Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Radiation conditions for the Helmholtz equation in a half plane filled by inhomogeneous periodic material

Guanghui Hu[1], Andreas Rathsfeld[2]

submitted: May 29, 2020

[1] School of Mathematical Sciences
Nankai University
300071 Tianjin
China
Email: ghhu@nankai.edu.cn

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: andreas.rathsfeld@wias-berlin.de

No. 2726
Berlin 2020

2010 *Mathematics Subject Classification.* 74J20, 76B15, 35J50, 35J08.

Key words and phrases. Half-space radiation condition, inhomogeneous medium, wave-mode expansion, scattering by grating, RCWA.

Part of this work was carried out when G. Hu visited the Weierstrass Institute in October 2019. The hospitality of the institute is greatly appreciated.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Radiation conditions for the Helmholtz equation in a half plane filled by inhomogeneous periodic material

Guanghui Hu, Andreas Rathsfeld

Abstract

In this paper we consider time-harmonic acoustic wave propagation in a half-plane filled by inhomogeneous periodic medium. If the refractive index depends on the horizontal coordinate only, we define upward and downward radiating modes by solving a one-dimensional Sturm-Liouville eigenvalue problem with a complex-valued periodic coefficient. The upward and downward radiation conditions are introduced based on a generalized Rayleigh series. Using the variational method, we then prove uniqueness and existence for the scattering of an incoming wave mode by a grating located between an upper and lower half plane with such inhomogeneous periodic media. Finally, we discuss the application of the new radiation conditions to the scattering matrix algorithm, i.e., to rigorous coupled wave analysis or Fourier modal method.

1 Introduction

Since Lord Rayleigh's original work [28] in 1907, time harmonic scattering problems by periodic and even by biperiodic gratings are well studied in both the physical and mathematical communities. The theory provides a Rayleigh expansion radiation condition over the half plane filled by homogeneous material. Using this, the acoustic, elastic and electromagnetic diffraction problems have been studied extensively concerning theoretical analysis and numerical approximation using integral equation and variational methods (cf. e.g. [1, 4, 5, 7, 8, 10–12, 23, 31, 36, 37]). We refer to [6, 32–34] for historical remarks and details of engineering applications, if the cover material in the half spaces above and the substrate material below the periodic surface structure of the grating is supposed to be homogeneous. However, special inhomogeneous materials are possible in applications. For instance, in the design of photonic crystals, the refractive index corresponding to materials of interest is a periodic function in different spatial directions. This paper is devoted to new radiation conditions for the Helmholtz equation and the corresponding solvability theory. This theory applies to the analysis of the scattering matrix algorithm even for the solution of classical scattering problems with homogeneous cover and substrate material.

To start the analysis, we consider the case of periodic gratings in the two-dimensional space contained in the layer $\{(x_1,x_2)^\top\in\mathbb{R}^2:\ b\le x_2\le d\}$, where the refractive index $(x_1,x_2)^\top\mapsto\mathrm{ind}(x_1)$ in the half planes $\{(x_1,x_2)^\top\in\mathbb{R}^2:\ d\le x_2\}$ of cover material and $\{(x_1,x_2)^\top\in\mathbb{R}^2:\ x_2\le b\}$ of substrate material is independent of the vertical x_2 and a periodic function with respect to the horizontal x_1. We

DOI 10.20347/WIAS.PREPRINT.2726 Berlin 2020

assume $\mathrm{ind}(x_1+p)=\mathrm{ind}(x_1)$ with the same period p as for the grating structure. Similarly to the homogeneous case, the radiation condition for these half planes is defined by expansions into a Rayleigh series of upgoing and downgoing wave modes. However, the wave modes will be of the form $(x_1, x_2)^\top \mapsto \exp(\lambda x_2)h(x_1)$, where λ is an eigenvalue and h an eigenfunction or a linear combination of associated eigenfunctions of a Sturm-Liouville differential operator. Using these natural conditions, we can show the Fredholm property for the boundary value problem modeling the scattering of an incoming wave mode by the grating. Uniqueness is shown for the propagating reflected and transmitted wave modes. The full solution is unique if the grating contains absorbing materials.

Our research is closest to the recent work [26], where a technical outgoing radiation condition was proposed to analyze the transmission problem between free space and an unbounded photonic crystal. In comparison with [26], we assume that the inhomogeneous material is invariant along the vertical coordinate x_2, leading to more explicit upward and downward radiating modes and stronger uniqueness and existence results. The methodology used in this work differs from other scattering problems arising from closed periodic waveguides [14] (see also [13]), infinite periodic cylinders [25] and in stratified media [24], which rely essentially on Floquet-Bloch transform and the limiting absorption principle. The materials in the aforementioned works are usually assumed to be periodic inside the waveguide and to be identical in the exterior, whereas in our settings, the inhomogeneous periodic material occupies a half plane. We also refer to [2, 3, 19, 35] for earlier studies on radiating modes in open and semi-infinite waveguides.

