\eta(\alpha) - w(\alpha) - \mathcal{J}(u' + v^\eta)(\alpha), z) = a_\alpha(w^\eta(\alpha) - w(\alpha), z) - a_\alpha(\mathcal{J}(u' + v^\eta)(\alpha), z)\,.$$

Since the right-hand sides of the *Reference Problem* and *Perturbed Problem* are equal, we conclude that for each $\alpha \in \Lambda$, we have $a_\alpha(w(\alpha), z) = a_\alpha^\eta(w^\eta, z)$. This leads to

$$\begin{aligned} &a_\alpha(w^\eta(\alpha) - w(\alpha) - \mathcal{J}(u' + v^\eta)(\alpha), z) \\ &\qquad = -\left(a_\alpha^\eta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z)\right) - a_\alpha(\mathcal{J}(u' + v^\eta)(\alpha), z)\,. \end{aligned} \tag{5.11}$$

We begin by computing the first term in (5.11). Using the definition of the sesquilinear forms a_α and a_α^η given in the *Reference Problem* and *Perturbed Problem*, we have

$$\begin{aligned} a_\alpha^\eta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z) = &\left\langle (A^\eta - I)\nabla_x(\mathcal{J}^{-1}w^\eta), \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} \\ &- k^2\left\langle (c^\eta - 1)\mathcal{J}^{-1}w^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}}. \end{aligned} \tag{5.12}$$

The asymptotic representations of A^η and c^η in (A.4) and (A.5) give us $c^\eta - 1 = \eta\partial_{x_2}\beta_h^\eta$ and

$$\begin{aligned} &\left((A^\eta - I)\nabla_x(\mathcal{J}^{-1}w^\eta)\right)\cdot\overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})} \\ &\quad = -\operatorname{div}\left\{v^\eta\overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})} + \eta\beta_h^\eta\nabla_x(\mathcal{J}^{-1}w^\eta)\overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})} - \eta\beta_h^\eta e_2(\nabla_x(\mathcal{J}^{-1}w^\eta)\cdot\overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})})\right\} \\ &\qquad + v^\eta\overline{\Delta_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})} + \eta\beta_h^\eta\Delta_x(\mathcal{J}^{-1}w^\eta)\overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty}\to 0\,. \end{aligned}$$

Substituting the above representations into (5.12) and using the divergence theorem, we obtain

$$\begin{aligned} a_\alpha^\eta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z) = &-\left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma_H^{2\pi}} + \left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\ &- \left\langle \eta\beta_h^\eta\nabla_x(\mathcal{J}^{-1}w^\eta), \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma_H^{2\pi}} \\ &+ \left\langle \eta\beta_h^\eta\nabla_x(\mathcal{J}^{-1}w^\eta), \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\ &+ \left\langle v^\eta, \overline{\Delta_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} + \left\langle \eta\beta_h^\eta\Delta_x(\mathcal{J}^{-1}w^\eta), \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} \\ &- k^2\left\langle \eta(\partial_{x_2}\beta_h^\eta)\mathcal{J}^{-1}w^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \text{ as } \|\eta\|_{1,\infty}\to 0\,, \end{aligned}$$

where $n = (n_1, n_2)^\top$ denotes the outward unit normal vector. From the definition of β_h^η in (4.5), we know that $\beta_h^\eta = 0$ on the top surface $\Gamma_H^{2\pi}$ and $\beta_h^\eta = 1$ on the bottom surface $\Gamma^{2\pi}$. This yields

$$\begin{aligned} a_\alpha^\eta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z) = &-\left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma_H^{2\pi}} + \left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\ &+ \left\langle \eta\nabla_x(\mathcal{J}^{-1}w^\eta), \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\ &+ \left\langle v^\eta, \overline{\Delta_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} + \left\langle \eta\beta_h^\eta\Delta_x(\mathcal{J}^{-1}w^\eta), \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} \\ &- k^2\left\langle \eta(\partial_{x_2}\beta_h^\eta)\mathcal{J}^{-1}w^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \text{ as } \|\eta\|_{1,\infty}\to 0\,. \end{aligned}$$

Applying Green's first identity to $\left\langle v^\eta, \overline{\Delta_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}}$ gives

$$\begin{aligned}
a_\alpha^\eta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z) = &-\left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma_H^{2\pi}} + \left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\
&+ \left\langle \eta\nabla_x(\mathcal{J}^{-1}w^\eta), \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\
&- \left\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} + \left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma_H^{2\pi}} \\
&- \left\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} + \left\langle \eta\beta_h^\eta\Delta_x(\mathcal{J}^{-1}w^\eta), \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} \\
&- k^2\left\langle \eta(\partial_{x_2}\beta_h^\eta)\mathcal{J}^{-1}w^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \text{ as } \|\eta\|_{1,\infty} \to 0\,.
\end{aligned}$$

Some of the boundary terms cancel each other out and we can further simplify the terms above, by using $\eta\Delta_x(\mathcal{J}^{-1}w^\eta) = -\eta k^2(\mathcal{J}^{-1}w^\eta) + \mathcal{O}(\|\eta\|_{1,\infty}^2)$ as $\|\eta\|_{1,\infty} \to 0$ (see Remark A.4). This leads to

$$\begin{aligned}
a_\alpha^\eta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z) = &\left\langle \eta\nabla_x(\mathcal{J}^{-1}w^\eta), \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\
&- \left\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} - k^2\left\langle \eta\beta_h^\eta\mathcal{J}^{-1}w^\eta, \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} \\
&- k^2\left\langle \eta(\partial_{x_2}\beta_h^\eta)\mathcal{J}^{-1}w^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2)\,.
\end{aligned} \tag{5.13}$$

Now, the first term in the equation above vanishes, since

$$\begin{aligned}
&\left\langle \eta\nabla_x(\mathcal{J}^{-1}w^\eta), \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\
&\quad= \left\langle n_1\eta\partial_{x_1}(\mathcal{J}^{-1}w^\eta) + n_2\eta\partial_{x_2}(\mathcal{J}^{-1}w^\eta), \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\
&\qquad- \left\langle n_2\eta\partial_{x_1}(\mathcal{J}^{-1}w^\eta), \overline{\partial_{x_1}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} - \left\langle n_2\eta\partial_{x_2}(\mathcal{J}^{-1}w^\eta), \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\
&\quad= \left\langle \eta\partial_{x_1}(\mathcal{J}^{-1}w^\eta), \overline{n_1\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\partial_{x_1}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Gamma^{2\pi}} \\
&\quad= \left\langle \eta\partial_{x_1}(\mathcal{J}^{-1}w^\eta), \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})\cdot n^\perp}\right\rangle_{\Gamma^{2\pi}} = 0\,,
\end{aligned}$$

in which the last equality is obtained by the fact that $z\mathrm{e}^{\mathrm{i}\alpha x_1}\big|_{\Gamma^{2\pi}}$ is constantly zero and hence its tangential derivative is zero. Substituting the result above into (5.13), yields

$$\begin{aligned}
a_\alpha^\eta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z) = &-\left\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}} \\
&- k^2\left\langle \eta\beta_h^\eta\mathcal{J}^{-1}w^\eta, \overline{\partial_{x_2}z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}} \\
&- k^2\left\langle \eta(\partial_{x_2}\beta_h^\eta)\mathcal{J}^{-1}w^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}} \\
&+ \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0\,.
\end{aligned} \tag{5.14}$$

Now, we are going to compute the second term of (5.11). By using the definition of the sesquilinear form a_α given in the *Reference Problem*, this term is written as follows

$$a_\alpha(\mathcal{J}(u' + v^\eta)(\alpha), z) = a_\alpha(\mathcal{J}u'(\alpha), z) + a_\alpha(\mathcal{J}v^\eta(\alpha), z)\,. \tag{5.15}$$

Applying the FB transform to (5.9) and writing its variational form, we see that $\mathcal{J}u'$ satisfies

$$a_\alpha(\mathcal{J}u'(\alpha), z) = 0 \quad \text{for all } z \in \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi}) \text{ and almost all } \alpha \in \Lambda\,.$$

Thus, since this term vanishes in equation (5.15), it only remains to compute $a_\alpha(\mathcal{J}v^\eta(\alpha), z)$. From the definition of $v^\eta(x) = \eta(x_1)\beta_h^\eta(x)\partial_{x_2}u$, we know that v^η is compactly supported in $\Omega_H^{2\pi}$, since the local perturbation η is supported in this cell. This leads to $\mathcal{J}v^\eta = v^\eta \mathrm{e}^{-\mathrm{i}\alpha x_1}$ and $\nabla_x(\mathcal{J}v^\eta) = (\nabla_x v^\eta - \mathrm{i}\alpha e_1 v^\eta)\mathrm{e}^{-\mathrm{i}\alpha x_1}$ for $e_1 \coloneqq (1,0)^\top$. Moreover, from the definition of β_h^η in (4.5), we obtain that $\mathcal{T}_\alpha^+(\mathcal{J}v^\eta) = 0$ as $\beta_h^\eta = 0$ on $\Gamma_H^{2\pi}$. Taking the above properties into account and substituting them into equation (5.15), we obtain

$$\begin{aligned}
a_\alpha(\mathcal{J}(u'+v^\eta)(\alpha), z) &= \Big\langle \nabla_x(\mathcal{J}v^\eta), \overline{\nabla_x z}\Big\rangle_{\Omega_H^{2\pi}} - 2\mathrm{i}\alpha\langle \partial_{x_1}(\mathcal{J}v^\eta), \overline{z}\rangle_{\Omega_H^{2\pi}} \\
&\quad - (k^2-\alpha^2)\langle \mathcal{J}v^\eta, \overline{z}\rangle_{\Omega_H^{2\pi}} - \Big\langle \mathcal{T}_\alpha^+(\mathcal{J}v^\eta), \overline{z}\Big\rangle_{\Gamma_H^{2\pi}} \\
&= \Big\langle (\nabla_x v^\eta), \overline{\mathrm{e}^{\mathrm{i}\alpha x_1}\nabla_x z}\Big\rangle_{\Omega_H^{2\pi}} - \mathrm{i}\alpha\Big\langle v^\eta, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1}\partial_{x_1} z}\Big\rangle_{\Omega_H^{2\pi}} \\
&\quad - 2\mathrm{i}\alpha\Big\langle \partial_{x_1} v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}} - 2\alpha^2\Big\langle v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}} \\
&\quad - (k^2-\alpha^2)\Big\langle v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}}\,.
\end{aligned} \tag{5.16}$$

Substituting (5.14) and (5.16) into (5.11), shows that

$$\begin{aligned}
&a_\alpha(w^\eta(\alpha) - w(\alpha) - \mathcal{J}(u'+v^\eta)(\alpha), z) \\
&\quad = -(a_\alpha^\delta(w^\eta, z) - a_\alpha(w^\eta(\alpha), z)) - a_\alpha((\mathcal{J}u' + \mathcal{J}v^\eta)(\alpha), z) \\
&\quad = \Big\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{2\pi}} + k^2\Big\langle \eta\beta_h^\eta \mathcal{J}^{-1}w^\eta, \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{2\pi}} \\
&\qquad + k^2\Big\langle \eta(\partial_{x_2}\beta_h^\eta)\mathcal{J}^{-1}w^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}} - \Big\langle (\nabla_x v^\eta), \overline{\mathrm{e}^{\mathrm{i}\alpha x_1}\nabla_x z}\Big\rangle_{\Omega_H^{2\pi}} \\
&\qquad + \mathrm{i}\alpha\Big\langle v^\eta, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1}\partial_{x_1} z}\Big\rangle_{\Omega_H^{2\pi}} + 2\mathrm{i}\alpha\Big\langle \partial_{x_1} v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}} + \alpha^2\Big\langle v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}} \\
&\qquad + k^2\Big\langle v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0\,.
\end{aligned}$$

Since

$$\begin{aligned}
\mathrm{i}\alpha\Big\langle v^\eta \mathrm{e}^{-\mathrm{i}\alpha x_1}, \overline{\partial_{x_1} z}\Big\rangle_{\Omega_H^{2\pi}} + \alpha^2\Big\langle v^\eta \mathrm{e}^{-\mathrm{i}\alpha x_1}, \overline{z}\Big\rangle_{\Omega_H^{2\pi}} &= \mathrm{i}\alpha\Big\langle v^\eta, \overline{\partial_{x_1}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{2\pi}} \\
&= -\mathrm{i}\alpha\Big\langle \partial_{x_1} v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}}\,,
\end{aligned}$$

all terms which do not depend on k cancel each other out. We are left with

$$\begin{aligned}
&a_\alpha(w^\eta(\alpha) - w(\alpha) - \mathcal{J}(u'+v^\eta)(\alpha), z) \\
&\quad = k^2 \int_{\Omega_H^{2\pi}} \partial_{x_2}(\eta\beta_h^\eta(\mathcal{J}^{-1}w^\eta)\overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}})\,\mathrm{d}x + \mathcal{O}(\|\eta\|_{1,\infty}^2) \\
&\quad = k^2 \int_{\Omega_H^{2\pi}} \operatorname{div}(e_2\eta\beta_h^\eta(\mathcal{J}^{-1}w^\eta)\overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}})\,\mathrm{d}x + \mathcal{O}(\|\eta\|_{1,\infty}^2)\,.
\end{aligned}$$

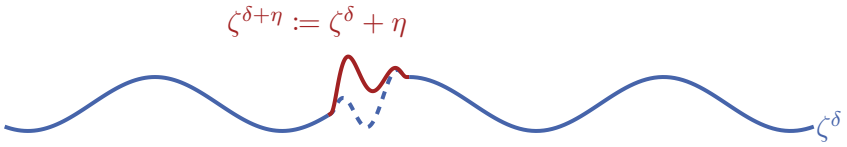


FIGURE 5.2. Locally perturbed functions ζ^δ and $\zeta^\delta + \eta$.

Using the divergence theorem, we end up with

$$\begin{aligned} a_\alpha(w^\eta(\alpha) - w(\alpha) - \mathcal{J}(u' + v^\eta)(\alpha), z) = k^2 \int_{\Gamma_H^{2\pi}} n_2 \eta \beta_h^\eta (\mathcal{J}^{-1} w^\eta) \overline{z \mathrm{e}^{\mathrm{i}\alpha x_1}} \, \mathrm{d}s \\ - k^2 \int_{\Gamma^{2\pi}} n_2 \eta \beta_h^\eta (\mathcal{J}^{-1} w^\eta) \overline{z \mathrm{e}^{\mathrm{i}\alpha x_1}} \, \mathrm{d}s \\ + \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0 \,. \end{aligned}$$

Using the fact that $z = 0$ on $\Gamma^{2\pi}$ and $\beta_h^\eta = 0$ on $\Gamma_H^{2\pi}$, we conclude

$$a_\alpha(w^\eta(\alpha) - w(\alpha) - \mathcal{J}(u' + v^\eta)(\alpha), z) = \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0$$

for all $z \in \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$ and almost all $\alpha \in \Lambda$. As $\widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$ is dense in $\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})$, we obtain

$$\|u^\eta_{\mathrm{tra}} - u - (u' + v^\eta)\|_{H^1(\Omega_H^{\mathrm{per}})} = \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0 \,.$$

This completes the proof because on the compact set K, we have $v^\eta = 0$ and $u^\eta_{\mathrm{tra}}|_K = u^\eta|_K$. □

Theorem 5.6. *Let the compact set K be defined as in Theorem 5.3 and $\eta \in X$. Then, the Fréchet derivative $\mathcal{S}'(\delta)\eta$ of $\mathcal{S}$ at δ in the direction η exists and is given by $u'|_K \in H^1(\Omega_H^\delta)$, where u' satisfies*

$$\Delta u' + k^2 u' = 0 \qquad \text{in } \Omega_H^\delta, \tag{5.17a}$$

$$\partial_{x_2} u' = \mathcal{T}^+ u' \qquad \text{on } \Gamma_H, \tag{5.17b}$$

$$u' = -\frac{\eta}{\sqrt{1 + (\zeta^\delta)'^2}} \partial_n u^\delta = -\eta \partial_{x_2} u^\delta \qquad \text{on } \Gamma^\delta \tag{5.17c}$$

with $\mathcal{T}^+$ defined as in (2.29) and the total field $u^\delta \in \widetilde{H}^2(\Omega_H^\delta)$ satisfying

$$\Delta u^\delta + k^2 u^\delta = 0 \qquad \text{in } \Omega_H^\delta \,, \tag{5.18a}$$

$$u^\delta = 0 \qquad \text{on } \Gamma^\delta \,, \tag{5.18b}$$

$$(\partial_{x_2} - \mathcal{T}^+) u^\delta = (\partial_{x_2} - \mathcal{T}^+) u^i \qquad \text{on } \Gamma_H \tag{5.18c}$$

for a non-periodic incident field $u^i \in H^2(\Omega_H^\delta)$.

Proof. Let u' be the solution of Problem (5.17) and the perturbation η be such that $\|\zeta^\delta + \eta\|_\infty < H$ (depicted in Figure 5.2). In this case, we can define the domain

$$\Omega_H^{\delta+\eta} := \{x \in \mathbb{R}^2 : \zeta^\delta + \eta < x_2 < H\} \,.$$

The solution of the scattering problem (5.18) corresponding to the perturbation $\eta+\delta$ is denoted by $u^{\delta+\eta}\in\widetilde{H}^2(\Omega_H^{\delta+\eta})$.

From the definition of the Fréchet derivative in (5.8), it is sufficient to show that

$$\left\|u^{\delta+\eta}-u^{\delta}-u'\right\|_{H^1(K)}=\mathcal{O}(\|\eta\|_{1,\infty}^2)\quad\text{as }\|\eta\|_{1,\infty}\to 0\,. \tag{5.19}$$

For this purpose, we first define the diffeomorphism $\Psi^\eta\colon\Omega_H^\delta\to\Omega_H^{\delta+\eta}$ as in (4.4) for the perturbation η. This diffeomorphism depends on $\beta_h^\eta(x)$, which is defined by replacing ζ^{per} with ζ^δ in (4.5). Moreover, choosing the parameter h as in the proof of the previous theorem yields $\Psi^\eta|_K=I$. In this case, defining $\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}:=u^{\delta+\eta}\circ\Psi^\eta\in\widetilde{H}^2(\Omega_H^\delta)$ and considering the fact that $u^{\delta+\eta}|_K=\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}|_K$, we have

$$\left\|u^{\delta+\eta}-u^{\delta}-u'\right\|_{H^1(K)}=\left\|\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}-u^{\delta}-u'\right\|_{H^1(K)}.$$

Since $\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}$ satisfies Problem (5.18), its variational form can be written as

$$a^{\delta+\eta}(\widetilde{u}_{\mathrm{tra}}^{\delta+\eta},z):=\left\langle(\partial_{x_2}-\mathcal{T}^+)u^i\circ\Psi^\eta,\overline{z}\right\rangle_{\Gamma_H}\text{ for all }z\in\widetilde{H}^2(\Omega_H^\delta)\,,$$

where

$$\begin{aligned}a^{\delta+\eta}(\phi,\psi)=&\left\langle\nabla\phi,\overline{\nabla\psi}\right\rangle_{\Omega_H^\delta}-k^2\left\langle\nabla\phi,\overline{\psi}\right\rangle_{\Omega_H^\delta}-\left\langle\mathcal{T}^+\phi,\overline{\psi}\right\rangle_{\Gamma_H}\\&+\left\langle(A^\eta-I)\nabla\phi,\overline{\nabla\psi}\right\rangle_{\Omega_H^\delta}-k^2\left\langle(c^\eta-1)\phi,\overline{\psi}\right\rangle_{\Omega_H^\delta}\end{aligned}$$

with the coefficients c^η and A^η defined as in (4.7) from the diffeomorphism Ψ^η.