One of the most popular numerical methods for the classical periodic gratings is the scattering matrix algorithm (SMA), which in its various versions is called rigorous coupled wave analysis (RCWA) or Fourier modal method (FMM) (cf. e.g. [17, 18, 27, 30, 32, 33]). In the two-dimensional case, the Helmholtz equation is considered as an ordinary differential equation (ODE) with respect to the height x_2 over the surface, where the solution takes values in function spaces with respect to the horizontal variable. A clever numerical algorithm has been designed to integrate the ODE. A partition of the grating domain into slices (layers) parallel to the surface is introduced, the Helmholtz equation is solved over each slice, and the coupling through the common boundary of neighbour slices is realized by a stable recursive iteration. The discretization in the horizontal direction is based on Fourier series expansions.

Unfortunately, there is no analysis available so far. The technique of ODEs is difficult to apply since differential operators with piecewise constant coefficients act on the horizontal functions. Instead, the spaces and theorems for the Helmholtz equations should be used. On the planar upper and lower boundaries of the slices an expansion into upgoing and downgoing wave modes is used. In other words, there appear the above mentioned radiation conditions for inhomogeneous media. The S-matrices appearing in the recursive iteration are nothing else than the discretized boundary potentials for the Helmholtz solvers over the slice, which map the waves incoming to the slices to the reflected and transmitted waves. So the following program is the natural approach: The recursive iteration should be considered on the non-discretized level. The results on boundary values problems including inhomogeneous cover or substrate material should be used for the non-discretized S-matrices, to derive conditions for the applicability of the non-discretized scattering matrix algorithm. Afterwards, the discretization in form of RCWA or FMM should be discussed. We shall address only a few of the problems.

For instance, a reliable numerical algorithm might have to deal with the existence of wave modes including associated eigenvalues of rank larger than one. It might have to deal with the case that some operator, which is discretized and inverted, is only a Fredholm operator.

We introduce the inhomogeneous half spaces with cover and substrate material as well as the corresponding boundary value problems in Sect. 2. In Sect. 3, supposing non-absorbing materials, we define the radiation condition by Fourier series expansion with respect to x_1 and by solving a function valued ODE with techniques of the functional analysis. Alternatively, we solve the ODE with operator valued coefficient by an eigenvalue decomposition for this coefficient operator acting on quasiperiodic functions with respect to x_1. In the Subsects. 4.2 and 4.3 we discuss the eigenvalues, eigenfunctions, and associated eigenfunctions for the coefficient operator, which is a Sturm-Liouville operator. This decomposition is used to define upward and downward radiating wave modes and the radiation condition in Subsect. 4.4. In Sect. 5 we introduce the boundary value problem for gratings between an upper and lower half space of inhomogeneous media. We present the variational formulation and analyze the solution. Sect. 6 introduces the scattering matrix algorithm, shows the connection to the boundary value problems of Sect. 5, and addresses some of the problems for the numerical algorithm.

2 Quasiperiodic boundary value problem in an inhomogeneous half space

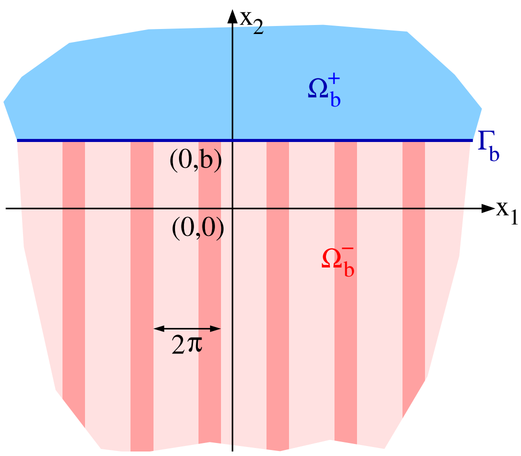

Figure 1: The geometry settings.

Denoting the points in two-dimensional space by $x=(x_1,x_2)^\top$, we suppose that the lower half space $\Omega_b^- := \{x\in\mathbb{R}^2 :\ x_2<b\}$ is illuminated by an incoming wave from the upper half space $\Omega_b^+ := \{x\in\mathbb{R}^2 :\ x_2>b\}$ with the wave number $k>0$. In this paper it is assumed that Ω_b^- is occupied by an inhomogeneous periodic medium modeled by the squared refractive index $q\in L^\infty(\Omega_b^-)$ (cf. Fig. 1). Further, q is assumed to be independent of x_2 and 2π-periodic in x_1, i.e.,

$$q(x) = q(x_1),\ q(x_1+2\pi n) = q(x_1) \qquad \text{for a.e. } x_1 \in \mathbb{R} \text{ and all } n \in \mathbb{Z}. \tag{2.1}$$

For physical reasons, we suppose that there is a $c_q > 0$ such that either $q(x_1) \geq c_q$ or $\operatorname{Im} q(x_1) \geq c_q$ for a.e. $x_1 \in \mathbb{R}$.

Then the time-harmonic acoustic wave propagation in Ω_b^- is governed by the Helmholtz equation $\Delta u + k^2 q u = 0$ in Ω_b^-, where $u = u(x)$ denotes the acoustic pressure or a component of an electromagnetic field. Since the lower half space in unbounded, we need a radiation condition of u as $x_2 \to -\infty$ to ensure well-posedness of the scattering problem. To mathematically formulate the scattering problem, we need the concept of quasiperiodic functions and Sobolev spaces.

Definition 2.1. *The function u is called quasiperiodic in x_1 with the parameter $\alpha \in [0,1)$ (that is, α-quasiperiodic), if $x_1 \mapsto u(x_1, x_2) e^{-i\alpha x_1}$ is 2π-periodic in x_1 for any fixed x_2.*

Define the quasiperiodic Sobolev spaces on Ω_b^- and $\mathbb{R}$ by

$$\begin{aligned} H^1_\alpha(\Omega_b^-) &:= \{u \in H^1_{loc}(\Omega_b^-) : u \text{ is } \alpha\text{-quasiperiodic in } x_1\} \\ H^{1/2}_\alpha(\mathbb{R}) &:= \{f \in H^{1/2}_{loc}(\mathbb{R}) : e^{-i\alpha x_1} f(x_1) \text{ is } 2\pi\text{-periodic in } x_1\}. \end{aligned}$$

Note that our $H^1_{loc}(\Omega_b^-)$ is the space of all functions u over Ω_b^- such that, for any radius $r > 0$, the restriction of u to $\Omega_{b,r}^- := \{x \in \Omega_b^- : |x| < r\}$ is in $H^1(\Omega_{b,r}^-)$. If the incoming wave is a plane wave of the form $u^{in}(x) := \exp(ik(x_1 \sin\theta - x_2 \cos\theta))$ with the incident angle $\theta \in (-\pi/2, \pi/2)$, we set $\alpha_0 := k \sin\theta$ and get an α-quasiperiodic function u^{in} with α the unique number such that $\alpha \in [0,1)$ and $\alpha - \alpha_0$ is an integer. In the case $q \equiv 1$ in Ω_b^-, we recall that a Helmholtz solution u is called downward radiating if u admits a Rayleigh expansion (see, e.g., [1, 12, 23])

$$u(x) = \sum_{n \in \mathbb{Z}} c_n \, e^{i(\alpha_n x_1 - \beta_n (x_2 - b))}, \qquad x_2 < b, \tag{2.2}$$

where the $c_n \in \mathbb{C}$ are called Rayleigh coefficients and

$$\alpha_n := n + \alpha_0, \quad \beta_n := \begin{cases} \sqrt{k^2 - \alpha_n^2} & \text{if } |\alpha_n| \leq k, \\ i\sqrt{\alpha_n^2 - k^2} & \text{if } |\alpha_n| > k. \end{cases} \tag{2.3}$$

The existence of coefficients c_n with Equ. (2.2) is called the radiation condition for the lower half plane Ω_b^-. The upward radiation condition in Ω_b^+ filled by a homogeneous medium can be defined analogously. Obviously, the Rayleigh expansion (2.2) consists of a finite number of propagating waves corresponding to n with $|\alpha_n| \leq k$ and an infinite number of evanescent waves for $|\alpha_n| > k$, which decay exponentially when $|x_2| \to \infty$. It has been widely used in the literature to prove well-posedness and design numerical schemes for time-harmonic acoustic, elastic and electromagnetic scattering by periodic surface structures located between half spaces occupied by homogeneous media [1, 4, 5, 7, 8, 10–12, 21–23, 31]. One of the main subjects of the present paper is to define downward and upward radiation conditions in an inhomogeneous medium, which will generalize the above Rayleigh expansion from a homogeneous periodic medium to the inhomogeneous case of (2.1).

Consider the boundary value problem in an inhomogeneous half space

$$\text{(BVP):} \quad \begin{cases} \Delta u + k^2 q u = 0 & \text{in } \Omega_b^-, \\ u = f & \text{on } \Gamma_b := \{x \in \mathbb{R}^2 : x_2 = b\}, \end{cases} \tag{2.4}$$

where $f \in H_\alpha^{1/2}(\mathbb{R})$. We shall define an 'appropriate' downward radiation condition over Ω_b^- and prove, under some additional assumptions, that the boundary value problem (2.4) combined with the radiation condition has a unique solution $u \in H_\alpha^1(\Omega_b^-)$ for any given $f \in H_\alpha^{1/2}(\Gamma_b)$.

To get a first version of a Rayleigh expansion in an inhomogeneous medium, we look at the Fourier expansion of the solution. Since u is α-quasiperiodic, it admits the expansion

$$e^{-i\alpha x_1} u(x_1, x_2) = \sum_{n\in\mathbb{Z}} u_n(x_2) e^{inx_1}, \quad x_2 < b,$$

or equivalently,

$$u(x_1, x_2) = \sum_{n\in\mathbb{Z}} u_n(x_2) e^{i\alpha_n x_1}, \quad x_2 < b. \tag{2.5}$$

Inserting (2.5) into the Helmholtz equation we find that

$$\sum_{n\in\mathbb{Z}} \left[u_n''(x_2) + \left(k^2 q(x_1) - \alpha_n^2\right) u_n(x_2) \right] e^{i\alpha_n x_1} = 0. \tag{2.6}$$

Thus the coefficients u_n are solutions of $u_n''(x_2) + (k^2 q(x_1) - \alpha_n^2)\, u_n(x_2) = 0$. In other words, in the inhomogeneous case we have to replace the Rayleigh modes $e^{i(\alpha_n x_1 - \beta_n(x_2 - b))}$ in (2.2) by $e^{i\alpha_n x_1} u_n(x_2)$ with u_n the solution of a second-order ODE.