Defining $v^\eta(x)=\eta(x_1)\beta_h^\eta(x)\partial_{x_2}u^\delta(x)$, it suffices to prove

$$\left\|\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}-u^{\delta}-(u'+v^\eta)\right\|_{H^1(\Omega_H^\delta)}=\mathcal{O}(\|\eta\|_{1,\infty}^2)\quad\text{as }\|\eta\|_{1,\infty}\to 0\,,$$

since $v^\eta=0$ on K. To be able to apply the FB transform, we further transform these functions to the periodic domain Ω_H^{per} using the diffeomorphism Ψ^δ from (4.4). Considering

$$u_{\mathrm{tra}}^{\delta+\eta}:=\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}\circ\Psi^\delta\in\widetilde{H}^2(\Omega_H^{\mathrm{per}})\,,$$

we see that its FB transform, denoted by $w^{\delta+\eta}:=\mathcal{J}u_{\mathrm{tra}}^{\delta+\eta}\in L^2(\Lambda;\widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi}))$, satisfies

$$a_\alpha^{\delta+\eta}(w^{\delta+\eta},z_{\mathrm{tra}})=\left\langle(\partial_{x_2}-\mathcal{T}_\alpha^+)\mathcal{J}u_{\mathrm{tra}}^i(\alpha),\overline{z_{\mathrm{tra}}}\right\rangle_{\Gamma_H^{2\pi}}, \tag{5.20}$$

for all $z_{\mathrm{tra}}\in\widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$ and almost all $\alpha\in\Lambda$, where

$$\begin{aligned}a_\alpha^{\delta+\eta}(w^{\delta+\eta},z_{\mathrm{tra}})=&\,a_\alpha^{\delta}(w^{\delta+\eta},z_{\mathrm{tra}})\\&+\left\langle c^\delta(\nabla\Psi^\delta)^{-1}(A^\eta-I)(\nabla\Psi^\delta)^{-\top}\nabla_x(\mathcal{J}^{-1}w^{\delta+\eta}),\overline{\nabla_x(z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}}\\&-k^2\left\langle c^\delta(c^\eta-1)\mathcal{J}^{-1}w^{\delta+\eta},\overline{z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}},\end{aligned} \tag{5.21}$$

with a_α^δ as defined in the *Perturbed Problem* and $c^\delta=|\det\nabla\Psi^\delta|$.

Similarly as in the proof of the previous theorem, using the mapping property of the FB transform, there exists a constant C such that

$$\left\|u_{\mathrm{tra}}^{\delta+\eta}-u_{\mathrm{tra}}^{\delta}-(u'_{\mathrm{tra}}+v_{\mathrm{tra}}^{\eta})\right\|_{H^1(\Omega_H^{\mathrm{per}})}\leq C\left\|w^{\delta+\eta}-w^{\delta}-\mathcal{J}(u'_{\mathrm{tra}}+v_{\mathrm{tra}}^{\eta})\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))},$$

where $u_{\mathrm{tra}}^{\delta}\coloneqq u^{\delta}\circ\Psi^{\delta}$, $u'_{\mathrm{tra}}\coloneqq u'\circ\Psi^{\delta}$, $v_{\mathrm{tra}}^{\eta}\coloneqq v^{\eta}\circ\Psi^{\delta}$ and $w^{\delta}\coloneqq\mathcal{J}u_{\mathrm{tra}}^{\delta}$. Using (2.15) for each $\alpha\in\Lambda$, we obtain

$$\left\|w^{\delta+\eta}-w^{\delta}-\mathcal{J}(u'_{\mathrm{tra}}+v_{\mathrm{tra}}^{\eta})\right\|_{L^2(\Lambda;\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi}))}\leq\sup_{\substack{z\in\widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})\\ z\neq 0}}\frac{a_\alpha^{\delta}(w^{\delta+\eta}-w^{\delta}-\mathcal{J}(u'_{\mathrm{tra}}+v_{\mathrm{tra}}^{\eta}),z)}{C_{\mathrm{infsup}}\|z\|_{H^1(\Omega_H^{2\pi})}},$$

where C_{infsup} is the inf-sup constant. Then, it is enough to show that for all $z_{\mathrm{tra}}\in\widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$ and almost all $\alpha\in\Lambda$, we have

$$a_\alpha^{\delta}(w^{\delta+\eta}-w^{\delta}-\mathcal{J}(u'_{\mathrm{tra}}+v_{\mathrm{tra}}^{\eta}),z_{\mathrm{tra}})=\mathcal{O}(\|\eta\|_{1,\infty}^2)\quad\text{as }\|\eta\|_{1,\infty}\to 0\,.$$

Since the right-hand side of equation (5.20) and the *Perturbed Problem* are equal, we obtain $a_\alpha^{\delta}(w^{\delta},z_{\mathrm{tra}})=a_\alpha^{\delta+\eta}(w^{\delta+\eta},z_{\mathrm{tra}})$. Therefore, we can compute for $z_{\mathrm{tra}}\in\widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$

$$\begin{aligned}&a_\alpha^{\delta}(w^{\delta+\eta}-w^{\delta}-\mathcal{J}(u'_{\mathrm{tra}}+v_{\mathrm{tra}}^{\eta}),z_{\mathrm{tra}})\\&\quad=-\left(a_\alpha^{\delta+\eta}(w^{\delta+\eta},z_{\mathrm{tra}})-a_\alpha^{\delta}(w^{\delta+\eta},z_{\mathrm{tra}})\right)-a_\alpha^{\delta}(\mathcal{J}(u'_{\mathrm{tra}}+v_{\mathrm{tra}}^{\eta}),z_{\mathrm{tra}})\,.\end{aligned}\tag{5.22}$$

We start with computing the first term in equation (5.22). Using the definition of the sesquilinear form $a_\alpha^{\delta+\eta}$ in (5.21) and recalling $u_{\mathrm{tra}}^{\delta+\eta}=\mathcal{J}^{-1}w^{\delta+\eta}$, we obtain

$$\begin{aligned}a_\alpha^{\delta+\eta}(w^{\delta+\eta},z_{\mathrm{tra}})-a_\alpha^{\delta}(w^{\delta+\eta},z_{\mathrm{tra}})&=\left\langle c^{\delta}(\nabla\Psi^{\delta})^{-1}(A^{\eta}-I)(\nabla\Psi^{\delta})^{-\top}\nabla u_{\mathrm{tra}}^{\delta+\eta},\overline{\nabla(z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{2\pi}}\\&\quad-k^2\left\langle c^{\delta}(c^{\eta}-1)u_{\mathrm{tra}}^{\delta+\eta},\overline{z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{2\pi}}.\end{aligned}$$

Changing the variables back to the perturbed cell $\Omega_H^{\delta,2\pi}\coloneqq\Psi^{\delta}(\Omega_H^{2\pi})$ and recalling the functions $\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}=u_{\mathrm{tra}}^{\delta+\eta}\circ(\Psi^{\delta})^{-1}$ and $z=z_{\mathrm{tra}}\circ(\Psi^{\delta})^{-1}$, we can see that

$$\begin{aligned}a_\alpha^{\delta+\eta}(w^{\delta+\eta},z_{\mathrm{tra}})-a_\alpha^{\delta}(w^{\delta+\eta},z_{\mathrm{tra}})&=\left\langle(A^{\eta}-I)\nabla\widetilde{u}_{\mathrm{tra}}^{\delta+\eta},\overline{\nabla(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\right\rangle_{\Omega_H^{\delta,2\pi}}\\&\quad-k^2\left\langle(c^{\eta}-1)\widetilde{u}_{\mathrm{tra}}^{\delta+\eta},\overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\right\rangle_{\Omega_H^{\delta,2\pi}}.\end{aligned}\tag{5.23}$$

Using the estimates (A.4) and (A.5) leads to $c^{\eta}-1=\eta\partial_{x_2}\beta_h^{\eta}$ and

$$\begin{aligned}&\left((A^{\eta}-I)\nabla\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}\right)\cdot\overline{\nabla(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\\&\quad=-\operatorname{div}\left\{v^{\eta}\overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}+\eta\beta_h^{\eta}\nabla\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}\overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}-\eta\beta_h^{\eta}e_2(\nabla\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}\cdot\overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})})\right\}\\&\qquad+v^{\eta}\overline{\Delta_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}+\eta\beta_h^{\eta}\Delta\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}\overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}+\mathcal{O}(\|\eta\|_{1,\infty}^2)\quad\text{as }\|\eta\|_{1,\infty}\to 0\,.\end{aligned}$$

Substituting the terms above into (5.23), applying the divergence theorem and using the fact

that $\beta_h^\eta = 0$ on $\Gamma_H^{2\pi}$ and $\beta_h^\eta = 1$ on $\Gamma^{\delta,2\pi}$, we obtain

$$\begin{aligned}
&a_\alpha^{\delta+\eta}(w^{\delta+\eta}, z_{\mathrm{tra}}) - a_\alpha^{\delta}(w^{\delta+\eta}, z_{\mathrm{tra}}) \\
&\quad = -\Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma_H^{2\pi}} - \Big\langle \eta\beta_h^\eta \nabla \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma_H^{2\pi}} \\
&\qquad + \Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} + \Big\langle \eta\beta_h^\eta \nabla \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} \\
&\qquad + \Big\langle v^\eta, \overline{\Delta_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} + \Big\langle \eta\beta_h^\eta \Delta \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} \\
&\qquad - k^2\Big\langle \eta(\partial_{x_2}\beta_h^\eta)\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \\
&\quad = -\Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma_H^{2\pi}} + \Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} \\
&\qquad + \Big\langle \eta\nabla_x \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} \\
&\qquad + \Big\langle v^\eta, \overline{\Delta_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} + \Big\langle \eta\beta_h^\eta \Delta \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} \\
&\qquad - k^2\Big\langle \eta(\partial_{x_2}\beta_h^\eta)\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0\,,
\end{aligned}$$

where $n = (n_1, n_2)^\top$ denotes the unit outward normal vector. Applying Green's first identity to the term $\Big\langle v^\eta, \overline{\Delta(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}}$ leads to

$$\begin{aligned}
a_\alpha^{\delta+\eta}(w^{\delta+\eta}, z_{\mathrm{tra}}) - a_\alpha^{\delta}(w^{\delta+\eta}, z_{\mathrm{tra}}) &= -\Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma_H^{2\pi}} + \Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} \\
&\quad + \Big\langle \eta\nabla_x \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} \\
&\quad - \Big\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} + \Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma_H^{2\pi}} \\
&\quad - \Big\langle nv^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} + \Big\langle \eta\beta_h^\eta \Delta_x \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} \\
&\quad - k^2\Big\langle \eta(\partial_{x_2}\beta_h^\eta)\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2)\,.
\end{aligned}$$

Some boundary terms cancel out and we are able to further simplify the equation above using the fact that $\eta\Delta\widetilde{u}_{\mathrm{tra}}^{\delta+\eta} = -\eta k^2 \widetilde{u}_{\mathrm{tra}}^{\delta+\eta} + \mathcal{O}(\|\eta\|_{1,\infty}^2)$, as $\|\eta\|_\infty \to 0$ (see Remark A.4). This yields

$$\begin{aligned}
&a_\alpha^{\delta+\eta}(w^{\delta+\eta}, z_{\mathrm{tra}}) - a_\alpha^{\delta}(w^{\delta+\eta}, z_{\mathrm{tra}}) \\
&\quad = \Big\langle \eta\nabla_x \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Gamma^{\delta,2\pi}} - \Big\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} \\
&\qquad - k^2\Big\langle \eta\beta_h^\eta \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} - k^2\Big\langle \eta(\partial_{x_2}\beta_h^\eta)\widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}} \\
&\qquad + \mathcal{O}(\|\eta\|_{1,\infty}^2) \text{ as } \|\eta\|_{1,\infty} \to 0\,.
\end{aligned} \tag{5.24}$$

Due to the fact that

$$n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1}) = e_1(n_1\partial_{x_2} z\mathrm{e}^{\mathrm{i}\alpha x_1} - n_2\partial_{x_1} z\mathrm{e}^{\mathrm{i}\alpha x_1})\,,$$

the first term in (5.24) can be simplified as follows

$$\begin{aligned}
&\Big\langle \eta \nabla_x \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{n\partial_{x_2}(z\mathrm{e}^{\mathrm{i}\alpha x_1}) - n_2 \nabla_x (z\mathrm{e}^{\mathrm{i}\alpha x_1})} \Big\rangle_{\Gamma^{\delta,2\pi}} \\
&\quad = \Big\langle \eta \partial_{x_1} \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{n_1 \partial_{x_2} z\mathrm{e}^{\mathrm{i}\alpha x_1} - n_2 \partial_{x_1} z\mathrm{e}^{\mathrm{i}\alpha x_1}} \Big\rangle_{\Gamma^{\delta,2\pi}} \\
&\quad = \Big\langle \eta \partial_{x_1} \widetilde{u}_{\mathrm{tra}}^{\delta+\eta}, \overline{\nabla_x (z\mathrm{e}^{\mathrm{i}\alpha x_1}) \cdot n^{\perp}} \Big\rangle_{\Gamma^{\delta,2\pi}} = 0\,,
\end{aligned}$$

where the last equality holds, since $z\mathrm{e}^{\mathrm{i}\alpha x_1}\big|_{\Gamma^{\delta,2\pi}}$ is constantly zero and hence its tangential derivative is zero. Substituting the above result into (5.24), we obtain

$$\begin{aligned}
&a_\alpha^{\delta+\eta}(w^{\delta+\eta}, z_{\mathrm{tra}}) - a_\alpha^{\delta}(w^{\delta+\eta}, z_{\mathrm{tra}}) \\
&\quad = -\Big\langle \nabla_x v^\eta, \overline{\nabla_x (z\mathrm{e}^{\mathrm{i}\alpha x_1})} \Big\rangle_{\Omega_H^{\delta,2\pi}} - k^2 \Big\langle \eta \beta_h^\eta \mathcal{J}^{-1} w^\delta, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1} \partial_{x_2} z} \Big\rangle_{\Omega_H^{\delta,2\pi}} \\
&\qquad - k^2 \Big\langle \eta (\partial_{x_2} \beta_h^\eta) \mathcal{J}^{-1} w^\delta, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1} z} \Big\rangle_{\Omega_H^{\delta,2\pi}} + \mathcal{O}(\|\eta\|_{1,\infty}^2) \quad \text{as } \|\eta\|_{1,\infty} \to 0\,.
\end{aligned} \tag{5.25}$$

Now, we are going to compute the second term of (5.22). For all $z_{\mathrm{tra}} \in \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$, we have

$$a_\alpha^\delta(\mathcal{J}(u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}}), z_{\mathrm{tra}}) = a_\alpha^\delta(\mathcal{J}u'_{\mathrm{tra}}, z_{\mathrm{tra}}) + a_\alpha^\delta(\mathcal{J}v^\eta_{\mathrm{tra}}, z_{\mathrm{tra}})\,, \tag{5.26}$$

where a_α^δ is defined as in the *Perturbed Problem*. Recall that $u'_{\mathrm{tra}} = u' \circ \Psi^\delta \in H^2(\Omega_H^{\mathrm{per}})$, with u' satisfying problem (5.17). Then, by applying the FB transform to u'_{tra}, we can see that $\mathcal{J}u'_{\mathrm{tra}}$ satisfies

$$a_\alpha^\delta(\mathcal{J}u'_{\mathrm{tra}}, z_{\mathrm{tra}}) = 0 \quad \text{for all } z_{\mathrm{tra}} \in \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi}) \text{ and almost all } \alpha \in \Lambda\,,$$

where a_α^δ is defined in the *Perturbed Problem*. Considering this fact, equation (5.26) can be simplified as

$$a_\alpha^\delta(\mathcal{J}(u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}}), z_{\mathrm{tra}}) = a_\alpha^\delta(\mathcal{J}v^\eta_{\mathrm{tra}}, z_{\mathrm{tra}})\,.$$