3 Radiation condition for real-valued potentials

In this section we suppose that the squared refractive index function q with $q(x) = q(x_1)$ and with $q \in L^\infty(0, 2\pi)$ is real-valued. Now we shall show that the Helmholtz equation is equivalent to an ODE in the space of sequences of Fourier coefficients.

In order to introduce norms for the trace of the solution to the boundary value problem (2.4), we may expand the Dirichlet data $f = u|_{\Gamma_b}$ into the Fourier series

$$f(x_1) = \sum_{n\in\mathbb{Z}} f_n \, e^{i\alpha_n x_1}, \quad f_n \in \mathbb{C}.$$

We introduce the weighted ℓ^2 space of sequences

$$X^s := \left\{ \mathbf{a} = (a_n)_{n\in\mathbb{Z}} : \sum_{n\in\mathbb{Z}} (1+n^2)^s |a_n|^2 < \infty \right\}$$

endowed with the inner product and norm

$$\langle \mathbf{a}, \mathbf{b} \rangle_s := \sum_{n\in\mathbb{Z}} (1+n^2)^s a_n \, \bar{b}_n, \quad \|a\|_{X^s} := \sqrt{\sum_{n\in\mathbb{Z}} (1+n^2)^s |a_n|^2}.$$

Then X^s is a Hilbert space for any $s\in\mathbb{R}$. The Fourier coefficients of f satisfy

$$\|f\|_{H_\alpha^{1/2}(\Gamma_b)} = \|\mathbf{f}\|_{X^{1/2}} < \infty, \quad \mathbf{f} := (f_n)_{n\in\mathbb{Z}}.$$

Applying Fourier expansion to the refractive index function, we have

$$q(x_1) = \sum_{m\in\mathbb{Z}} q_m e^{imx_1}, \quad q_m \in \mathbb{C}.$$

Obviously, we would have $q\equiv q_0$ if the medium of Ω_b^- is homogeneous. Inserting the above expansion into (2.6), it follows that

$$\sum_{n\in\mathbb{Z}} \left[\Big(u_n''(x_2) - \alpha_n^2 u_n(x_2) \Big) e^{i\alpha_n x_1} + k^2 \sum_{m\in\mathbb{Z}} q_m e^{i\alpha_{n+m}x_1} u_n(x_2) \right] = 0,\ x \in \Omega_b^-.$$

Multiplying the previous equation by $e^{-i\alpha_j x_1}$ and integrating over $(0,2\pi)$ with respect to x_1 lead to

$$u_j'' - \alpha_j^2 u_j + k^2 \sum_{m\in\mathbb{Z}} q_{j-m} u_m = 0, \quad j \in \mathbb{Z}.$$

We set $U(x_2) := (\cdots, u_{-1}(x_2), u_0(x_2), u_1(x_2), \cdots)$. Since the function $x_1 \mapsto u(x_1,x_2)$ is in $H_\alpha^{1/2}(\mathbb{R})$ for any $x_2 \le b$, it holds that $U(x_2) \in X^{1/2}$ for any fixed $x_2 \le b$. The previous equations can be rewritten as a second-order ODE in the form

$$U''(x_2) + A\,U(x_2) = 0, \quad x_2 < b, \tag{3.1}$$

where $A := (a_{jm})_{j,m\in\mathbb{Z}}$ is an infinite dimensional matrix, whose entries are given by

$$a_{jm} := \begin{cases} k^2 q_{j-m} & \text{if } j \ne m, \\ -\alpha_j^2 + k^2 q_0 & \text{if } j = m. \end{cases}$$

The matrix A can be written as $A = B + k^2\,C$, where $B := (b_{j,m})_{j,m\in\mathbb{Z}}$ is the diagonal matrix and $C := (c_{j,m})_{j,m\in\mathbb{Z}}$ the Toeplitz matrix defined by

$$b_{j,m} := \begin{cases} 0 & \text{if } j \ne m, \\ -\alpha_j^2 & \text{if } j = m. \end{cases} \qquad c_{j,m} := q_{j-m}.$$

Evidently, the operator $B\colon X^{1/2} \to X^{-1/2}$ is bounded. The embedding theorems together with the fact that $q \in L^\infty(0,2\pi)$ imply that the operator $C\colon X^{1/2} \to X^{-1/2}$ is compact. Since q is real-valued, we have $q_m = \bar{q}_{-m}$. It then follows that the matrix $A\colon X^{1/2} \to X^{-1/2}$ is a linear self-adjoint operator. Moreover, the spectrum $\sigma(A)$ of A is real.

Now the solution of the ODE (3.1) follows the classical theory of linear ODEs with constant coefficients. By the spectral theorem, we may express A as an integral over the spectrum with respect to a projection-valued measure, that is,

$$A = \int_{\sigma(A)} \lambda \, dP_\lambda.$$

For simplicity assume that $0\notin\sigma(A)$. We define $\chi_{\mathbb{R}^\pm}:\mathbb{R}\to\mathbb{R}$ to be the characteristic function of the half line $\mathbb{R}^\pm$ and

$$A^\pm:=\int_{\sigma(A)}\chi_{\mathbb{R}^\pm}(\lambda)\lambda\,dP_\lambda,\quad \sqrt{A^\pm}:=\int_{\sigma(A)}\chi_{\mathbb{R}^\pm}(\lambda)\sqrt{\pm\lambda}\,dP_\lambda.$$

Evidently, we have $A=A^+ + A^-$ and $\sqrt{A}=\sqrt{A^+}+i\sqrt{A^-}$. The general solution to (3.1) is of the form

$$\begin{aligned} U(x_2) &= e^{i\sqrt{A}\,x_2}\mathbf{a}^+ + e^{-i\sqrt{A}\,x_2}\mathbf{a}^- \\ &= (e^{i\sqrt{A^+}\,x_2}+e^{-\sqrt{A^-}\,x_2})\mathbf{a}^+ + (e^{-i\sqrt{A^+}\,x_2}+e^{\sqrt{A^-}\,x_2})\,\mathbf{a}^- \end{aligned} \tag{3.2}$$

with $\mathbf{a}^\pm\in X^{1/2}$ and with $e^{\pm i\sqrt{A}\,x_2}$ to be understood as the exponential of an operator. In fact, straightforward calculations show that

$$\begin{aligned} (e^{i\sqrt{A^\pm}x_2}\mathbf{a}^\pm)'' &= -A^\pm e^{i\sqrt{A^\pm}x_2}\mathbf{a}^\pm = \int_{\sigma(A)} -\chi_{\mathbb{R}^\pm}(\lambda)\,\lambda e^{i\sqrt{\pm\lambda}x_2}\,dP_\lambda\,\mathbf{a}^\pm \\ &= \int_{\sigma(A)}-\lambda\,dP_\lambda\int_{\sigma(A)}\chi_{\mathbb{R}^\pm}(\lambda)e^{i\sqrt{\pm\lambda}x_2}\,dP_\lambda\,\mathbf{a}^\pm \\ &= -A\,e^{i\sqrt{A^\pm}x_2}\,\mathbf{a}^\pm. \end{aligned}$$

This implies that

$$U''=(e^{i\sqrt{A^+}x_2}\mathbf{a}^+)''+(e^{i\sqrt{A^-}x_2}\mathbf{a}^-)''=-Ae^{i\sqrt{A^+}x_2}\mathbf{a}^+ - Ae^{i\sqrt{A^-}x_2}\mathbf{a}^- = -AU,$$

which proves that the function $U(x_2)$ given by (3.2) is a solution of the infinite dimensional system (3.1). Since u should be downward radiating, we require u not to contain upgoing plane waves $e^{i\sqrt{A^+}\,x_2}\mathbf{a}^+$ and to be bounded for $x_2<b$, i.e., $\mathbf{a}^+\equiv 0$. Recalling $u|_{\Gamma_b}=f$, it follows from (3.2) that $\mathbf{a}^-=e^{i\sqrt{A^-}b}\mathbf{f}$, $\mathbf{f}:=(f_n)_{n\in\mathbb{Z}}$. This implies that

$$U(x_2)=e^{-i\sqrt{A^-}(x_2-b)}\mathbf{f}.$$

Definition 3.1. *If $q(x)=q(x_1)$ and $q\in L^\infty(0,2\pi)$ is real-valued, then $u\in H^1_\alpha(\Omega_b^-)$ is said to be a downward radiating solution to the Helmholtz equation if*

$$u(x_1,x_2)=\sum_{n\in\mathbb{Z}}\left[e^{-i\sqrt{A^-}(x_2-b)}\boldsymbol{g}\right]_n e^{i\alpha_n x_1},\quad x_2\le b,$$

for some $\boldsymbol{g}\in X^{1/2}$. Here the notation $[\cdot]_n$ stands for the nth entry of an infinite dimensional vector.

The above radiation condition allows us to express the solution to the boundary value problem (2.4) as

$$u(x_1,x_2)=\sum_{n\in\mathbb{Z}}\left[e^{-i\sqrt{A^-}(x_2-b)}\mathbf{f}\right]_n e^{i\alpha_n x_1},\quad x_2\le b.$$

Remark 3.2. *If $q \equiv q_0 = 1$, all the off-diagonal terms of A vanish and the diagonal terms take the form $a_{nn} = k^2 - \alpha_n^2$ for all $n \in \mathbb{Z}$. This implies that $(\sqrt{A^-}\,)_{nn} = \beta_n$, where $\beta_n \in \mathbb{C}$ is defined in (2.3). Hence, we have*

$$\left[e^{-i\sqrt{A^-}(x_2-b)}\boldsymbol{f}\right]_n = e^{-i\beta_n\,(x_2-b)}\,f_n,$$

that is, u takes the same form as (2.2). The new radiation condition in Def. 3.1 is a generalization of the classical radiation condition for periodic gratings with homogeneous cover and substrate material.

We remark that, the real-valued bounded index function q gives rise to a self-adjoint operator A and particularly excludes eigenvalues with generalized eigenfunctions in the spectrum of A. This has significantly simplified the arguments in comparison to the complex-valued potentials, which will be presented below.