Therefore, it only remains to compute $a_\alpha^\delta(\mathcal{J}v^\eta_{\mathrm{tra}}, z_{\mathrm{tra}})$. From the definition of $v^\eta_{\mathrm{tra}} = \eta\beta_h^\eta \partial_{x_2} u^\delta_{\mathrm{tra}}$, it follows that v^η_{tra} is compactly supported in $\Omega_H^{2\pi}$, since η is a local perturbation supported in this cell. This yields $\mathcal{J}v^\eta_{\mathrm{tra}} = v^\eta_{\mathrm{tra}} \exp(-\mathrm{i}\alpha x_1)$. Moreover, $v^\eta_{\mathrm{tra}}|_{\Gamma_H^{2\pi}} = 0$, as $\beta_h^\eta(x) = 0$ on $\Gamma_H^{2\pi}$. Hence, using $\nabla_x(\mathcal{J}v^\eta_{\mathrm{tra}}) = (\nabla_x v^\eta_{\mathrm{tra}} - \mathrm{i}\alpha e_1 v^\eta_{\mathrm{tra}})\mathrm{e}^{-\mathrm{i}\alpha x_1}$ and $\mathcal{T}_\alpha^+ \mathcal{J}v^\eta_{\mathrm{tra}} = 0$, we obtain

$$\begin{aligned}
&a_\alpha^\delta(\mathcal{J}(u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}}), z_{\mathrm{tra}}) \\
&\quad = \Big\langle \nabla_x(\mathcal{J}v^\eta_{\mathrm{tra}}), \overline{\nabla_x z_{\mathrm{tra}}} \Big\rangle_{\Omega_H^{2\pi}} - 2\mathrm{i}\alpha \langle \partial_{x_1}(\mathcal{J}v^\eta_{\mathrm{tra}}), \overline{z_{\mathrm{tra}}} \rangle_{\Omega_H^{2\pi}} - (k^2 - \alpha^2) \langle \mathcal{J}v^\eta_{\mathrm{tra}}, \overline{z_{\mathrm{tra}}} \rangle_{\Omega_H^{2\pi}} \\
&\qquad + \Big\langle (A^\delta - I)\nabla v^\eta_{\mathrm{tra}}, \overline{\nabla(z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1})} \Big\rangle_{\Omega_H^{2\pi}} - k^2 \Big\langle (c^\delta - 1) v^\eta_{\mathrm{tra}}, \overline{z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1}} \Big\rangle_{\Omega_H^{2\pi}} - \Big\langle \mathcal{T}_\alpha^+ \mathcal{J}v^\eta_{\mathrm{tra}}, \overline{z_{\mathrm{tra}}} \Big\rangle_{\Gamma_H^{2\pi}} \\
&\quad = -\mathrm{i}\alpha \Big\langle v^\eta_{\mathrm{tra}}, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1} \partial_{x_1} z_{\mathrm{tra}}} \Big\rangle_{\Omega_H^{2\pi}} - \mathrm{i}\alpha \Big\langle \partial_{x_1} v^\eta_{\mathrm{tra}}, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1} z_{\mathrm{tra}}} \Big\rangle_{\Omega_H^{2\pi}} - \alpha^2 \Big\langle v^\eta_{\mathrm{tra}}, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1} z_{\mathrm{tra}}} \Big\rangle_{\Omega_H^{2\pi}} \\
&\qquad + \Big\langle A^\delta \nabla v^\eta_{\mathrm{tra}}, \overline{\nabla(z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1})} \Big\rangle_{\Omega_H^{2\pi}} - k^2 \Big\langle c^\delta v^\eta_{\mathrm{tra}}, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1} z_{\mathrm{tra}}} \Big\rangle_{\Omega_H^{2\pi}}\,.
\end{aligned}$$

Using

$$\begin{aligned}\mathrm{i}\alpha\Big\langle v^\eta_{\mathrm{tra}}, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1}\partial_{x_1} z_{\mathrm{tra}}}\Big\rangle_{\Omega_H^{2\pi}} + \alpha^2\Big\langle v^\eta_{\mathrm{tra}}, \overline{\mathrm{e}^{\mathrm{i}\alpha x_1} z_{\mathrm{tra}}}\Big\rangle_{\Omega_H^{2\pi}} &= \mathrm{i}\alpha\Big\langle v^\eta_{\mathrm{tra}}, \overline{\partial_{x_1}(z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{2\pi}} \\ &= -\mathrm{i}\alpha\Big\langle \partial_{x_1} v^\eta_{\mathrm{tra}}, \overline{z_{\mathrm{tra}}\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{2\pi}},\end{aligned}$$

and changing the variables back to the perturbed cell $\Omega_H^{\delta,2\pi}$, we get

$$a^\delta_\alpha(\mathcal{J}(u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}}), z_{\mathrm{tra}}) = \Big\langle \nabla v^\eta, \overline{\nabla(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} - k^2\Big\langle v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}}. \tag{5.27}$$

By substituting equations (5.25) and (5.27) into (5.22), we obtain

$$\begin{aligned}&a^\delta_\alpha(w^{\delta+\eta} - w^\delta - \mathcal{J}(u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}}), z_{\mathrm{tra}}) \\ &\quad= \Big\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} + k^2\Big\langle \eta\beta^\eta_h \widetilde{u}^{\delta+\eta}_{\mathrm{tra}}, \overline{\partial_{x_2} z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}} + k^2\Big\langle \eta(\partial_{x_2}\beta^\eta_h)\widetilde{u}^{\delta+\eta}_{\mathrm{tra}}, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}} \\ &\qquad- \Big\langle \nabla_x v^\eta, \overline{\nabla_x(z\mathrm{e}^{\mathrm{i}\alpha x_1})}\Big\rangle_{\Omega_H^{\delta,2\pi}} + k^2\Big\langle v^\eta, \overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\Big\rangle_{\Omega_H^{\delta,2\pi}} + \mathcal{O}(\|\eta\|^2_{1,\infty}) \quad \text{as } \|\eta\|_{1,\infty} \to 0.\end{aligned}$$

Clearly the terms which do not depend on k cancel each other out and we are left with

$$\begin{aligned}a^\delta_\alpha(w^{\delta+\eta} - w^\delta - \mathcal{J}(u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}}), z_{\mathrm{tra}}) &= k^2\int_{\Omega_H^{\delta,2\pi}} \partial_{x_2}(\eta\beta^\eta_h \widetilde{u}^{\delta+\eta}_{\mathrm{tra}}\overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}})\,\mathrm{d}x + \mathcal{O}(\|\eta\|^2_{1,\infty}) \\ &= k^2\int_{\Omega_H^{\delta,2\pi}} \operatorname{div}(e_2\eta\beta^\eta_h \widetilde{u}^{\delta+\eta}_{\mathrm{tra}}\overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}})\,\mathrm{d}x + \mathcal{O}(\|\eta\|^2_{1,\infty}) \\ &= k^2\int_{\Gamma_H^{2\pi}} n_2\eta\beta^\eta_h \widetilde{u}^{\delta+\eta}_{\mathrm{tra}}\overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\,\mathrm{d}s \\ &\quad - k^2\int_{\Gamma^{\delta,2\pi}} n_2\eta\beta^\eta_h \widetilde{u}^{\delta+\eta}_{\mathrm{tra}}\overline{z\mathrm{e}^{\mathrm{i}\alpha x_1}}\,\mathrm{d}s + \mathcal{O}(\|\eta\|^2_{1,\infty}).\end{aligned}$$

Using the fact that $z = 0$ on $\Gamma^{\delta,2\pi}$ and $\beta^\eta_h = 0$ on $\Gamma_H^{2\pi}$, we conclude for all $z_{\mathrm{tra}} \in \widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$

$$a^\delta_\alpha(w^{\delta+\eta} - w^\delta - \mathcal{J}(u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}}), z_{\mathrm{tra}}) = \mathcal{O}(\|\eta\|^2_{1,\infty}) \quad \text{as } \|\eta\|_{1,\infty} \to 0.$$

As $\widetilde{H}^2_{\mathrm{per}}(\Omega_H^{2\pi})$ is dense in $\widetilde{H}^1_{\mathrm{per}}(\Omega_H^{2\pi})$, we have

$$\Big\|u^{\delta+\eta}_{\mathrm{tra}} - u^\delta_{\mathrm{tra}} - (u'_{\mathrm{tra}} + v^\eta_{\mathrm{tra}})\Big\|_{H^1(\Omega_H^{\mathrm{per}})} = \mathcal{O}(\|\eta\|^2_{1,\infty}) \quad \text{as } \|\eta\|_{1,\infty} \to 0.$$

Finally, using the mapping property of the FB transform and transforming all functions back to Ω^δ_H, we complete the proof because $v^\eta_{\mathrm{tra}} = 0$ on the compact set K. □

In the following theorems, we aim to show some properties of the operator $\mathcal{S}'(\delta)$.

Theorem 5.7. *The operator $\mathcal{S}'(\delta)\colon X \to L^2(\Gamma_H^{2\pi})$ is locally compact for sufficiently small $\delta \in X$.*

Proof. Let $\delta \in X$ be sufficiently small. We need to show that for any bounded set $U \subset X$, $\mathcal{S}'(\delta)U$ is relatively compact in $L^2(\Gamma_H^{2\pi})$. We first prove that $\mathcal{S}'(\delta)\colon X \to H^{1/2}(\Gamma_H^{2\pi})$ is continuous. Afterwards, we use the compact Sobolev embedding theorem to complete the proof.

To show the continuity of $\mathcal{S}'(\delta)$, we consider a sufficiently small ball B around zero such that $\mathcal{S}(\delta+\eta)$ exists for every $\eta \in B$. Using Theorem 5.6, there exists a constant C_1 such that

$$\|\mathcal{S}(\delta+\eta) - \mathcal{S}(\delta) - \mathcal{S}'(\delta)\eta\|_{H^{1/2}(\Gamma_H^{2\pi})} \leq C_1 \|\eta\|_{1,\infty}^2 .$$

Moreover, from the continuity of the scattering operator $\mathcal{S}$ (shown in Theorem 5.3), we have

$$\|\mathcal{S}(\delta+\eta) - \mathcal{S}(\delta)\|_{H^{1/2}(\Gamma_H^{2\pi})} \leq C_2 \|\eta\|_{1,\infty} .$$

Using the triangle inequality, we can write

$$\begin{aligned}
\|\mathcal{S}'(\delta)\eta\|_{H^{1/2}(\Gamma_H^{2\pi})} &= \|\mathcal{S}'(\delta)\eta + \mathcal{S}(\delta+\eta) - \mathcal{S}(\delta) - \mathcal{S}(\delta+\eta) + \mathcal{S}(\delta)\|_{H^{1/2}(\Gamma_H^{2\pi})} \\
&\leq \|\mathcal{S}'(\delta)\eta - \mathcal{S}(\delta+\eta) + \mathcal{S}(\delta)\|_{H^{1/2}(\Gamma_H^{2\pi})} + \|\mathcal{S}(\delta+\eta) - \mathcal{S}(\delta)\|_{H^{1/2}(\Gamma_H^{2\pi})} \\
&\leq C_1 \|\eta\|_{1,\infty}^2 + C_2 \|\eta\|_{1,\infty} \leq C_3 \|\eta\|_{1,\infty} ,
\end{aligned}$$

where the last inequality holds since η is in the bounded set B. Using the compact embedding theorem in the fractional Sobolev spaces (see [32, Cor. 7.2]), we obtain that the embedding $\mathscr{I}\colon H^{1/2}(\Gamma_H^{2\pi}) \hookrightarrow L^2(\Gamma_H^{2\pi})$ is compact. Then, the set $\mathcal{S}'(\delta)U$ is embedded compactly into $L^2(\Gamma_H^{2\pi})$. □

Theorem 5.8. *The operator $\mathcal{S}'(\delta)$ is injective for sufficiently small $\delta \in X$.*

Proof. The proof follows the approach used in the periodic case [56, Cor. 3.2] and the case of bounded scatterers [68, Lem. 2.2].

Let u' satisfy problem (5.17). Due to the linearity of the operator $\mathcal{S}'(\delta)$, it is sufficient to prove that the kernel is trivial. Therefore, we consider $\eta \in X$ such that the derivative of the scattering operator $\mathcal{S}'(\delta)\eta = u'|_{\Gamma_H} = 0$. Using (5.17b) and the linearity of $\mathcal{T}^+$, gives us $\partial_n u'|_{\Gamma_H} = \partial_{x_2} u'|_{\Gamma_H} = 0$. According to Holmgren's theorem (see [31, Thm. 2.3]), since u' and its normal derivative are zero on an open subset of the boundary, we conclude that $u' = 0$ in Ω_H^δ.

Substituting $u' = 0$ in the boundary condition (5.17c), we obtain $\eta \partial_{x_2} u^\delta = 0$ on Γ^δ. Since the total field u^δ is not the trivial solution in Ω_H^δ and satisfies $u^\delta \neq 0$ on Ω_H^δ, again using Holmgren's theorem we conclude that $\partial_{x_2} u^\delta$ is different from the zero function in every relatively open subset of Γ^δ, which yields $\eta = 0$. □

Theorem 5.9. *The inverse of the operator $\mathcal{S}'(\delta)\colon X \to L^2(\Gamma_H^{2\pi})$ exists and is unbounded.*

Proof. Let U be the range of $\mathcal{S}'(\delta)$. Since $\mathcal{S}'(\delta)$ is injective (see Theorem 5.8), the operator $\mathcal{S}'(\delta)\colon X \to U$ is bijective. Hence its inverse is well defined. According to the proof of [31, Thm. 4.2], as the linear operator $\mathcal{S}'(\delta)$ is compact and X is infinite dimensional, the inverse is not bounded. □

5.3. Regularization, Discretization and Reconstruction

As shown in Theorem 5.4, the inverse problem (5.2) is ill-posed. To obtain a stable approximation of the solution, we are going to regularize the problem (5.2) and then apply the Gauss–Newton

method as described for instance in [94, Sec. 10.2]. However, other iterative schemes, for example, the nonlinear Landweber iteration [64, Sec. 2] and the inexact Newton method [87, 95] can also be applied to solve the mentioned problem.

To regularize (5.2), we add a penalty term to the objective function as follows

$$\delta^* \coloneqq \arg\min_{\delta \in X} \|\mathcal{S}(\delta) - D\|^2_{L^2(\Gamma_H^{2\pi})} + \left(\alpha_{\mathrm{reg}} \|\psi(\delta)\|_{L^2([-\pi,\pi])}\right)^2, \tag{5.28}$$

where α_{reg} is a constant called regularization parameter and $\psi(\delta)$ denotes the penalty term. A common choice for the penalty term is the curvature of the bottom surface Γ^δ. However, in our case, since the bottom surface is the graph of the function ζ^δ, it suffices to penalize the second order derivative of the perturbation. Therefore, we define the penalty term $\psi\colon X \to \mathbb{R}$ as

$$\psi(\delta) = \delta'' \,. \tag{5.29}$$

To find a minimizer δ^* for the functional (5.28) with the Gauss–Newton method, we need to compute the Fréchet derivative of the scattering operator $\mathcal{S}$ and the penalty function $\psi(\delta)$ for any sufficiently small perturbation $\delta \in X$. The former has been obtained in Theorem 5.5, and the latter will be computed in the next lemma.

Lemma 5.10. *Let $\eta \in X$ and ψ be defined as above. The Fréchet derivative of ψ, denoted by ψ', is given by*

$$\psi'(\delta)(\eta) = \eta'' \,.$$

Proof. Using the definition of the Fréchet derivative and substituting the expression ψ' defined as above leads to

$$\|\psi(\delta+\eta) - \psi(\delta) - \psi'(\delta)(\eta)\|_{L^2([-\pi,\pi])} = \|(\delta+\eta)'' - \delta'' - \eta''\|_{L^2([-\pi,\pi])} = 0 \,.$$

Since the second derivative is a linear operator, then ψ' is the Fréchet derivative of ψ. □

Now, we have all necessary tools to numerically reconstruct the perturbation δ, which satisfies the minimization problem (5.28) depending on the given near-field data D.

Discretization and Reconstruction: We discretize the space of the admissible perturbations X by the following space of splines

$$X_N \coloneqq \operatorname{span}\{\phi_1, \ldots, \phi_N\} \subset X \,,$$

where ϕ_j are cubic B-splines for a uniform subdivision of $[-\pi, \pi]$ and N denotes the number of splines. In the discrete setting, we therefore seek $\delta \in X_N$, that is

$$\delta(x) = \sum_{n=1}^{N} \delta_n \phi_n(x) \,,$$

where $\boldsymbol{\delta}_N \coloneqq (\delta_1, \ldots, \delta_N)^\top \in \mathbb{R}^N$.

In real applications, the near field is not available as an L^2-function; instead, M detectors are placed on $\Gamma_H^{2\pi}$. Therefore, we assume that M observations of the near field are available on equidistant points on the top surface. We still use the notation $D \in \mathbb{C}^M$ for these measurements.

To write the discrete form of the regularized inverse problem (5.28), we start by introducing the operator generating the curve from the coefficients

$$\begin{aligned} \mathcal{C} \colon \mathbb{R}^N &\to X_N \\ \boldsymbol{\delta}_N &\mapsto \delta \,. \end{aligned}$$

Moreover, we consider the projection operator $\mathcal{P} \colon L^2(\Gamma_H^{2\pi}) \to \mathbb{C}^M$, which models measurements of the total field on M observation points. Using these operators, the nonlinear scattering operator is discretized as $\mathcal{S}_N = \mathcal{P} \circ \mathcal{S} \circ \mathcal{C}$, mapping $\mathbb{R}^N$ to $\mathbb{C}^M$.

We can now derive the discrete version of the regularized optimization problem (5.28).

Discrete Inverse Problem: Find $\boldsymbol{\delta}_N^* \in \mathbb{R}^N$ such that

$$\boldsymbol{\delta}_N^* = \operatorname*{arg\,min}_{\boldsymbol{\delta}_N \in \mathbb{R}^N} \left(\|\mathcal{S}_N(\boldsymbol{\delta}_N) - D\|_2^2 + \|\alpha_{\mathrm{reg}} \psi(\mathcal{C}(\boldsymbol{\delta}_N))\|_{L^2([-\pi,\pi])}^2 \right) . \tag{5.30}$$

The term $\mathcal{R}_1(\boldsymbol{\delta}_N) \coloneqq \mathcal{S}_N(\boldsymbol{\delta}_N) - D$ is a complex vector of length M and the penalty function (5.29) can be written as

$$\psi(\mathcal{C}(\boldsymbol{\delta}_N))(x) = \sum_{n=1}^{N} \delta_n \phi_n''(x) \,,$$

where ϕ_n'' are piecewise linear functions. We can hence compute the regularization term in (5.30) by using a composite trapezoidal rule for the L^2-norm of ψ

$$\|\alpha_{\mathrm{reg}} \psi(\mathcal{C}(\boldsymbol{\delta}_N))\|_{L^2([-\pi,\pi])}^2 = \sum_{\ell=1}^{N+4} \left(\alpha_{\mathrm{reg}} \omega_\ell \sum_{n=1}^{N} \delta_n \phi_n''(x_\ell) \right)^2 ,$$

where the nodes and weights are given by $x_\ell = -\pi + 2\pi(\ell - 1)/(N+3)$ for $\ell = 1, \ldots, N+4$ and

$$\omega_\ell^2 = \begin{cases} \dfrac{\pi}{N+3} & \text{for } \ell = 1, N+4 \,, \\ \dfrac{2\pi}{N+3} & \text{for } \ell = 2, \ldots, N+3 \,. \end{cases}$$

Considering

$$\mathcal{R}_2(\boldsymbol{\delta}_N) \coloneqq \left(\alpha_{\mathrm{reg}} \omega_\ell \sum_{n=1}^{N} \delta_n \phi_n''(x_\ell) \right)_{\ell=1}^{N+4} ,$$

the objective function of (5.30) can be written as the scalar product

$$\left\langle \mathcal{R}_1(\boldsymbol{\delta}_N), \overline{\mathcal{R}_1(\boldsymbol{\delta}_N)} \right\rangle + \langle \mathcal{R}_2(\boldsymbol{\delta}_N), \mathcal{R}_2(\boldsymbol{\delta}_N) \rangle \,.$$

To simplify the numerical implementation, we split up the real and imaginary parts of $\mathcal{R}_1(\boldsymbol{\delta}_N)$

and consider the vector

$$\mathcal{R}(\boldsymbol{\delta}_N) \coloneqq \begin{bmatrix} \operatorname{Re}\mathcal{R}_1(\boldsymbol{\delta}_N) \\ \operatorname{Im}\mathcal{R}_1(\boldsymbol{\delta}_N) \\ \mathcal{R}_2(\boldsymbol{\delta}_N) \end{bmatrix} \in \mathbb{R}^{2M+N+4}\,.$$

This allows us to rewrite the objective function as

$$\begin{aligned}\left\langle \mathcal{R}_1(\boldsymbol{\delta}_N), \overline{\mathcal{R}_1(\boldsymbol{\delta}_N)} \right\rangle + \langle \mathcal{R}_2(\boldsymbol{\delta}_N), \mathcal{R}_2(\boldsymbol{\delta}_N)\rangle &= \langle \operatorname{Re}\mathcal{R}_1(\boldsymbol{\delta}_N), \operatorname{Re}\mathcal{R}_1(\boldsymbol{\delta}_N)\rangle \\ &\quad + \langle \operatorname{Im}\mathcal{R}_1(\boldsymbol{\delta}_N), \operatorname{Im}\mathcal{R}_1(\boldsymbol{\delta}_N)\rangle + \langle \mathcal{R}_2(\boldsymbol{\delta}_N), \mathcal{R}_2(\boldsymbol{\delta}_N)\rangle \\ &= \langle \mathcal{R}(\boldsymbol{\delta}_N), \mathcal{R}(\boldsymbol{\delta}_N)\rangle\,,\end{aligned}$$

hence we can rewrite the optimization problem (5.30) as $\arg\min_{\boldsymbol{\delta}_N \in \mathbb{R}^N} \|\mathcal{R}(\boldsymbol{\delta}_N)\|_2^2$. To solve this optimization problem, we apply the Gauss–Newton method, which is based on the linearization of the operator $\mathcal{R}$. To this end, we first introduce the discrete version of the Fréchet derivative of $\mathcal{S}_N$ as $\mathcal{S}_N'(\boldsymbol{\delta}_N) = \mathcal{P} \circ \mathcal{S}'(\mathcal{C}(\boldsymbol{\delta}_N)) \circ \mathcal{C}$, which is a linear mapping from $\mathbb{R}^N$ to $\mathbb{R}^M$. The Fréchet derivative of the operator $\mathcal{R}_1$ represents the Jacobian matrix $\mathbf{J}_{\mathcal{R}_1}(\boldsymbol{\delta}_N)$ with dimension $M \times N$ whose columns are obtained by

$$(\mathbf{J}_{\mathcal{R}_1}(\boldsymbol{\delta}_N))_{(:,n)} = \mathcal{P} \circ \mathcal{S}'(\mathcal{C}(\boldsymbol{\delta}_N))\phi_n \quad \text{for } n \in \{1, \dots, N\}\,.$$

We can now write the linearization of the operator $\mathcal{R}$ as follows

$$\mathcal{R}(\boldsymbol{\delta}_N + \boldsymbol{\eta}) = \mathcal{R}(\boldsymbol{\delta}_N) + \mathbf{J}_{\mathcal{R}}(\boldsymbol{\delta}_N)\boldsymbol{\eta} + \mathcal{O}(\boldsymbol{\eta}^2)\,,$$

where $\boldsymbol{\eta} \in \mathbb{R}^N$ and $\mathbf{J}_{\mathcal{R}}$ denotes the Jacobian matrix of $\mathcal{R}$ with dimension $(2M + N + 4) \times N$ whose n-th column is given by

$$(\mathbf{J}_{\mathcal{R}}(\boldsymbol{\delta}_N))_{(:,n)} \coloneqq \begin{bmatrix} \operatorname{Re}(\mathbf{J}_{\mathcal{R}_1}(\boldsymbol{\delta}_N))_{(:,n)} \\ \operatorname{Im}(\mathbf{J}_{\mathcal{R}_1}(\boldsymbol{\delta}_N))_{(:,n)} \\ (\mathbf{J}_{\mathcal{R}_2}(\boldsymbol{\delta}_N))_{(:,n)} \end{bmatrix}, \tag{5.31}$$

with $\mathbf{J}_{\mathcal{R}_2}(\boldsymbol{\delta}_N)\boldsymbol{\eta} = \mathcal{R}_2(\boldsymbol{\eta})$ as in Lemma 5.10. To apply the iterative Gauss–Newton method, we need to compute the update $\boldsymbol{\eta}$ with respect to the current reconstruction by

$$\boldsymbol{\eta} = -(\mathbf{J}_{\mathcal{R}}(\boldsymbol{\delta}_N))^{\dagger}\mathcal{R}(\boldsymbol{\delta}_N)\,,$$

where $(\mathbf{J}_{\mathcal{R}}(\boldsymbol{\delta}_N))^{\dagger} = -(\mathbf{J}_{\mathcal{R}}^{\top}(\boldsymbol{\delta}_N)\mathbf{J}_{\mathcal{R}}(\boldsymbol{\delta}_N))^{-1}\mathbf{J}_{\mathcal{R}}^{\top}(\boldsymbol{\delta}_N)$ is the pseudo-inverse of $\mathbf{J}_{\mathcal{R}}(\boldsymbol{\delta}_N)$.

We describe how to reconstruct the unknown perturbation $\boldsymbol{\delta}_N$, using the Gauss–Newton method, in Algorithm 4 (inspired by [72]).

In numerical experiments, it turns out that the regularization parameter has a significant effect on the accuracy of the reconstruction. More specifically, on one hand, if the value of this parameter is chosen to be too high, this leads to an inaccurate reconstruction due to the high impact of the penalty term in determining the update $\boldsymbol{\eta}$; on the other hand a regularization parameter that is too small is not sufficient for regularizing the optimization problem and due to its ill-posedness the iterative method may not converge. To mitigate the risk of an a priori

Algorithm 4: reconstruction of a local perturbation

Input: measured data D, stopping `tolerance`, regularization parameter α_{reg}

1 Choose an initial guess $\boldsymbol{\delta}_N$;
2 **for** $\ell = 1, \dots, \ell_{\max}$ **do**
3 Compute $\mathcal{S}(\boldsymbol{\delta}_N)$ by solving the direct problem as proposed in Chapter 4;
4 Compute the residual $\mathcal{R}(\boldsymbol{\delta}_N)$ from the solution $\mathcal{S}(\boldsymbol{\delta}_N)$ and measured data D;
  %Assemble the Jacobian matrix $\mathbf{J}_{\mathcal{R}}$ column-by-column.
  for $n = 1, \dots, N$ **do**
5 Use the direct solver to compute the Fréchet derivative $\mathcal{S}'(\boldsymbol{\delta}_N)\phi_n$ for the n-th B-spline ϕ_n;
6 Construct the n-th column of $\mathbf{J}_{\mathcal{R}}$ as in equation (5.31);
  %Determine the Gauss-Newton search direction.
7 $\boldsymbol{\eta} \leftarrow -\left(\mathbf{J}_{\mathcal{R}}^{\top}(\boldsymbol{\delta}_N)\mathbf{J}_{\mathcal{R}}(\boldsymbol{\delta}_N)\right)^{-1}\mathbf{J}_{\mathcal{R}}^{\top}(\boldsymbol{\delta}_N)\mathcal{R}(\boldsymbol{\delta}_N)$;
  %Calculate the movement of the reconstruction.
8 `Movement` $\leftarrow \|\boldsymbol{\eta}\|_2 / \|\boldsymbol{\delta}_N\|_2$;
9 **if** `Movement` $>$ `tolerance` **then**
10 Update $\boldsymbol{\delta}_N \leftarrow \boldsymbol{\delta}_N + \boldsymbol{\eta}$;
11 **else if** $\|\mathcal{R}_2\|_2 > \|\mathcal{R}_1\|_2$ **then**
    %The residual $\|\mathcal{R}\|_2$ is dominated by the penalty term.
12 Reduce $\alpha_{\text{reg}} \leftarrow \alpha_{\text{reg}}/2$;
13 **else**
    %The residual $\|\mathcal{R}\|_2$ is dominated by $\|\mathcal{R}_1\|_2$.
14 Stop the iterations;
15 **return** $\boldsymbol{\delta}_N$

selection of this parameter, we propose an a posteriori selection procedure. That is, we start with an a priori upper bound. In each iteration, we compute the contribution of the penalty term to the residual $\mathcal{R}$. If this exceeds half the norm of the residual, we reduce the regularization term by halving its value. The regularization parameter determined by the described selection procedure turns out to provide a good reconstruction of the perturbation.

For the stopping criterion, we define the *movement* as the ratio of the norms of the update $\boldsymbol{\eta}$ and the current reconstruction $\boldsymbol{\delta}_N$ (see Algorithm 4, line 8). The iterative method stops when the movement is less than a given tolerance and the regularization parameter is not updated.

5.4. NUMERICAL RESULTS

To illustrate the efficiency of the proposed reconstruction method, we focus here on the downward propagating Green's function as an incident field with the point source above the locally perturbed surface.

We consider the same examples as given in Section 4.6. The periodic functions

$$\zeta_1^{\text{per}}(x) = 1 + \frac{\cos(x)}{4}, \quad x \in \mathbb{R},$$
$$\zeta_2^{\text{per}}(x) = 1.5 + \frac{\sin(x)}{3} - \frac{\cos(x)}{4}, \quad x \in \mathbb{R},$$

are given whereas the perturbations

$$\delta_1(x) = \frac{1}{2}\exp\left(\frac{1}{x(x+2)}\right)\left(\cos\left(\frac{\pi(x+2)}{2}\right)+1\right)\chi_{[-2,0]}(x), \quad x \in \mathbb{R}$$
$$\delta_2(x) = \frac{3}{2}\exp\left(\frac{1}{x^2-1}\right)\sin(\pi(x+1))\chi_{[-1,1]}(x) \quad x \in \mathbb{R}.$$

are used to compute the measured data on the flat surface $\Gamma_H^{2\pi}$ using the direct solver from the previous chapter. The knowledge of the exact perturbation δ_1 and δ_2 allows us to report the error of the numerical reconstruction.

We set $H = 2.5$ and $k = 1.4$, and use the regularized Gauss–Newton method presented in Algorithm 4 with the following inputs.

- The dimension of X_N is $N = 30$.
- We consider $M = 60$ detectors on $\Gamma_H^{2\pi}$ for measuring the near-field data.
- The initial value of the regularization parameter is $\alpha_{\text{reg}} = 0.6$.
- The stopping `tolerance` is 10^{-3}.
- The direct solver proposed in Chapter 4 is used with the PML thickness $\lambda = 1.5$, the PML parameter $\rho = 20$ and the number of Floquet parameters $N_\alpha = 20$.
- As an initial guess, we choose $\boldsymbol{\delta}_N = 0$ corresponding to the periodic bottom surface.

To check the accuracy of the numerical reconstruction, we compute the following relative error

$$E_{\text{rec}} = \frac{\|\boldsymbol{\delta}_N - \boldsymbol{\delta}_{\text{exact}}\|_\infty}{\|\boldsymbol{\delta}_{\text{exact}}\|_\infty},$$

where $\boldsymbol{\delta}_{\text{exact}}$ and $\boldsymbol{\delta}_N$ denote the exact and numerical reconstructions.

Remark 5.11. In our case, exact solutions to the direct scattering problem or experimental measured data are not available. Therefore, we must solve the inverse problem typically based on synthetic near-field data, which is obtained by solving the direct problem. To avoid the inverse crime–the trivial inversion of a discretized problem (see [31, p. 179])–the synthetic near-field data must be generated with a direct solver that is independent of the inverse solver. Accordingly, we compute the near-field data using the exact DtN formulation and the proposed numerical scheme introduced in Chapter 3, employing twice as many discretization points as those used in the PML-based inverse solver.

We provide the numerical reconstructions from noise-free data for the perturbations δ_1 and δ_2 in Figures 5.3 and 5.4, with the point source located at $y = (-1, 2.5)^\top$ and $y = (0, 3)^\top$, respectively.

As shown in Figure 5.3, at iteration 8, the numerical reconstruction (red dashed line) is clearly approaching the exact perturbation δ_1 (blue line). At iteration 22, a good reconstruction of the perturbation δ_1 is obtained, where the maximum error E_{rec} is around 10^{-2} as seen in the left image of Figure 5.5. At this point, the algorithm stops, since the reconstruction does not improve beyond the required precision.

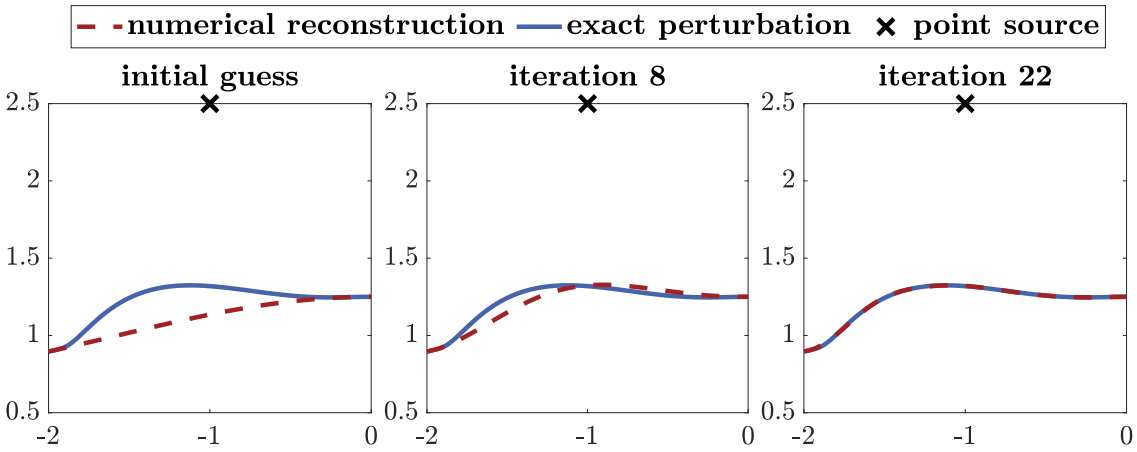


FIGURE 5.3. Numerical reconstruction of δ_1 from noise-free data.

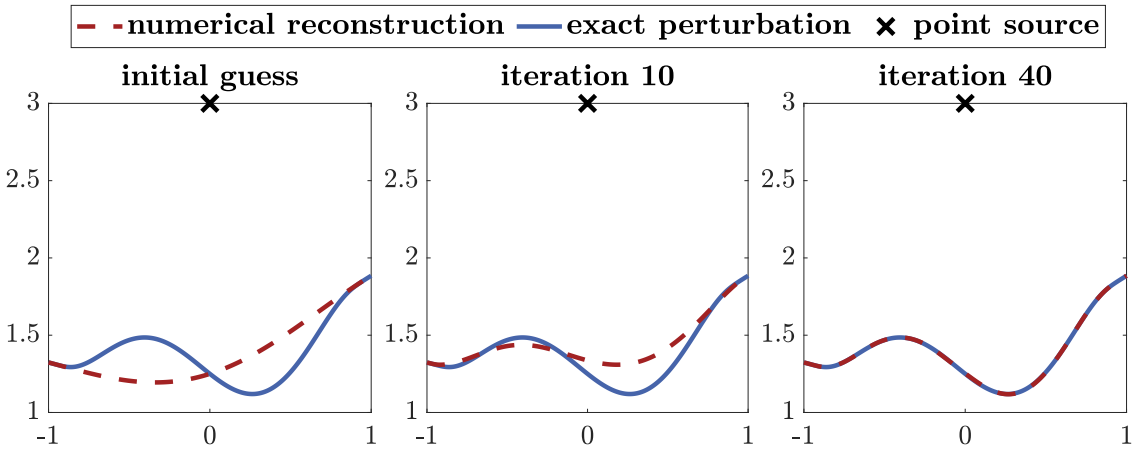


FIGURE 5.4. Numerical reconstruction of δ_2 from noise-free data.

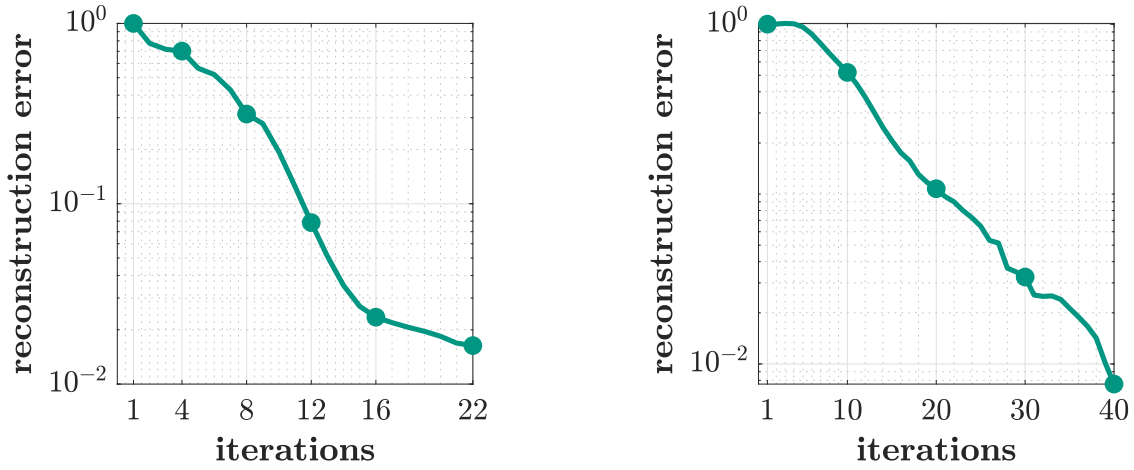


Figure 5.5. Reconstruction error of δ_1 (left) and δ_2 (right) from noise-free data.

The reconstruction shown in Figure 5.4 demonstrates a clear convergence toward the exact perturbation δ_2 by iteration 10. At iteration 40, the reconstruction achieves high accuracy, with a maximum reconstruction error E_{rec} of roughly 10^{-2}, as depicted in the right image of Figure 5.5. Due to the structure of the perturbation δ_2, it requires relatively more iterations to achieve an accurate reconstruction.

To simulate more realistic measurements, we introduce uniformly distributed noise to the near-field data. More specifically, we add 5% noise on the measured data for δ_1 and 2% noise for δ_2. In Figures 5.6 and 5.7, we illustrate the corresponding numerical reconstructions.