4 Radiation condition for complex-valued potentials

Assume that $q(x) = q(x_1)$, where $q \in L^\infty(0, 2\pi)$ is complex-valued. We shall derive a different Rayleigh expansion into wave modes of the form $e^{\lambda x_2} h(x_1)$ instead of the $e^{i(\alpha_n x_1 - \beta_n(x_2-b))}$ in (2.2) or the $e^{i\alpha_n x_1} u_n(x_2)$ in (2.5). The functions h will be quasiperiodic eigenfunctions of a special ODE with respect to x_1, and the λ will be the corresponding eigenvalues. We shall consider the Helmholtz equation in Ω_b^- as a second-order ODE with respect to $x_2 \in (\infty, b)$, where the solution takes the function $\mathbb{R} \ni x_1 \mapsto u(x_1, x_2)$ as values at x_2. As usually, the second-order ODE is equivalent to a linear first-order 2-by-2 ODE system. The coefficient M, an ordinary differential operator with respect to x_1, is independent of x_2. Using the eigenvalues and generalized eigenfunctions of M, we can represent any solution as a Rayleigh series of wave modes, where, roughly speaking, each mode is the product of a generalized eigenfunction depending on x_1 times an exponential $e^{\lambda x_2}$ with λ the eigenvalue. In other words, in this section we write the Helmholtz equation as a linear second-order ODE with constant operator coefficient L. In Subsect. 4.1 we shall derive the equivalent first-order ODE with operator coefficient M. This 2-by-2 operator contains L in one of its entries. We shall analyze eigenvalues and eigenfunctions for L and M and special wave modes in Subsects. 4.2 and 4.3. Finally, we shall define the wave modes for the Rayleigh series and the radiation conditions in Subsect. 4.4.

4.1 Ordinary differential equation with respect to x_1

To get an equivalent first-order ODE, we set $\partial_j u = \partial u / \partial x_j$ $(j = 1, 2)$, $v := \partial_2 u$, and $W := (u, v)^\top$. Clearly, introducing the second-order ordinary differential operator

$$(Lf)(x_1) := -\frac{\mathrm{d}^2 f(x_1)}{\mathrm{d}x_1^2} - k^2 q(x_1) f(x_1), \tag{4.1}$$

the Helmholtz equation $(\Delta + k^2 qI)u = 0$ is equivalent to the function-valued second-order ODE $\partial_2^2 u(\cdot, x_2) - Lu(\cdot, x_2) = 0$, or equivalently, $\partial_2 v = Lu$. Hence, the Helmholtz equation can be written in the matrix-vector form

$$\partial_2 W = M\,W, \quad M := \begin{pmatrix} 0 & I \\ L & 0 \end{pmatrix}. \tag{4.2}$$

The domain of L is defined as

$$\mathcal{D} \;:=\; \Big\{ f \in L^2(0,2\pi) \colon f, f' \text{ are absolute continuous and } \alpha\text{-quasiperiodic}, \, Lf \in L^2(0,2\pi) \Big\}.$$

Note that L is self-adjoint over $\mathcal{D}$ if and only if the potential q is real-valued. It is well-known that the spectrum of L is purely discrete. In the Subsects. 4.2 and 4.3 we shall investigate the relation between the spectra of M and L. The eigenvalues and associated eigenfunctions of L and M are defined as follows.

Definition 4.1. *A number $\lambda \in \mathbb{C}$ is called an eigenvalue of the differential operator M combined with α-quasiperiodic boundary conditions, if the α-quasiperiodic boundary value problem $MW = \lambda W$ has at least one non-trivial solution $W = (w,v)^\top \in \mathcal{D}^2$. The function W is called eigenfunction corresponding to λ. Furthermore, we define associated eigenfunction of rank $m \geq 1$ by induction. A function $W \in \mathcal{D}^2$ is called associated eigenfunction of rank one of M corresponding to λ if it is an eigenfunction corresponding to λ. For $m > 1$, a function $W \in \mathcal{D}^2$ is called associated eigenfunction of rank m of M corresponding to λ if $W' := (M - \lambda \boldsymbol{I})\, W$ is a non-trivial associated eigenfunction of rank $m-1$ corresponding to λ. Here $\boldsymbol{I}$ denotes the 2-by-2 identity matrix. The functions $W^{(j)} := (M - \lambda \boldsymbol{I})^j W$ with $j \geq 0$ and $W^{(0)} := W$ will be referred to as the chain of associated eigenfunctions generated by W.*

Definition 4.2. *A number $\mu \in \mathbb{C}$ is called an eigenvalue of the differential operator L combined with α-quasiperiodic boundary conditions, if the α-quasiperiodic boundary value problem $Lh = \mu\, h$ has at least one non-trivial solution $h \in \mathcal{D}$. The function h is called eigenfunction corresponding to μ. Furthermore, we define associated eigenfunction of rank $m \geq 1$ by induction. A function $h \in \mathcal{D}$ is called associated eigenfunction of rank one of L corresponding to μ if it is an eigenfunction of L corresponding to μ. For $m > 1$, a function $h \in \mathcal{D}$ is called associated eigenfunction of rank m of L corresponding to μ if the function $h^{(1)} := (L - \mu I)\, h$ is a non-trivial associated eigenfunction of rank $m-1$ corresponding to μ. The functions $h^{(j)} := (L - \mu I)^j h$ with $j \geq 0$ and $h^{(0)} := h$ will be referred to as the chain of associated eigenfunctions generated by h.*

We conclude this subsection presenting an example of eigenvalues and eigenfunctions for L, where $k = 1$ and q is a piecewise constant function. For the proofs we refer to the techniques in [29]. We fix numbers $q_j \in \mathbb{C},\; j = 0,1$ and consider the squared refractive-index function

$$q(x_1) \;:=\; \begin{cases} q_0 & \text{if } 0 < x_1 < \pi \\ q_1 & \text{if } \pi < x_1 < 2\pi \end{cases}.$$

If μ is sufficiently large, then there are no associated eigenfunctions of rank greater one. For an eigenvalue μ, the eigenfunction h is given by

$$h(x_1) := \begin{cases} a\dfrac{\sin\left(\sqrt{q_0+\mu}\, x_1\right)}{\sqrt{q_0+\mu}} + \cos\left(\sqrt{q_0+\mu}\, x_1\right) & \text{if } 0 < x_1 < \pi \\ e^{\mathbf{i}\alpha 2\pi} \left\{ a \dfrac{\sin\!\left(\sqrt{q_1+\mu}\,(x_1 - 2\pi)\right)}{\sqrt{q_1+\mu}} + \cos\!\left(\sqrt{q_1+\mu}\,(x_1 - 2\pi)\right) \right\} & \text{if } \pi < x_1 < 2\pi \end{cases},$$

$$a \;:=\; e^{\mathbf{i}\alpha 2\pi} \cos\left(\sqrt{q_1+\mu}\,\pi\right) - \cos\left(\sqrt{q_0+\mu}\,\pi\right) \;=\; h'(0). \tag{4.3}$$

j	asymptotics of $\mu_{j,\pm}$	$\mu_{j,+}$	$\mu_{j,-}$
1	-0.43750	-0.51990	-0.36619
2	2.51562	2.4851	2.5457
3	7.50694	7.4901	7.5237
4	14.50391	14.493	14.515
5	23.50250	23.494	23.512
6	34.50174	34.501	34.502
7	47.50128	47.501	47.502
8	62.50098	62.501	62.501
9	79.50077	79.501	79.501
10	98.50062	98.501	98.501

Table 1: First ten eigenvalues for the case $\alpha=0$, $q_0=1$, and $q_1=2$.

Note that it does not matter which sign for the square root $\sqrt{q_0+\mu}$ and $\sqrt{q_1+\mu}$ is taken. Clearly, the formula (4.3) for h requires $\sqrt{q_j+\mu}\neq 0$. If $\sqrt{q_0+\mu}=0$ or $\sqrt{q_1+\mu}=0$, then we define $\sin(\sqrt{q_j+\mu}\,x_1)/\sqrt{q_j+\mu}=x_1$ and the formula remains true. The eigenvalues are those μ for which h and h' are α-quasiperiodic function. Thus they are the zeros of the function

$$\begin{aligned}\det(\mu) \;:=\; & -1-e^{\mathbf{i}\alpha 4\pi}+2e^{\mathbf{i}\alpha 2\pi}\cos(\sqrt{q_0+\mu}\,\pi+\sqrt{q_1+\mu}\,\pi)\\ & -e^{\mathbf{i}\alpha 2\pi}\frac{\sin(\sqrt{q_0+\mu}\,\pi)\sin(\sqrt{q_1+\mu}\,\pi)}{4(\sqrt{q_0+\mu}+\sqrt{q_1+\mu}\,)^2\sqrt{q_0+\mu}\sqrt{q_1+\mu}}\,.\end{aligned}$$

We obtain the asymptotics for the zeros $\mu_{j,\pm}$, $j\in\mathbb{Z}$ (cf. a special case in Tab. 1) given by

$$\mu_{j,\pm} \;:=\; (j\pm\alpha)^2-\frac{q_0+q_1}{2}+\mathcal{O}\big(|j|^{-\kappa}\big)\,,\;|j|\to\infty.$$

Here we have $\kappa:=1.5$ for $\alpha\neq 1/2$ and $\kappa:=0.5$ else. Moreover, $\mu_{j,+}\neq\mu_{j,-}$ for sufficiently large $|j|$.

4.2 Spectra of non-zero eigenvalues

Suppose $\mu\in\mathbb{C}$ is a non-zero eigenvalue of L. To state the relation between the spectra of L and M, we need to define the sequence γ_n, $n\in\mathbb{N}^+$ recursively by

$$\gamma_1 := \frac{1}{2\lambda},\quad \gamma_n := -\frac{\sum_{j=1}^{n-1}\gamma_j\gamma_{n-j}}{2\lambda},\quad n\geq 2, \tag{4.4}$$

where $\lambda=\lambda^{\pm}:=\pm\sqrt{\mu}$ is non-zero. Obviously,

$$\gamma_2=-\frac{1}{8\lambda^3},\;\gamma_3=\frac{1}{16\lambda^5},\;\gamma_4=-\frac{5}{128\lambda^7},\;\cdots$$

For the following lemma, recall that $h^{(j)}$ $(j=0,1,\cdots)$ is the chain generated by h (see Def. 4.2).