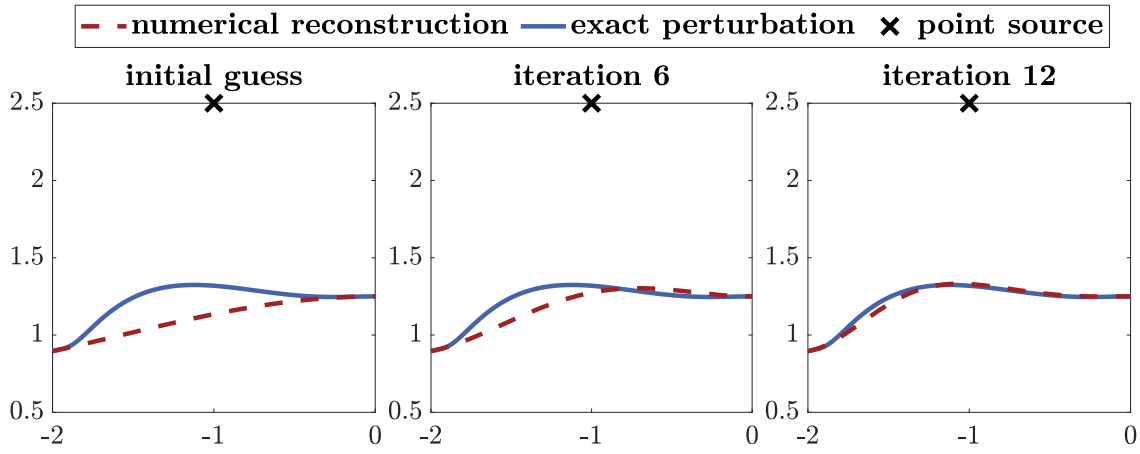


Figure 5.6. Numerical reconstruction of δ_1 from noisy data with 5% noise.

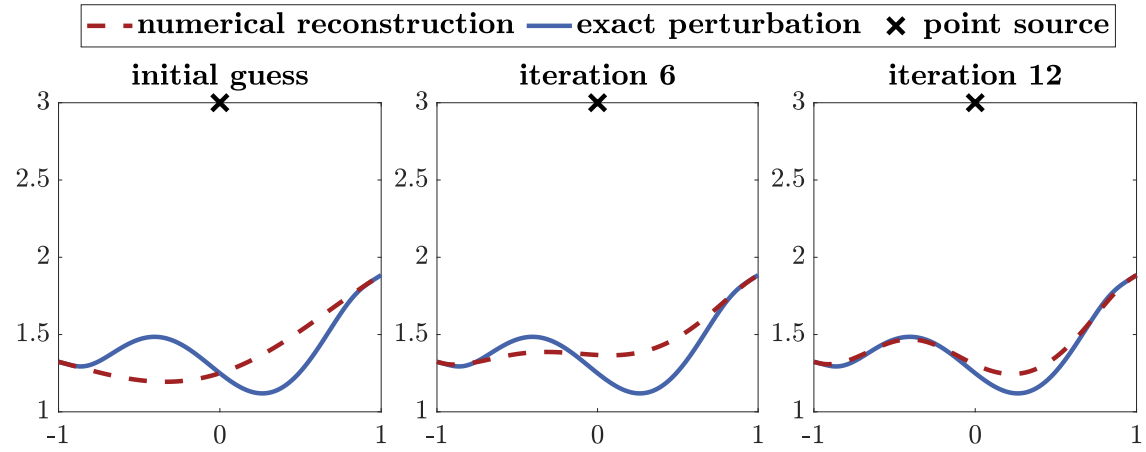


FIGURE 5.7. Numerical reconstruction of δ_2 from noisy data with 2% noise.

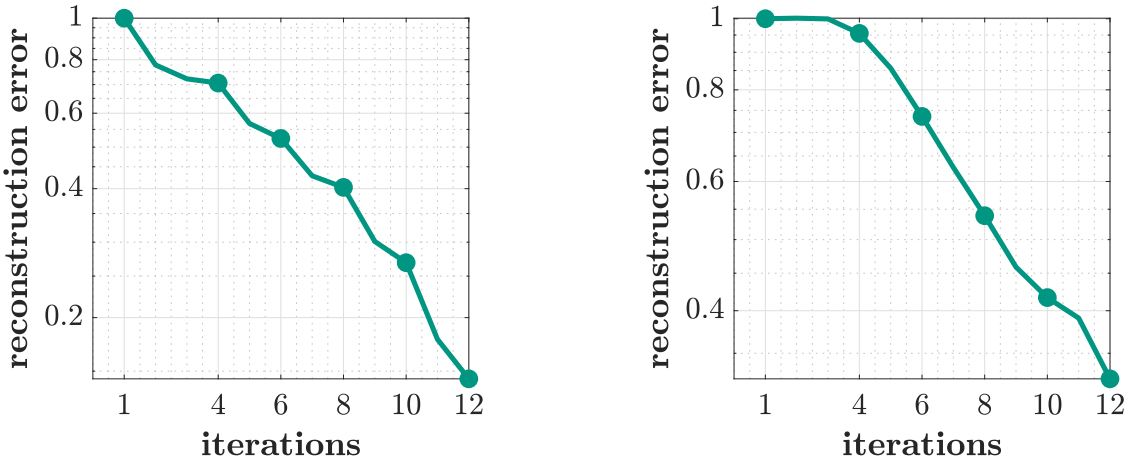


FIGURE 5.8. Reconstruction error of δ_1 (left) and δ_2 (right) from noisy data.

We see in Figure 5.6 and Figure 5.7 that by iteration 6, the numerical reconstruction of the perturbation δ_1 and δ_2, respectively, are already closely approximating the exact perturbations. Around iteration 12 the reconstruction remains satisfactory in both cases. The reconstruction error for both cases can be seen in Figure 5.8.

In conclusion, these results show that the proposed method allows us to accurately reconstruct unknown perturbations on the periodic scatterer.

APPENDIX A

Some Technical Computations

Here we provide some estimates for derivatives of the square root function used in Chapter 3. We also compute the coefficients c^δ and A^δ in the sesquilinear form (4.6) and use these results in Chapters 4 and 5. For simplification, we restrict ourselves to the two-dimensional case.

Estimates for Derivatives of the Square Root Function

Lemma A.1. *Let $s \in \mathbb{C}$, $\alpha \in \mathbb{R}$ such that $\alpha \neq s$. Then, for any $\ell \in \mathbb{N}$,*

$$\left| \frac{\mathrm{d}^\ell}{\mathrm{d}\alpha^\ell} \sqrt{s \pm \alpha} \right| \leq \ell! \, |s \pm \alpha|^{1/2-\ell} .$$

Proof. For any $\ell \geq 0$, a direct calculation yields

$$\left| \frac{\mathrm{d}^\ell}{\mathrm{d}\alpha^\ell} \sqrt{s \pm \alpha} \right| = \frac{|(2\ell-3)!!|}{2^\ell} \, |s \pm \alpha|^{1/2-\ell} \leq \frac{(2\ell)!!}{2^\ell} |s \pm \alpha|^{1/2-\ell} = \ell! \, |s \pm \alpha|^{1/2-\ell} ,$$

where the double factorial $\ell!! \coloneqq \prod_{j=0}^{\lceil \ell/2 \rceil - 1} (\ell - 2j)$; here empty products are equal to 1. □

Lemma A.2. *Let $\nu \in \{1, 2\}$. For any fixed $\ell \in \mathbb{N}$ and $k \in \mathbb{R}_{>0}$, there is a constant C such that*

$$\left| \frac{\partial^\ell \sqrt{k^2 - |\alpha|^2}}{\partial \alpha_\nu^\ell} \right| \leq \frac{C \, \ell! \, |k + |\alpha||^{1/2}}{|k - |\alpha||^{\ell - 1/2}}$$

for all $\alpha \in \mathbb{R}^2$ such that $|\alpha| \neq k$.

Proof. Without loss of generality, we treat the case $\nu = 1$. Consider $\sqrt{k^2 - |\alpha|^2} = \sqrt{s^2 - \alpha_1^2}$ where $s \coloneqq \sqrt{k^2 - \alpha_2^2}$. From the Leibniz formula, we have

$$\left| \frac{\partial^\ell \sqrt{s^2 - \alpha_1^2}}{\partial \alpha_1^\ell} \right| \leq \sum_{n=0}^{\ell} \binom{\ell}{n} \left| \frac{\partial^n \sqrt{s + \alpha_1}}{\partial \alpha_1^n} \right| \left| \frac{\partial^{\ell-n} \sqrt{s - \alpha_1}}{\partial \alpha_1^{\ell-n}} \right| .$$

Using Lemma A.1 leads to

$$\left|\frac{\partial^\ell \sqrt{s^2-\alpha_1^2}}{\partial \alpha_1^\ell}\right| \le \sum_{n=0}^{\ell} \binom{\ell}{n} n!\,(\ell-n)!\,|s+\alpha_1|^{1/2-n}|s-\alpha_1|^{1/2-\ell+n}$$
$$\le \frac{C\,\ell!\sqrt{|s^2-\alpha_1^2|}}{\left(\min\{|s+\alpha_1|,|s-\alpha_1|\}\right)^\ell}\,. \tag{A.1}$$

Now, it remains to estimate $\min\{|s+\alpha_1|,|s-\alpha_1|\}$, and we can distinguish two cases as follows:

(a) If $|\alpha_2| \ge k$, then $s = \mathrm{i}\sqrt{\alpha_2^2-k^2}$. Hence,

$$|s+\alpha_1| = |s-\alpha_1| = \sqrt{\alpha_2^2-k^2+\alpha_1^2} = \sqrt{\left|k^2-|\alpha|^2\right|} \ge |k-|\alpha||\,.$$

(b) If $|\alpha_2| < k$, then $s = \sqrt{k^2-\alpha_2^2} > 0$. In this case, we write

$$\min\{|s+\alpha_1|,|s-\alpha_1|\} = |s-|\alpha_1|| = \frac{\left|k^2-|\alpha|^2\right|}{\sqrt{k^2-\alpha_2^2}+|\alpha_1|}\,.$$

We conclude that

$$\min\{|s+\alpha_1|,|s-\alpha_1|\} \ge \frac{\left|k^2-|\alpha|^2\right|}{k+|\alpha|} = |k-|\alpha||\,.$$

In both cases, we find by substituting $s^2 = k^2-\alpha_2^2$ into (A.1) that

$$\left|\frac{\partial^\ell \sqrt{k^2-|\alpha|^2}}{\partial \alpha_1^\ell}\right| \le \frac{C\,\ell!\,|k+|\alpha||^{1/2}}{|k-|\alpha||^{\ell-1/2}}\,.$$

□

Error Estimate for the Gauss Quadrature rule in Two Dimensions

In Section 3.3, we require an error estimate for the Gauss–Legendre quadrature rule in two dimensions. As a starting point, we recall the standard one-dimensional estimate for the Gauss–Legendre quadrature rule of order n, as stated in [76, Thm. 9.20].

Let $f \in C^{2n+2}([a,b])$, then the error of the Gaussian quadrature formula of order n is given by

$$\left|\int_a^b f(x)\,\mathrm{d}x - \sum_{k=0}^{n} a_k f(x_k)\right| = \frac{f^{(2n+2)}(\psi)}{(2n+2)!}\int_a^b [q_{n+1}(x)]^2\,\mathrm{d}x \qquad \text{for some } \psi \in [a,b]\,,$$

with the orthogonal polynomial q_{n+1} of degree $n+1$.

We assume that $[a,b] = [-1,1]$ and q_{n+1} are the Legendre polynomials. From [93, Sec. 18.2.5], we have

$$\int_{-1}^{1} [q_{n+1}(x)]^2\,\mathrm{d}x = \frac{2}{2n+3}\,.$$

Then, the error of the Gaussian quadrature formula in this case is estimated as

$$\left| \int_{-1}^{1} f(x)\,\mathrm{d}x \; - \sum_{k=0}^{n} w_k f(x_k) \right| \leq \frac{2}{(2n+3)!} \left\| f^{(2n+2)} \right\|_{\infty} .$$

Let us transfer the reference interval $[-1,1]$ to an interval $[-h/2, h/2]$, then we obtain

$$\left| \int_{-h/2}^{h/2} f(t)\,\mathrm{d}t \; - \sum_{k=0}^{n} \frac{h}{2} w_k f\left(\frac{h}{2} x_k\right) \right| \leq \frac{2}{(2n+3)!} \left(\frac{h}{2}\right)^{2n+3} \left\| f^{(2n+2)} \right\|_{\infty} .$$

Now, for the two-dimensional case, we consider $Q \coloneqq [-h/2, h/2]^2$ and $f \in C^{2n+2}(Q)$. Then,

$$\begin{aligned}
&\left| \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} f(t)\,\mathrm{d}t \; - \sum_{j,k=0}^{n} \frac{h^2}{4} w_j w_k f\left(\frac{h}{2} x_j, \frac{h}{2} x_k\right) \right| \\
&\quad \leq \left| \int_{-h/2}^{h/2} \left(\int_{-h/2}^{h/2} f(t_1, t_2)\,\mathrm{d}t_2 \; - \sum_{k=0}^{n} \frac{h}{2} w_k f\left(t_1, \frac{h}{2} x_k\right) \right) \mathrm{d}t_1 \right| \\
&\qquad + \sum_{k=0}^{n} \frac{h}{2} w_k \left| \int_{-h/2}^{h/2} f\left(t_1, \frac{h}{2} x_k\right) \mathrm{d}t_1 \; - \sum_{j=0}^{n} \frac{h}{2} w_j f\left(\frac{h}{2} x_j, \frac{h}{2} x_k\right) \right| \\
&\quad \leq \left| \int_{-h/2}^{h/2} \frac{2}{(2n+3)!} \left(\frac{h}{2}\right)^{2n+3} \left\| \partial_{t_2}^{2n+2} f(t_1, \cdot) \right\|_{\infty} \mathrm{d}t_1 \right| \\
&\qquad + \sum_{k=0}^{n} \frac{h}{2} w_k \frac{2}{(2n+3)!} \left(\frac{h}{2}\right)^{2n+3} \left\| \partial_{t_1}^{2n+2} f\left(\cdot, \frac{h}{2} x_k\right) \right\|_{\infty} \\
&\quad \leq \frac{4}{(2n+3)!} \left(\frac{h}{2}\right)^{2n+4} \max\left\{ \left\| \partial_{t_1}^{2n+2} f \right\|_{\infty}, \left\| \partial_{t_2}^{2n+2} f \right\|_{\infty} \right\} .
\end{aligned} \tag{A.2}$$

Asymptotic Representation of the Coefficients in the Sesquilinear Form (4.6)
In the following lemma, we provide some necessary technical computations for the coefficients of the sesquilinear form (4.6).

Lemma A.3. *Let $\delta \in C^2(\mathbb{R})$ be sufficiently small and the diffeomorphism Ψ^δ and the function β_h^δ be as in (4.4) and (4.5), respectively. Moreover, let $c^\delta \coloneqq \left|\det \nabla \Psi^\delta\right|$ and $A^\delta \coloneqq c^\delta (\nabla \Psi^\delta)^{-1} (\nabla \Psi^\delta)^{-\top}$ as in (4.7). Then, $c^\delta = 1 + \delta \partial_{x_2} \beta_h^\delta$ and*

$$A^\delta = \begin{bmatrix} 1 + \delta \partial_{x_2} \beta_h^\delta & -\delta \partial_{x_1} \beta_h^\delta - \delta' \beta_h^\delta \\ -\delta \partial_{x_1} \beta_h^\delta - \delta' \beta_h^\delta & 1 - \delta \partial_{x_2} \beta_h^\delta \end{bmatrix} + \mathcal{O}(\|\delta\|_{1,\infty}^2) \quad \text{as } \|\delta\|_{1,\infty} \to 0 . \tag{A.3}$$

Proof. Using the definition of Ψ^δ and a straightforward computation, we have

$$c^\delta = \left|\det \nabla \Psi^\delta\right| = \left| \det \begin{bmatrix} 1 & 0 \\ \delta \partial_{x_1} \beta_h^\delta + \delta' \beta_h^\delta & 1 + \delta \partial_{x_2} \beta_h^\delta \end{bmatrix} \right| = 1 + \delta \partial_{x_2} \beta_h^\delta , \tag{A.4}$$

where the absolute value in the last equality is removed since for sufficiently small δ, the

determinant is positive. Moreover, we can write

$$\begin{aligned} A^\delta &= c^\delta (\nabla \Psi^\delta)^{-1} (\nabla \Psi^\delta)^{-\top} \\ &= \frac{1}{1+\delta\partial_{x_2}\beta_h^\delta} \begin{bmatrix} (1+\delta\partial_{x_2}\beta_h^\delta)^2 & -(1+\delta\partial_2\beta_h^\delta)(\delta\partial_{x_1}\beta_h^\delta+\delta'\beta_h^\delta) \\ -(1+\delta\partial_2\beta_h^\delta)(\delta\partial_{x_1}\beta_h^\delta+\delta'\beta_h^\delta) & 1+(\delta'\beta_h^\delta+\delta\partial_1\beta_h^\delta)^2 \end{bmatrix} \\ &= \begin{bmatrix} 1+\delta\partial_{x_2}\beta_h^\delta & -\delta\partial_{x_1}\beta_h^\delta-\delta'\beta_h^\delta \\ -\delta\partial_{x_1}\beta_h^\delta-\delta'\beta_h^\delta & \dfrac{1}{1+\delta\partial_{x_2}\beta_h^\delta} \end{bmatrix} + \mathcal{O}(\|\delta\|_{1,\infty}^2) \quad \text{as } \|\delta\|_{1,\infty}\to 0\,. \end{aligned}$$

Using the fact that $\dfrac{1}{1+\delta\partial_{x_2}\beta_h^\delta} = 1-\delta\partial_{x_2}\beta_h^\delta + \mathcal{O}(\|\delta\|_{1,\infty}^2)$ completes the proof. □

Remark A.4. From Lemma A.3, it follows that $c^\delta = 1+\mathcal{O}(\|\delta\|_{1,\infty})$ and $A^\delta = I+\mathcal{O}(\|\delta\|_{1,\infty})$ as $\|\delta\|_{1,\infty}\to 0$. Therefore, for $u\in H^2(\mathbb{R}^2)$, we have $\nabla\cdot(A^\delta\nabla u) = \Delta u + \mathcal{O}(\|\delta\|_{1,\infty})$ and $c^\delta u = u+\mathcal{O}(\|\delta\|_{1,\infty})$. This leads to

$$\nabla\cdot(A^\delta\nabla u)+k^2c^\delta u = \Delta u + k^2 u + \mathcal{O}(\|\delta\|_{1,\infty}) \quad \text{as } \|\delta\|_{1,\infty}\to 0\,.$$

In the following lemma, we provide some necessary computations for Theorem 5.5. This can be obtained by performing a lengthy application of the product rule.

Lemma A.5. *Let δ, A^δ as in Lemma A.3, $u,\phi\in H^2(\mathbb{R}^2)$ and $v^\delta := \delta\beta_h^\delta(x)\partial_{x_2}u$. Then, we have*

$$\begin{aligned} ((A^\delta-I)\nabla u)\cdot\overline{\nabla\phi} = &-\operatorname{div}\Big\{v^\delta\overline{\nabla\phi}+\delta\beta_h^\delta\nabla u\overline{\partial_{x_2}\phi}-\delta\beta_h^\delta(\nabla u\cdot\overline{\nabla\phi})e_2\Big\} \\ &+v^\delta\overline{\Delta\phi}+\delta\partial_{x_2}\beta_h^\delta(\nabla u\cdot\overline{\nabla\phi})+\mathcal{O}(\|\delta\|_{1,\infty}^2) \quad \text{as } \|\delta\|_{1,\infty}\to 0\,, \end{aligned} \tag{A.5}$$

where β_h^δ is defined as in (4.5) and $e_2 := (0,1)^\top$.

Proof. First using the asymptotic representation of A^δ given in (A.3) and then adding and subtracting $\delta\partial_{x_2}\beta_h^\delta\partial_{x_2}u\overline{\partial_{x_2}\phi}$ and $\delta\beta_h^\delta\nabla(\partial_{x_2}u)\cdot\overline{\nabla\phi}$, we obtain

$$\begin{aligned} ((A^\delta-I)\nabla u)\cdot\overline{\nabla\phi} = &\,\delta\partial_{x_2}\beta_h^\delta(\nabla u\cdot\overline{\nabla\phi})-\delta\left(\nabla\beta_h^\delta\cdot\nabla u\right)\overline{\partial_{x_2}\phi}-\delta'\beta_h^\delta(\partial_{x_1}u)\overline{\partial_{x_2}\phi} \\ &-\delta\left(\nabla\beta_h^\delta\cdot(\partial_{x_2}u\overline{\nabla\phi})\right)-\delta'\beta_h^\delta(\partial_{x_2}u)\overline{\partial_{x_1}\phi} \\ &-\delta\beta_h^\delta\nabla(\partial_{x_2}u)\cdot\overline{\nabla\phi}+\delta\beta_h^\delta\nabla(\partial_{x_2}u)\cdot\overline{\nabla\phi}+\mathcal{O}(\|\delta\|_{1,\infty}^2) \quad \text{as } \|\delta\|_{1,\infty}\to 0\,. \end{aligned}$$

Using the definition of v^δ and the fact that

$$\nabla v^\delta\cdot\overline{\nabla\phi} = \delta\left(\nabla\beta_h^\delta\cdot(\partial_{x_2}u\overline{\nabla\phi})\right)+\delta'\beta_h^\delta\partial_{x_2}u\overline{\partial_{x_1}\phi}\,,+\delta\beta_h^\delta\nabla(\partial_{x_2}u)\cdot\overline{\nabla\phi}\,,$$

we get

$$\begin{aligned} ((A^\delta-I)\nabla u)\cdot\overline{\nabla\phi} = &\,\delta\partial_{x_2}\beta_h^\delta\nabla u\cdot\overline{\nabla\phi}-\delta\left(\nabla\beta_h^\delta\cdot\nabla u\right)\overline{\partial_{x_2}\phi}-\delta'\beta_h^\delta\partial_{x_1}u\overline{\partial_{x_2}\phi} \\ &-\nabla v^\delta\cdot\overline{\nabla\phi}+\delta\beta_h^\delta\nabla(\partial_{x_2}u)\cdot\overline{\nabla\phi}+\mathcal{O}(\|\delta\|_{1,\infty}^2) \quad \text{as } \|\delta\|_{1,\infty}\to 0\,. \end{aligned}$$

Adding and subtracting $\delta\beta_h^\delta \nabla u \cdot \overline{\nabla(\partial_{x_2}\phi)}$ to the above equation and considering

$$\operatorname{div}\Big\{\delta\beta_h^\delta(\nabla u \cdot \overline{\nabla\phi})e_2\Big\} = \delta\partial_{x_2}\beta_h^\delta \nabla u \cdot \overline{\nabla\phi} + \delta\beta_h^\delta \nabla(\partial_{x_2}u)\cdot\overline{\nabla\phi} + \delta\beta_h^\delta\nabla u\cdot\overline{\nabla(\partial_{x_2}\phi)}\,,$$

we arrive at

$$\begin{aligned}((A^\delta - I)\nabla u)\cdot\overline{\nabla\phi} &= \delta\partial_{x_2}\beta_h^\delta\nabla u\cdot\overline{\nabla\phi} - \delta\left(\nabla\beta_h^\delta\cdot\nabla u\right)\overline{\partial_{x_2}\phi} - \delta'\beta_h^\delta(\partial_{x_1}u)\overline{\partial_{x_2}\phi}\\ &\quad - \nabla v^\delta\cdot\overline{\nabla\phi} + \delta\beta_h^\delta\nabla(\partial_{x_2}u)\cdot\overline{\nabla\phi} + \delta\beta_h^\delta\nabla u\cdot\overline{\nabla(\partial_{x_2}\phi)}\\ &\quad - \delta\beta_h^\delta\nabla u\cdot\overline{\nabla(\partial_{x_2}\phi)} + \mathcal{O}(\|\delta\|_{1,\infty}^2)\\ &= -\delta\left(\nabla\beta_h^\delta\cdot\nabla u\right)\overline{\partial_{x_2}\phi} - \delta'\beta_h^\delta(\partial_{x_1}u)\overline{\partial_{x_2}\phi} - \nabla v^\delta\cdot\overline{\nabla\phi}\\ &\quad + \operatorname{div}\Big\{\delta\beta_h^\delta(\nabla u\cdot\overline{\nabla\phi})e_2\Big\} - \delta\beta_h^\delta\nabla u\cdot\overline{\nabla(\partial_{x_2}\phi)} + \mathcal{O}(\|\delta\|_{1,\infty}^2) \text{ as } \|\delta\|_{1,\infty}\to 0\,.\end{aligned}$$

Adding and subtracting $\delta\partial_{x_2}\beta_h^\delta(\nabla u\cdot\overline{\nabla\phi})$ yields

$$\begin{aligned}((A^\delta - I)\nabla u)\cdot\overline{\nabla\phi} &= -\delta\left(\nabla\beta_h^\delta\cdot\nabla u\right)\overline{\partial_{x_2}\phi} - \delta'\beta_h^\delta\partial_{x_1}u\overline{\partial_{x_2}\phi} - \nabla v^\delta\cdot\overline{\nabla\phi}\\ &\quad + \operatorname{div}\Big\{\delta\beta_h^\delta(\nabla u\cdot\overline{\nabla\phi})e_2\Big\} - \delta\beta_h^\delta\nabla u\cdot\overline{\nabla(\partial_{x_2}\phi)}\\ &\quad + \delta\partial_{x_2}\beta_h^\delta\nabla u\cdot\overline{\nabla\phi} - \delta\partial_{x_2}\beta_h^\delta\nabla u\cdot\overline{\nabla\phi} + \mathcal{O}(\|\delta\|_{1,\infty}^2) \quad \text{as } \|\delta\|_{1,\infty}\to 0\,.\end{aligned}$$

Since

$$\begin{aligned}\operatorname{div}\Big\{h\beta_h^\delta\nabla u\overline{\partial_{x_2}\phi}\Big\} &= \delta\beta_h^\delta\nabla u\cdot\overline{\nabla(\partial_{x_2}\phi)} + \delta\left(\nabla\beta_h^\delta\cdot\nabla u\right)\overline{\partial_{x_2}\phi}\\ &\quad + \delta'\beta_h^\delta\partial_{x_1}u\overline{\partial_{x_2}\phi} + \delta\partial_{x_2}\beta_h^\delta\nabla u\cdot\overline{\nabla\phi}\,,\end{aligned}$$

we obtain

$$\begin{aligned}((A^\delta - I)\nabla u)\cdot\overline{\nabla\phi} &= -\nabla v^\delta\cdot\overline{\nabla\phi} - \operatorname{div}\Big\{h\beta_h^\delta\nabla u\overline{\partial_{x_2}\phi}\Big\} + \operatorname{div}\Big\{h\beta_h^\delta(\nabla u\cdot\overline{\nabla\phi})e_2\Big\}\\ &\quad + \delta\partial_{x_2}\beta_h^\delta\nabla u\cdot\overline{\nabla\phi} + \mathcal{O}(\|\delta\|_{1,\infty}^2) \quad \text{as } \|\delta\|_{1,\infty}\to 0\,.\end{aligned}$$

Finally, by adding and subtracting $v^\delta\overline{\Delta\phi}$ and taking into account that

$$\operatorname{div}\Big\{v^\delta\overline{\nabla\phi}\Big\} = \nabla v^\delta\cdot\overline{\nabla\phi} + v^\delta\overline{\Delta\phi}\,,$$

we complete the proof. □

APPENDIX B

Computational Complexity of Direct and Iterative Solvers

In Algorithm 3, we proposed a fast iterative solver for solving the linear system (4.46). In this Appendix, we will compare the computational cost of the iterative solver with the direct solver introduced in [98, Sec. 2] using the Sherman–Morrison–Woodbury formula. To this end, we first estimate the complexity of each method separately and then compare them together.

We consider the following notations to represent the cost of operations on square matrices of size $N_\Delta \times N_\Delta$ or vectors of length N_Δ:

- $E_{m,\times}$: inverting a matrix or multiplying two matrices;
- $E_{m,+}$: summing two matrices or multiplying a matrix by a diagonal matrix;
- $E_{v,\times}$: multiplying a vector by a matrix;
- $E_{v,+}$: summing two vectors or multiplying a vector by a diagonal matrix.

Computational Cost of the Iterative Solver Proposed in Algorithm 3
As shown in Chapter 4, we reformulate the linear system given in (4.46) using the Schur complement recursively. This yields

$$\left(\mathbf{I} - \sum_{j=1}^{N_\alpha} \mathbf{C}_j \mathbf{A}_j^{-1} \mathbf{B}_j\right) U = -\sum_{j=1}^{N_\alpha} \mathbf{C}_j \mathbf{A}_j^{-1} F_j \,,$$

where $\mathbf{A}_j, \mathbf{B}_j$ are $N_\Delta \times N_\Delta$ sparse matrices, $\mathbf{C}_j$ is a diagonal matrix and F_j is a vector of length N_Δ. In what follows, we estimate the cost of the iterative solver proposed in Algorithm 3 to obtain U.

As a preliminary step, we compute the inverse of the matrices $\mathbf{A}_j$ for $j \in \{1, \ldots, N_\alpha\}$, having a cost of $N_\alpha E_{m,\times}$. We use these inverse matrices on both the left and right-hand sides of the above equation. Now, we estimate separately the complexity of computing the right-hand side and one iteration of the left-hand side.

- On the right-hand side, computing $\mathbf{A}_j^{-1}F_j$ requires a total cost of $N_\alpha E_{v,\times}$. Subsequently, computing the products $\mathbf{C}_j\mathbf{A}_j^{-1}F_j$ has a cost of $N_\alpha E_{v,+}$. Finally, computing the sum needs an additional cost of $N_\alpha E_{v,+}$. These steps add up to $N_\alpha(E_{v,\times}+2E_{v,+})$.
- On the left-hand side, performing the matrix vector multiplication $\mathbf{B}_j U$ has a complexity of $N_\alpha E_{v,\times}$. Afterwards, the multiplication $\mathbf{A}_j^{-1}\mathbf{B}_j U$ adds a cost of $N_\alpha E_{v,\times}$. Multiplying by $\mathbf{C}_j$ and summing the resulting terms requires $2N_\alpha E_{v,+}$. This leads to the total cost of $2N_\alpha(E_{v,\times}+E_{v,+})$ for each iteration.

Now, the total complexity of the proposed iterative method, denoted by E_{iter} is estimated as

$$E_{\text{iter}} = N_\alpha(E_{m,\times}+E_{v,\times}+2E_{v,+}) + 2N_{\text{iter}}N_\alpha(E_{v,\times}+E_{v,+})\,.$$

where N_{iter} denotes the number of required iterations in the iterative solver.

Computational Cost of the Direct Solver Proposed in [98, Thm. 2.2]
The main idea in [98, Thm. 2.2] is to use the Sherman–Morrison–Woodbury formula to compute the inverse of the block arrowhead matrix in (4.46). Following the same approach as [98], the inverse of the coefficient matrix (4.46), denoted by $\widehat{\mathbf{A}}^{-1}$, is obtained by

$$\widehat{\mathbf{A}}^{-1} \coloneqq \mathbf{D}^{-1} + \begin{bmatrix} \mathbf{A}_1^{-1}\mathbf{B}_1 \\ \vdots \\ \mathbf{A}_{N_\alpha}^{-1}\mathbf{B}_{N_\alpha} \\ -\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Z}\mathbf{C}_1\mathbf{A}_1^{-1} & \dots & \mathbf{Z}\mathbf{C}_{N_\alpha}\mathbf{A}_{N_\alpha}^{-1} & -\mathbf{Z} \end{bmatrix},$$

where $\mathbf{D} \coloneqq \operatorname{diag}(\mathbf{A}_1,\dots,\mathbf{A}_{N_\alpha},\mathbf{0})$ and $\mathbf{Z} \coloneqq \left(\mathbf{I}-\sum_{j=1}^{N_\alpha}\mathbf{C}_j\mathbf{A}_j^{-1}\mathbf{B}_j\right)^{-1}$.

According to [98, Sec. 3], the computational complexity of $\widehat{\mathbf{A}}^{-1}$ is estimated as follows:

(a) similar to the previous case, inverting the matrices $\mathbf{A}_j$ requires $N_\alpha E_{m,\times}$ operations;

(b) computing $\mathbf{A}_j^{-1}\mathbf{B}_j$ adds $N_\alpha E_{m,\times}$ operations;

(c) the term $\mathbf{I}-\sum_{j=1}^{N_\alpha}\mathbf{C}_j\mathbf{A}_j^{-1}\mathbf{B}_j$ can be obtained by multiplying by a diagonal matrix and summing each block, which brings the cost of $2N_\alpha E_{m,+}$.

(d) computing the inverse of the previous term, which gives $\mathbf{Z}$, has a cost of $E_{m,\times}$;

(e) performing $\mathbf{C}_j\mathbf{A}_j^{-1}$ requires $N_\alpha E_{m,+}$ operations;

(f) the products $\mathbf{Z}\mathbf{C}_j\mathbf{A}_j^{-1}$ have the cost of $N_\alpha E_{m,\times}$;

(g) the products $\mathbf{A}_j^{-1}\mathbf{B}_j\mathbf{Z}\mathbf{C}_s\mathbf{A}_s^{-1}$ for $j,s\in\{1,\dots,N_\alpha\}$ use Items (b) and (f) and require $N_\alpha^2 E_{m,\times}$ operations;

In conclusion, the total complexity of computing the inverse of (4.46) is estimated as

$$E_{\text{direct}} = E_{m,\times} + N_\alpha(3E_{m,\times}+3E_{m,+}+N_\alpha E_{m,\times})\,.$$

Comparison of the Computational Costs of the Direct and Iterative Solvers

To compare the cost of these two approaches, we assume $N_{\text{iter}} \approx cN_\alpha$ for a small constant c and $N_\Delta \gg N_\alpha$. This assumption is reasonable based on the results reported in Table 4.1. Due to the fact that $E_{v,\times} > E_{v,+}$, we obtain

$$\begin{aligned}\frac{E_{\text{direct}}}{E_{\text{iter}}} &= \frac{(1+3N_\alpha+N_\alpha^2)E_{m,\times}+3N_\alpha E_{m,+}}{N_\alpha E_{m,\times}+N_\alpha(1+2N_{\text{iter}})E_{v,\times}+2N_\alpha(1+N_{\text{iter}})E_{v,+}} \\ &\geq \frac{(1+3N_\alpha+N_\alpha^2)E_{m,\times}}{N_\alpha E_{m,\times}+N_\alpha(3+4cN_\alpha)E_{v,\times}}\,.\end{aligned}$$

For a $N_\Delta \times N_\Delta$ matrix $\mathbf{A}_j$, it holds $E_{m,\times} = \mathcal{O}(N_\Delta^3)$ and $E_{m,+} = \mathcal{O}(N_\Delta^2)$. Taking into account that $N_\alpha(3+4cN_\alpha) < N_\Delta$, we obtain

$$\frac{E_{\text{direct}}}{E_{\text{iter}}} > \frac{(1+3N_\alpha+N_\alpha^2)E_{m,\times}}{N_\alpha E_{m,\times}+N_\Delta E_{v,\times}} \geq \frac{(N_\alpha+1)^2N_\Delta^3}{(N_\alpha+1)N_\Delta^3} = N_\alpha+1\,.$$

As we have shown above, the proposed iterative solver is faster than computing the inverse of the block arrowhead matrix in (4.46).

APPENDIX C

GREEN'S FUNCTION AND ITS PROPERTIES

In Chapters 3 and 4, we select the Dirichlet Green's function in the upper half space to show the efficiency of the proposed methods. In this appendix, we aim to summerize some properties of this function and clarify why we consider Green's function as an incident field instead of the fundamental solution. Moreover, we show how to compute the FB transform of the Green's function.

The Dirichlet Green's function in the upper half space is defined by

$$G(x,y) \coloneqq \Phi(x,y) - \Phi(x,y')\,, \quad x \neq y \in \mathbb{R}^d_+ \coloneqq \left\{x \in \mathbb{R}^d : x_d > 0\right\},$$

with the reflected point source $y' \coloneqq (y_1, \ldots, y_{d-1}, -y_d)^\top$ and the fundamental solution of the Helmholtz equation

$$\Phi(x,y) = \begin{cases} \dfrac{\mathrm{i}}{4} H_0^{(1)}(k|x-y|) & x,y \in \mathbb{R}^2, x \neq y\,, \\ \dfrac{1}{4\pi} \dfrac{\mathrm{e}^{\mathrm{i}k|x-y|}}{|x-y|} & x,y \in \mathbb{R}^3, x \neq y\,, \end{cases} \tag{C.1}$$

where $H_0^{(1)}$ is the Hankel function of the first kind of order zero.

Remark C.1. Note that for $x \neq y$ the three-dimensional fundamental solution can be written based on the Hankel function of the first kind of order $-1/2$ as follows

$$\frac{\mathrm{e}^{\mathrm{i}k|x-y|}}{|x-y|} = \sqrt{\frac{k\pi}{2|x-y|}} H^{(1)}_{-1/2}(k|x-y|)\,.$$

To test the efficiency of the proposed numerical methods, instead of the fundamental solution, we use Green's function because of its faster decay.

Lemma C.2. *Let $x = (\widetilde{x}, x_d) \in \mathbb{R}^d$ with $\widetilde{x} = (x_1, \ldots, x_{d-1})$.*

(a) The fundamental solution Φ decays as $|x_1|^{-1/2}$ in $\mathbb{R}^2$ and as $|\widetilde{x}|^{-1}$ in $\mathbb{R}^3$.

(b) The decay rate of Green's function G is $|x_1|^{-3/2}$ in $\mathbb{R}^2$ and $|\widetilde{x}|^{-2}$ in $\mathbb{R}^3$.

Proof. For the two-dimensional case, using the asymptotic behavior of the Hankel function [93, Sec. 10.17], we obtain

$$H_0^{(1)}(k|x-y|) \sim \sqrt{\frac{2}{\pi k|x-y|}}\mathrm{e}^{\mathrm{i}(k|x-y|-\pi/4)}\sum_{\ell=0}^{\infty}\frac{\mathrm{i}^\ell a_\ell}{(k|x-y|)^\ell}\,,$$

where a_ℓ are coefficients only depending on ℓ.

(a) For real k, we have $|\exp(\mathrm{i}k|x-y|-\mathrm{i}\pi/4)|=1$. Then,

$$H_0^{(1)}(k|x-y|) \sim c|x_1|^{-1/2}+\mathcal{O}(|x_1|^{-3/2}) \quad \text{as } |x_1|\to\infty\,.$$

For the three-dimensional case, the statement is clear from the definition.

(b) From the asymptotic behavior of the Hankel function, we have

$$\begin{aligned}|G(x,y)| &= \left|H_0^{(1)}(k|x-y|)-H_0^{(1)}(k|x-y'|)\right| \\ &\sim \sqrt{\frac{2}{k\pi}}\left|\frac{\mathrm{e}^{\mathrm{i}(k|x-y|-\pi/4)}}{\sqrt{|x-y|}}\sum_{\ell=0}^{\infty}\frac{\mathrm{i}^\ell a_\ell}{(k|x-y|)^\ell}-\frac{\mathrm{e}^{\mathrm{i}(k|x-y'|-\pi/4)}}{\sqrt{|x-y'|}}\sum_{\ell=0}^{\infty}\frac{\mathrm{i}^\ell a_\ell}{(k|x-y'|)^\ell}\right| \\ &= c(k)\left|\frac{\mathrm{e}^{\mathrm{i}k|x-y|}}{\sqrt{|x-y|}}-\frac{\mathrm{e}^{\mathrm{i}k|x-y'|}}{\sqrt{|x-y'|}}\right|+\mathcal{O}(|x_1|^{-3/2}) \quad \text{as } |x_1|\to\infty\,.\end{aligned}$$

Let $z \coloneqq x_1-y_1$, $c_- \coloneqq x_2-y_2$ and $c_+ \coloneqq x_2+y_2$. Then, we obtain

$$\begin{aligned}\left|\frac{\mathrm{e}^{\mathrm{i}k|x-y|}}{\sqrt{|x-y|}}-\frac{\mathrm{e}^{\mathrm{i}k|x-y'|}}{\sqrt{|x-y'|}}\right| &= \left|\frac{\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_-^2}}}{\sqrt[4]{z^2+c_-^2}}-\frac{\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_+^2}}}{\sqrt[4]{z^2+c_+^2}}\right| \\ &= \frac{1}{\sqrt{|z|}}\left|\frac{\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_-^2}}}{\sqrt[4]{1+(\frac{c_-}{z})^2}}-\frac{\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_+^2}}}{\sqrt[4]{1+(\frac{c_+}{z})^2}}\right| \\ &= \frac{1}{\sqrt{|z|}}\left|\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_-^2}}-\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_+^2}}\right|+\mathcal{O}(|z|^{-5/2}) \quad \text{as } |z|\to\infty\,.\end{aligned}$$

Using the identity

$$|\exp(\mathrm{i}s)-\exp(\mathrm{i}t)| = |\exp(\mathrm{i}st)||\exp(\mathrm{i}/t)-\exp(\mathrm{i}/s)| = \left|2\sin\left(\frac{1}{2t}-\frac{1}{2s}\right)\right|,$$

for $s,t\in\mathbb{R}$, we obtain as before

$$\begin{aligned}\left|\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_-^2}}-\mathrm{e}^{\mathrm{i}k\sqrt{z^2+c_+^2}}\right| &= 2\sin\left(\left|\frac{1}{2k\sqrt{z^2+c_-^2}}-\frac{1}{2k\sqrt{z^2+c_+^2}}\right|\right) \\ &= 2\sin\left(\frac{1}{2k|z|}+\mathcal{O}(|z|^{-2})\right)=\mathcal{O}(|z|^{-1}) \quad \text{as } |z|\to\infty\,.\end{aligned}$$

Combining the estimates above and using the definition of z leads to $|G(x,y)|=\mathcal{O}(|x_1|^{-3/2})$ as $|x_1|\to\infty$. The proof for three dimensions is given in [24, Eq. (2.11)]. □

To compute the right-hand sides of (3.33) and (4.46), it is required to compute the FB transform of Green's function, which has been obtained in the following lemma.

Lemma C.3. *The FB transform of Green's function is obtained by*

$$(\mathcal{J}G)(\alpha;x) = C_d\,\mathrm{e}^{-\mathrm{i}\alpha\cdot\widetilde{x}}\sum_{j\in\mathbb{Z}^d}\mathrm{e}^{\mathrm{i}(\alpha+j)\cdot(\widetilde{x}-\widetilde{y})}\begin{cases}\mathrm{e}^{\mathrm{i}\gamma_j x_d}\,\mathrm{sinc}\,(\gamma_j y_d)y_d & y_d < x_d\,,\\ \mathrm{e}^{\mathrm{i}\gamma_j y_d}\,\mathrm{sinc}\,(\gamma_j x_d)x_d & \textit{otherwise}\,,\end{cases}$$

where $\gamma_j \coloneqq \sqrt{k^2-|\alpha+j|^2}$, $C_d = 1/2\pi$ *for* $d=2$ *and* $C_d = 1$ *for* $d=3$.

Proof. We first consider the two-dimensional case $d=2$. Using the definition of the FB transform, we have

$$\begin{aligned}(\mathcal{J}G)(\alpha;x) &= \frac{\mathrm{i}}{4}\sum_{j\in\mathbb{Z}}G(x_1+2\pi j,x_2)\mathrm{e}^{-\mathrm{i}\alpha(x_1+2\pi j)}\\ &= \frac{\mathrm{i}}{4}\sum_{j\in\mathbb{Z}}\left(H_0^{(1)}(k|x-y+2\pi j e_1|)-H_0^{(1)}(k|x-y+2\pi j e_1|)\right)\mathrm{e}^{-\mathrm{i}\alpha(x_1+2\pi j)}\,,\end{aligned}$$

where $e_1 \coloneqq (1,0)^\top$. Using [5, Eq. (2.7)] and the poisson summation formula, the Fourier series expansion of the fundamental solution is obtained by

$$\begin{aligned}(\mathcal{J}G)(\alpha;x) &= \frac{\mathrm{i}}{4\pi}\mathrm{e}^{-\mathrm{i}\alpha x_1}\sum_{j\in\mathbb{Z}}\frac{1}{\gamma_j}\left[\mathrm{e}^{\mathrm{i}(\alpha+j)(x_1-y_1)+\mathrm{i}\gamma_j|x_2-y_2|}-\mathrm{e}^{\mathrm{i}(\alpha+j)(x_1-y_1)+\mathrm{i}\gamma_j|x_2+y_2|}\right]\\ &= \frac{\mathrm{i}}{4\pi}\mathrm{e}^{-\mathrm{i}\alpha x_1}\sum_{j\in\mathbb{Z}}\frac{\mathrm{e}^{\mathrm{i}(\alpha+j)(x_1-y_1)}}{\gamma_j}\begin{cases}\mathrm{e}^{\mathrm{i}\gamma_j x_2}(\mathrm{e}^{-\mathrm{i}\gamma_j y_2}-\mathrm{e}^{\mathrm{i}\gamma_j y_2}) & \text{if } x_2>y_2\,,\\ \mathrm{e}^{\mathrm{i}\gamma_j y_2}(\mathrm{e}^{-\mathrm{i}\gamma_j x_2}-\mathrm{e}^{\mathrm{i}\gamma_j x_2}) & \text{otherwise}\,.\end{cases}\end{aligned}$$

Now using the fact that $\mathrm{sinc}\,(\gamma_j y_2) = (\mathrm{e}^{\mathrm{i}\gamma_j y_2}-\mathrm{e}^{-\mathrm{i}\gamma_j y_2})/2\mathrm{i}\gamma_j y_2$, we obtain

$$(\mathcal{J}G)(\alpha;x) = \frac{1}{2\pi}\mathrm{e}^{-\mathrm{i}\alpha x_1}\sum_{j\in\mathbb{Z}}\mathrm{e}^{\mathrm{i}(\alpha+j)(x_1-y_1)}\begin{cases}\mathrm{e}^{\mathrm{i}\gamma_j x_2}\,\mathrm{sinc}\,(\gamma_j y_2)y_2 & \text{if } x_2>y_2\,,\\ \mathrm{e}^{\mathrm{i}\gamma_j y_2}\,\mathrm{sinc}\,(\gamma_j x_2)x_2 & \text{otherwise}\,.\end{cases}$$

For the three-dimensional case, we refer to [84, Eqs. (52)-(54)]. □

Bibliography

[1] H. Ammari. Uniqueness theorems for an inverse problem in a doubly periodic structure. *Inverse Problems*, 11(4):823–833, 1995. DOI: 10.1088/0266-5611/11/4/013.

[2] H. Ammari, S. Barandun, and A. Uhlmann. Truncated Floquet-Bloch transform for computing the spectral properties of large finite systems of resonators. Preprint. 2024. DOI: 10.48550/arXiv.2410.17597.

[3] T. Arens. Scattering by Biperiodic Layered Media: The Integral Equation Approach. Habilitationschrift. Karlsruher Institut für Technologie, Karslruhe, 2010. DOI: 10.5445/IR/1000016241.

[4] T. Arens and T. Hohage. On radiation conditions for rough surface scattering problems. *IMA J. Appl. Math.*, 70(6):839–847, 2005. DOI: 10.1093/imamat/hxh065.

[5] T. Arens, K. Sandfort, S. Schmitt, and A. Lechleiter. Analysing Ewald's method for the evaluation of Green's functions for periodic media. *IMA J. Appl. Math.*, 78(3):405–431, 2013. DOI: 10.1093/imamat/hxr057.

[6] T. Arens, N. Shafieeabyaneh, and R. Zhang. A high-order numerical method for solving non-periodic scattering problems in three-dimensional bi-periodic structures. *ZAMM Z. Angew. Math. Mech.*, 104(9):Paper No. e202300650, 21, 2024. DOI: 10.1002/zamm.202300650.

[7] T. Arens and R. Zhang. A nonuniform mesh method in the Floquet parameter domain for wave scattering by periodic surfaces. *Math. Methods Appl. Sci.*, 48(4):4289–4309, 2025. DOI: 10.1002/mma.10548.

[8] K. Atkinson. *Introduction to Numerical Analysis*. Wiley, New York, 2008.

[9] K. Atkinson and W. Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer, New York, 2005. DOI: 10.1007/978-1-4419-0458-4.

[10] G. Bao. Finite element approximation of time harmonic waves in periodic structures. *SIAM J. Numer. Anal.*, 32(4):1155–1169, 1995. DOI: 10.1137/0732053.

[11] G. Bao. A uniqueness theorem for an inverse problem in periodic diffractive optics. *Inverse Problems*, 10(2):335–345, 1994. DOI: 10.1088/0266-5611/10/2/009.

[12] G. Bao, L. Cowsar, and W. Masters. *Mathematical Modeling in Optical Science*. SIAM, 2001. DOI: 10.1137/1.9780898717594.

[13] J.-P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 114(2):185–200, 1994. DOI: 10.1006/jcph.1994.1159.

[14] M. Bixon and J. Jortner. Intramolecular radiationless transitions. *J. Chem. Phys.*, 48(2):715–726, 1968. DOI: 10.1063/1.1668703.

[15] F. Bloch. Über die Quantenmechanik der Elektronen in Kristallgittern. *Z. Physik*, 52:555–600, 1929. DOI: 10.1007/BF01339455.

[16] A.-S. Bonnet-Bendhia and F. Starling. Guided waves by electromagnetic gratings and non-uniqueness examples for the diffraction problem. *Math. Methods Appl. Sci.*, 17(5):305–338, 1994. DOI: 10.1002/mma.1670170502.

[17] L. Bourgeois and S. Fliss. On the identification of defects in a periodic waveguide from far field data. *Inverse Problems*, 30(9):095004, 31, 2014. DOI: 10.1088/0266-5611/30/9/095004.

[18] O. P. Bruno and F. Reitich. Numerical solution of diffraction problems: a method of variation of boundaries. *J. Opt. Soc. Am. A*, 10(6):1168–1175, 1993. DOI: 10.1364/JOSAA.10.002307.

[19] O. P. Bruno and F. Reitich. Solution of a boundary value problem for the Helmholtz equation via variation of the boundary into the complex domain. *Proc. R. Soc. Edinb. A: Math.*, 122(3-4):317–340, 1992. DOI: 10.1017/S0308210500021132.

[20] J. Bulling, B. Jurgelucks, J. Prager, and A. Walther. Experimental validation of an inverse method for defect reconstruction in a two-dimensional waveguide model. *J. Acoust. Soc. Am.*, 155(6):3794–3806, 2024. DOI: 10.1121/10.0025469.

[21] F. Cakoni, H. Haddar, and T.-P. Nguyen. Fast imaging of local perturbations in a unknown bi-periodic layered medium. *J. Comput. Phys.*, 501:Paper No. 112773, 19, 2024. DOI: 10.1016/j.jcp.2024.112773.

[22] F. Cakoni, H. Haddar, and T.-P. Nguyen. New interior transmission problem applied to a single Floquet-Bloch mode imaging of local perturbations in periodic media. *Inverse Problems*, 35(1):Paper No. 015009, 31, 2019. DOI: 10.1088/1361-6420/aaecfd.

[23] S. N. Chandler-Wilde and J. Elschner. Variational approach in weighted Sobolev spaces to scattering by unbounded rough surfaces. *SIAM J. Math. Anal.*, 42(6):2554–2580, 2010. DOI: 10.1137/090776111.

[24] S. N. Chandler-Wilde, E. Heinemeyer, and R. Potthast. Acoustic scattering by mildly rough unbounded surfaces in three dimensions. *SIAM J. Appl. Math.*, 66(3):1002–1026, 2006. DOI: 10.1137/050635262.

[25] S. N. Chandler-Wilde and P. Monk. Existence, uniqueness, and variational methods for scattering by unbounded rough surfaces. *SIAM J. Math. Anal.*, 37(2):598–618, 2005. DOI: 10.1137/040615523.

[26] S. N. Chandler-Wilde and P. Monk. The PML for rough surface scattering. *Appl. Numer. Math.*, 59(9):2131–2154, 2009. DOI: 10.1016/j.apnum.2008.12.007.

[27] S. N. Chandler-Wilde and B. Zhang. A uniqueness result for scattering by infinite rough surfaces. *SIAM J. Appl. Math.*, 58(6):1774–1790, 1998. DOI: 10.1137/S0036139996309722.

[28] Z. Chen and H. Wu. An adaptive finite element method with perfectly matched absorbing layers for the wave scattering by periodic structures. *SIAM J. Numer. Anal.*, 41(3):799–826, 2003. DOI: 10.1137/S0036142902400901.

[29] J. Coatléven. Helmholtz equation in periodic media with a line defect. *J. Comput. Phys.*, 231(4):1675–1704, 2012. DOI: 10.1016/j.jcp.2011.10.022.

[30] F. Collino and P. Monk. The perfectly matched layer in curvilinear coordinates. *SIAM J. Sci. Comput.*, 19(6):2061–2090, 1998. DOI: 10.1137/S1064827596301406.

[31] D. L. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. 4th edition. Vol. 93. Springer, Cham, 1998. DOI: 10.1007/978-3-030-30351-8.

[32] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *Bull. Sci. Math.*, 136(5):521–573, 2012. DOI: 10.1016/j.bulsci.2011.12.004.

[33] J. Diaz and P. Joly. A time domain analysis of PML models in acoustics. *Comput. Methods Appl. Mech. Eng.*, 195(29):3820–3853, 2006. DOI: 10.1016/j.cma.2005.02.031.

[34] J. J. Dongarra, I. S. Duff, D. C. Sorensen, H. A. Van der Vorst, et al. *Solving Linear Systems on Vector and Shared Memory Computers*. Vol. 10. Society for Industrial and Applied Mathematics Philadelphia, 1991.

[35] W. Dörfler, A. Lechleiter, M. Plum, G. Schneider, and C. Wieners. *Photonic Crystals: Mathematical Analysis and Numerical Approximation*. Vol. 42. Springer, Birkhäuser Basel, 2011. DOI: 10.1007/978-3-0348-0113-3.

[36] I. S. Duff, A. M. Erisman, and J. K. Reid. *Direct Methods for Sparse Matrices*. Oxford University Press, 2017. DOI: 10.1093/acprof:oso/9780198508380.001.0001.

[37] M. Ehrhardt, H. Han, and C. Zheng. Numerical simulation of waves in periodic structures. *Commun. Comput. Phys.*, 5:849–870, 2009.

[38] M. Ehrhardt, J. San, and C. Zheng. Evaluation of scattering operators for semi-infinite periodic arrays. *Commun. Comput. Phys.*, 7:347–364, 2009.

[39] J. Elschner, G. Schmidt, and M. Yamamoto. An inverse problem in periodic diffractive optics: global uniqueness with a single wavenumber. *Inverse Problems*, 19(3):779–787, 2003.

[40] J. Elschner and G. Hu. Global uniqueness in determining polygonal periodic structures with a minimal number of incident plane waves. *Inverse Problems*, 26(11):115002/1–115002/23, 2010. DOI: 10.1088/0266-5611/26/11/115002.

[41] J. Elschner and M. Yamamoto. An inverse problem in periodic diffractive optics: reconstruction of Lipschitz grating profiles. *Appl. Anal.*, 81(6):1307–1328, 2002. DOI: 10.1080/0003681021000035551.

[42] J. Elschner and M. Yamamoto. Uniqueness in determining polygonal periodic structures. *Z. Anal. Anwend.*, 26(2):165–177, 2007. DOI: 10.4171/ZAA/1316.

[43] A. Fichtner. *Full Seismic Waveform Modelling and Inversion*. Springer, Berlin, Heidelberg, 2010. DOI: 10.1007/978-3-642-15807-0.

[44] S. Fliss and P. Joly. Wave propagation in locally perturbed periodic media (case with absorption): Numerical aspects. *J. Comput. Phys.*, 231:1244–1271, 2012. DOI: 10.1016/j.jcp.2011.10.007.

[45] S. Fliss and P. Joly. Exact boundary conditions for time-harmonic wave propagation in locally perturbed periodic media. *Appl. Numer. Math.*, 59(9):2155–2178, 2009. DOI: 10.1016/j.apnum.2008.12.013.

[46] G. Floquet. Sur les équations différentielles linéaires à coefficients périodiques. *Annales scientifiques de l'École Normale Supérieure*, 12:47–88, 1883. DOI: 10.24033/asens.220.

[47] J. W. Gadzuk. Localized vibrational modes in Fermi liquids. General theory. *Phys. Rev. B*, 24:1651–1663, 1981. DOI: 10.1103/PhysRevB.24.1651.

[48] K. M. Giannoutakis and G. A. Gravvanis. High performance finite element approximate inverse preconditioning. *Appl. Math. Comput.*, 201(1):293–304, 2008. DOI: 10.1016/j.amc.2007.12.023.

[49] V. Girault and P.-A. Raviart. *Finite Element Approximation of the Navier–Stokes Equations*. Vol. 749. Springer, Berlin, 1979. DOI: 10.1007/BFb0063447.

[50] A. K. Goyal, H. S. Dutta, and S. Pal. Performance optimization of photonic crystal resonator based sensor. *Opt. Quantum Electron.*, 48(9):431, 2016. DOI: 10.1007/s11082-016-0701-0.

[51] G. A. Gravvanis. High performance inverse preconditioning. *Arch. Comput. Methods Eng.*, 16:77–108, 2009. DOI: 10.1007/s11831-008-9026-x.

[52] G. A. Gravvanis. Solving symmetric arrowhead and special tridiagonal linear systems by fast approximate inverse preconditioning. *J. Math. Model. Algorithms*, 1:269–282, 2002. DOI: 10.1023/A:1021630031889.

[53] G. A. Gravvanis. An approximate inverse matrix technique for arrowhead matrices. *Int. J. Comput. Math.*, 70(1):35–45, 1998.

[54] J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Yale University Press, 1923.

[55] H. Haddar and T. Nguyen. A volume integral method for solving scattering problems from locally perturbed infinite periodic layers. *Appl. Anal.*, 96:130–158, 2017. DOI: 10.1080/00036811.2016.1221942.

[56] F. Hettlich. Iterative regularization schemes in inverse scattering by periodic structures. *Inverse Problems*, 18(3):701–714, 2002. DOI: 10.1088/0266-5611/18/3/311.

[57] F. Hettlich and A. Kirsch. Schiffer's theorem in inverse scattering theory for periodic structures. *Inverse Problems*, 13(2):351–361, 1997. DOI: 10.1088/0266-5611/13/2/010.

[58] L. Hörmander. *Introduction to Complex Analysis in Several Variables*. Birkhäuser Cham, 1979. DOI: 10.1007/978-3-031-26428-3.

[59] G. Hu and A. Kirsch. Direct and inverse time-harmonic scattering by Dirichlet periodic curves with local perturbations. 2024. Preprint. DOI: 10.48550/arXiv.2403.07340.

[60] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*. Springer, New York, 1998. DOI: 10.1007/b98828.

[61] J. D. Joannopoulous, S. G. Johnson, J. Winn, and R. D. Meade. Photonic crystals: Molding the flow of light. 2008. DOI: 10.2307/j.ctvcm4gz9.

[62] S. Johnson and J. Joannopoulos. *Photonic Crystals: The Road from Theory to Practice*. Springer, New York, 2002.

[63] P. Joly, J.-R. Li, and S. Fliss. Exact boundary conditions for periodic waveguides containing a local perturbation. *Commun. Comput. Phys.*, 1:945–973, 2006.

[64] B. Kaltenbacher, A. Neubauer, and O. Scherzer. *Iterative regularization methods for nonlinear ill-posed problems*. Vol. 6. Radon Series on Computational and Applied Mathematics. Walter de Gruyter GmbH & Co. KG, Berlin, 2008. DOI: 10.1515/9783110208276.

[65] A. Kirsch and P. Monk. Convergence analysis of a coupled finite element and spectral method in acoustic scattering. *IMA J. Numer. Anal.*, 10(3):425–447, 1990. DOI: 10.1093/imanum/10.3.425.

[66] A. Kirsch. Diffraction by periodic structures. *Inverse Problems in Mathematical Physics*. Ed. by L. Päivärinta and E. Somersalo. Vol. 42. Lecture Notes in Physics. Springer, Berlin, Heidelberg, 1993.

[67] A. Kirsch. On the scattering of a plane wave by a perturbed open periodic waveguide. *Math. Methods Appl. Sci.*, 46(9):10698–10718, 2023. DOI: 10.1002/mma.9147.

[68] A. Kirsch. The domain derivative and two applications in inverse scattering theory. *Inverse Problems*, 9(1):81–96, 1993. DOI: 10.1088/0266-5611/9/1/005.

[69] A. Kirsch. Uniqueness theorems in inverse scattering theory for periodic structures. *Inverse Problems*, 10(1):145–152, 1994. DOI: 10.1088/0266-5611/10/1/011.

[70] A. Kirsch and F. Hettlich. *The mathematical theory of time-harmonic Maxwell's equations, volume 190 of Applied Mathematical Sciences*. Springer, Cham, 2015. DOI: 10.1007/978-3-319-11086-8.

[71] A. Kirsch and R. Zhang. The PML-method for a scattering problem for a local perturbation of an open periodic waveguide. *Numer. Math.*, 157(2):717–748, 2025. DOI: 10.1007/s00211-025-01456-9.

[72] M. Knöller. Electromagnetic scattering from thin tubular objects and an application in electromagnetic chirality. PhD thesis. Karlsruher Institut für Technologie, Karlsruhe, 2023. DOI: 10.5445/IR/1000161368.

[73] A. Konschin. Direkte und inverse elektromagnetische Streuprobleme fur lokal gestörte periodische Medien. PhD thesis. Universität Bremen, Bremen, 2019.

[74] A. Konschin. Electromagnetic wave scattering from locally perturbed periodic inhomogeneous layers. *Math. Methods Appl. Sci.*, 44(18):14126–14147, 2021. DOI: 10.1002/mma.7680.

[75] R. Kress. *Linear Integral Equations*. 3rd edition. Springer, New York, 2014. DOI: 10.1007/978-1-4614-9593-2.

[76] R. Kress. *Numerical Analysis*. Springer, New York, 2012. DOI: 10.1007/978-1-4612-0599-9.

[77] P. A. Kuchment. *Floquet Theory for Partial Differential Equations*. Vol. 60. Birkhäuser, Basel, 2012. DOI: 10.1007/978-3-0348-8573-7.

[78] H. T. Kung and B. W. Suter. A hub matrix theory and applications to wireless communications. *EURASIP J. Adv. Signal Process.*, 2007(1):Art. ID 13659, 2007. DOI: 10.1155/2007/13659.

[79] M. Lassas and E. Somersalo. Analysis of the PML equations in general convex geometry. *Proc. Roy. Soc. Edinburgh Sect. A*, 131(5):1183–1207, 2001. DOI: 10.1017/S0308210500001335.

[80] M. Lassas and E. Somersalo. On the existence and convergence of the solution of PML equations. *Computing*, 60(3):229–241, 1998. DOI: 10.1007/BF02684334.

[81] A. Lechleiter. The Floquet–Bloch transform and scattering from locally perturbed periodic surfaces. *J. Math. Anal. Appl.*, 446(1):605–627, 2017. DOI: 10.1016/j.jmaa.2016.08.055.

[82] A. Lechleiter and D.-L. Nguyen. Scattering of Herglotz waves from periodic structures and mapping properties of the Bloch transform. *Proc. R. Soc. Edinb. A: Math.*, 145:1283–1311, 2015. DOI: 10.1017/S0308210515000335.

[83] A. Lechleiter and R. Zhang. A Floquet-Bloch transform based numerical method for scattering from locally perturbed periodic surfaces. *SIAM J. Sci. Comput.*, 39(5):B819–B839, 2017. DOI: 10.1137/16M1104111.

[84] A. Lechleiter and R. Zhang. Non-periodic acoustic and electromagnetic, scattering from periodic structures in 3D. *Comput. Math. with Appl.*, 74(11):2723–2738, 2017. DOI: 10.1016/j.camwa.2017.08.042.

[85] A. Lechleiter and R. Zhang. A convergent numerical scheme for scattering of aperiodic waves from periodic surfaces based on the Floquet-Bloch transform. *SIAM J. Numer. Anal.*, 55(2):713–736, 2017. DOI: 10.1137/16M1067524.

[86] A. Lechleiter and R. Zhang. Reconstruction of local perturbations in periodic surfaces. *Inverse Problems*, 34(3):035006, 17, 2018. DOI: 10.1088/1361-6420/aaa7b1.

[87] F. J. Margotti. On Inexact Newton Methods for Inverse Problems in Banach Spaces. PhD thesis. Karlsruher Institut für Technologie, Karlsruhe, 2015. DOI: 10.5445/IR/1000048606.

[88] W. McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, Cambridge, 2000.

[89] A. Meier, T. Arens, S. N. Chandler-Wilde, and A. Kirsch. A Nyström method for a class of integral equations on the real line with applications to scattering by diffraction gratings and rough surfaces. *J. Integral Equations Appl.*, 12(3):281–321, 2000. DOI: 10.1216/jiea/1020282209.

[90] P. Monk. *Finite Element Methods for Maxwell's Equations*. Oxford University Press, New York, 2003. DOI: 10.1093/acprof:oso/9780198508885.001.0001.

[91] H. S. Najafi, S. Edalatpanah, and G. A. Gravvanis. An efficient method for computing the inverse of arrowhead matrices. *Appl. Math. Lett.*, 33:1–5, 2014. DOI: 10.1016/j.aml.2014.02.010.

[92] D. P. Nicholls and F. Reitich. Shape deformations in rough-surface scattering: improved algorithms. *J. Opt. Soc. Am. A*, 21(4):606–621, 2004. DOI: 10.1364/JOSAA.21.000606.

[93] NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov/, Release 1.2.3 of 2024-12-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[94] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 1999. DOI: 10.1007/978-0-387-40065-5.

[95] A. Rieder. On the regularization of nonlinear ill-posed problems via inexact Newton iterations. *Inverse Problems*, 15(1):309–327, 1999. DOI: 10.1088/0266-5611/15/1/028.

[96] G. Schmidt. On the Diffraction by Biperiodic Anisotropic Structures. *Appl. Anal.*, 82(1):75–92, 2003. DOI: 10.1080/0003681031000068275.

[97] Y. Shestopalov, Y. Smirnov, and E. Smolkin. *Optical waveguide theory*. Vol. 237. Springer, Singapore, 2022. DOI: 10.1007/978-981-19-0584-1.

[98] P. S. Stanimirović, V. N. Katsikis, and D. Kolundžija. Inversion and pseudoinversion of block arrowhead matrices. *Appl. Math. Comput.*, 341:379–401, 2019. DOI: 10.1016/j.amc.2018.09.006.

[99] F. Trèves. *Basic linear partial differential equations*. Reprint of the 1975 original. Dover Publications, Inc., Mineola, NY, 2006.

[100] A. Voronovich. *Wave Scattering from Rough Surfaces*. 2nd edition. Vol. 17. Springer, Berlin, 1999. DOI: 10.1007/978-3-642-59936-1.

[101] Q. Yi, C. Zhao, and P. Wang. Characteristics of defect states in periodic railway track structure. *J. low freq. noise vib. act. control*, 41(1):196–208, 2022. DOI: 10.1177/146134842110382.

[102] L. Yuan and Y. Y. Lu. A recursive-doubling Dirichlet-to-Neumann-map method for periodic waveguides. *J. Light. Technol.*, 25(11):3649–3656, 2007.

[103] Y. Yuan. Generalized inverse eigenvalue problems for symmetric arrow-head matrices. *Int. J. Comput. Math. Sci.*, 4(6):268–271, 2010.

[104] R. Zhang. A high order numerical method for scattering from locally perturbed periodic surfaces. *SIAM J. Sci. Comput.*, 40(4):A2286–A2314, 2018. DOI: 10.1137/17M1144945.

[105] R. Zhang. Exponential convergence of perfectly matched layers for scattering problems with periodic surfaces. *SIAM J. Numer. Anal.*, 60(2):804–823, 2022. DOI: 10.1137/21M1439043.

[106] R. Zhang. Fast convergent PML method for scattering with periodic surfaces: the exceptional case. Preprint. 2022. DOI: 10.48550/arXiv.2211.01229.

Notations

BASIC NOTATION

$\mathbb{R}^d$	d-dimensional real Euclidean space	11
$\widetilde{x}$	point $\widetilde{x} = (x_1, \ldots, x_{d-1})^\top$ in $\mathbb{R}^{d-1}$	11
x	point $x = (\widetilde{x}, x_d)^\top$ in $\mathbb{R}^d$	11
Ξ	complex stretched coordinate	19
k	wave number	2
κ	cutoff value	68
Σ	set of singular points	36
$\mathbf{J}$	all centers of circular arcs	36
λ	physical width of the PML	19
σ	virtual width of the PML	24
α	Floquet parameter	34
C_{infsup}	inf-sup constant	16
Λ	fundamental domain in α-space	28
α_{reg}	regularization parameter	111

FUNCTION SPACES

$L^s(\mathbb{R}^d)$	Lebesgue space of p-integrable functions on $\mathbb{R}^d$	10
$H^s(\mathbb{R}^d)$	Sobolev space of order s on $\mathbb{R}^d$	8
$H^s_\alpha(\mathbb{R}^d)$	Sobolev space of α-quasiperiodic functions on $\mathbb{R}^d$	9
$H^s_{\text{per}}(\mathbb{R}^d)$	Sobolev space of periodic functions on $\mathbb{R}^d$	9
$H^s_r(\mathbb{R}^d)$	Sobolev space of order s and decay rate r on $\mathbb{R}^d$	8
$H^s_r(\Omega)$	functions in $H^s_r(\mathbb{R}^d)$ restricted to the Lipschitz domain Ω	9
$\widetilde{H}^s_r(\Omega)$	functions in $\widetilde{H}^s_r(\Omega)$ which are zero on a subset of $\partial\Omega$	10
$H^1(\Delta, \Omega)$	functions in $H^1(\Omega)$ whose Laplacian is in $L^2(\Omega)$	10
$C^\omega(U; V)$	space of analytic functions from U to V	37
$(X(\Omega))^*$	dual space of $X(\Omega)$	7
$\langle \cdot, \cdot \rangle_\Omega$	bilinear dual pairing between $(X(\Omega))^*$ and $X(\Omega)$	8
$\mathcal{S}^*(\mathbb{R}^d)$	space of temperate distributions	8

GEOMETRY

ζ	generic function generating the bottom surface	2, 19
Γ	generic unbounded surface defined by ζ	1, 10
Ω	generic unbounded domain above the surface Γ	2, 10
Γ_H	flat surface parallel to the surface Γ at height H	11
Ω_H	unbounded domain between Γ and Γ_H	11
Ω_H^+	exterior domain above the surface Γ_H	11
ζ^{per}	periodic function	27
Γ^{per}	periodic surface generated by ζ^{per}	27
Ω^{per}	unbounded periodic domain above the surface Γ^{per}	33
Ω_H^{per}	unbounded periodic domain between Γ^{per} and Γ_H	27
δ	compactly supported perturbation	59
ζ^δ	locally perturbed function, sum of ζ^{per} and δ	59
Γ^δ	locally perturbed surface generated by the function ζ^δ	59
Ω^δ	unbounded locally perturbed domain above the surface Γ^δ	59
Ω_H^δ	unbounded perturbed domain between Γ^δ and Γ_H	59
$\Omega_H^{2\pi}$	bounded cell	27
$\Gamma^{2\pi}$	bottom surface of the bounded cell $\Omega_H^{2\pi}$	27
$\Gamma_H^{2\pi}$	top surface of the bounded cell $\Omega_H^{2\pi}$	27
Γ_-, Γ_+	lateral boundaries of the bounded cell $\Omega_H^{2\pi}$	27
$\Gamma_{H+\lambda}$	absorbing surface parallel to Γ at height $H+\lambda$	19
$\Omega_{H+\lambda}$	extended domain containing the PML	19
Ω_{PML}	PML region with thickness λ	19

FUNCTIONS

u^i	incident field	34
u^s	scattered field	54
u	total field	34
u_σ	PML approximation of the total field u in Ω_H^{per}	24
$u^{s,\delta}$	scattered field in the perturbed domain Ω_H^δ	60
$u_\sigma^{s,\delta}$	PML approximation of the scattered field $u^{s,\delta}$ in Ω_H^δ	63
u^δ	total field defined in the perturbed domain Ω_H^δ	60
u_{tra}^δ	total field u^δ transformed to the periodic domain Ω_H^{per}	61
u_σ^δ	PML approximation of the total field u^δ in Ω_H^δ	142
$u_{\mathrm{tra},\sigma}^\delta$	PML approximation of u_σ^δ transformed to Ω_H^{per}	64
w^δ	Floquet–Bloch transform of u_{tra}^δ	62
w_σ^δ	Floquet–Bloch transform of $u_{\mathrm{tra},\sigma}^\delta$	64
Ψ^δ	diffeomorphism mapping periodic to perturbed domain	61
Φ	fundamental solution of the Helmholtz equation	13
$H_0^{(1)}$	Hankel function of the first kind of order zero	13

OPERATORS

$\mathcal{F}$	Fourier transform	7
$\mathcal{J}$	Floquet-Bloch transform	28
γ_{D}	trace operator	10, 98
γ_{N}	conormal derivative	10
$\mathcal{T}^+$	Dirichlet-to-Neumann operator	15
$\mathcal{T}^+_\alpha$	Floquet–Bloch transform of $\mathcal{T}^+$	35
Δ_{PML}	PML operator	23
$\mathcal{T}^+_\sigma$	PML approximation of $\mathcal{T}^+$	24
$\mathcal{T}^+_{\alpha,\sigma}$	Floquet–Bloch transform of $\mathcal{T}^+_\sigma$	64
$\mathcal{S}$	scattering operator	93
$\mathcal{S}'$	Fréchet derivative of the scattering operator	99
$\mathscr{I}$	compact embedding operator	98

Index