Lemma 4.3. *The pair (h,μ) with $\mu\neq 0$ is an eigenpair of rank $m\geq 1$ of the differential operator L, if and only if the eigenpair (W,λ) with $\lambda=\pm\sqrt{\mu}$, $W=(h,v)^T$ and*

$$v(x_1) := \lambda\, h(x_1) + \sum_{j=1}^{m-1} \gamma_j\, h^{(j)}(x_1).$$

is an eigenpair of rank $m\geq 1$ of M.

Proof. We first consider the case $m=1$. If (W,λ) with $W=(w,v)^\top$ is an eigenpair of rank one of M, then it is easy to conclude from $MW=\lambda W$ that $Lw=\lambda v$ and $v=\lambda w$ implying $(L-\lambda^2 I)w=0$. Hence, $(h,\mu)=(w,\lambda^2)$ is an eigenpair of rank one of L. Similarly, it is easy to prove that, if (h,λ^2) is an eigenpair of rank one of L, then (W,λ) with $W=(h,\lambda h)^\top$ is an eigenpair of rank one of M.

Now suppose $m=2$. If (W,λ) with $W=(w,v)^\top$, is an eigenpair of rank two of M, then $\widetilde{W}:=(M-\lambda\mathbf{I})W=:(\tilde{w},\tilde{v})^T\neq 0$ is an eigenfunction of rank one of M. This implies that $\tilde{v}=\lambda\tilde{w}$ and $(\tilde{w},\lambda^2)$ is an eigenpair of rank one of L. From the definition of $\widetilde{W}$, it is easy to obtain that

$$-\lambda w + v = \tilde{w}, \quad Lw - \lambda v = \tilde{v}, \tag{4.5}$$
$$M^2\, W = \lambda M W + M\widetilde{W} = \lambda(\lambda W + \widetilde{W}) + M\widetilde{W} = \lambda^2 W + (M+\lambda\mathbf{I})\widetilde{W}, \tag{4.6}$$

where

$$M^2 = \begin{pmatrix} L & 0 \\ 0 & L \end{pmatrix}.$$

Using $\tilde{v}=\lambda\tilde{w}$, we deduce from (4.6) that

$$Lw = \lambda^2 w + (\lambda\tilde{w} + \tilde{v}) = \lambda^2 w + 2\lambda\tilde{w},$$

leading to the relations

$$(L-\lambda^2 I)^2 w = (L-\lambda^2 I)(2\lambda\tilde{w}) = 0,$$
$$\tilde{w} = \gamma_1\,(L-\lambda^2 I)\, w \neq 0, \quad \gamma_1 := 1/(2\lambda).$$

Therefore, (w,λ^2) is an eigenpair of rank two of L. From the first relation in (4.5) we obtain

$$v = \lambda w + \tilde{w} = \lambda w + \gamma_1\, w^{(1)}, \quad w^{(j)} := (L-\lambda^2 I)^j\, w.$$

Now we treat the general case $m>2$ by induction. Suppose the induction hypothesis

The pair (W,λ) with $W=(w,v)^T$ is eigenpair of rank m of M
$\Longleftrightarrow$ (w,λ^2) is an eigenpair of rank m of L and $v=\lambda w+\sum_{j=1}^{m-1}\gamma_j\, w^{(j)}$. (4.7)

is fulfilled. We have to show that (4.7) holds with m replaced by $m+1$.

$\Rightarrow$: Suppose that (W,λ) with $W=(w,v)^T$ is an eigenpair of rank $m+1$ of M. Then $(\widetilde{W},\lambda)$ with $\widetilde{W}:=(M-\lambda\mathbf{I})W$ and $\widetilde{W}=(\tilde{w},\tilde{v})^T\neq 0$ is an eigenpair of rank m of M. By induction hypotheses this implies that $(\tilde{w},\lambda^2)$ is an eigenpair of rank m of L and

$$\tilde{v}=\lambda\tilde{w}+\sum_{j=1}^{m-1}\gamma_j\,\tilde{w}^{(j)}.$$

Combining the previous relation with (4.6) yields (cf. (4.7))

$$Lw=\lambda^2 w+(\lambda\tilde{w}+\tilde{v})=\lambda^2 w+2\lambda\tilde{w}+\sum_{j=1}^{m-1}\gamma_j\,\tilde{w}^{(j)},$$

from which we obtain

$$w^{(1)}:=(L-\lambda^2 I)w=2\lambda\tilde{w}+\sum_{j=1}^{m-1}\gamma_j\,\tilde{w}^{(j)}. \tag{4.8}$$

Since $(L-\lambda^2 I)^m\tilde{w}=0$, it follows that

$$(L-\lambda^2 I)^{m+1}w=(L-\lambda^2 I)^m w^{(1)}=2\lambda(L-\lambda^2 I)^m\tilde{w}+\sum_{j=1}^{m-1}\gamma_j\,\tilde{w}^{(m+j)}=0$$

and

$$(L-\lambda^2 I)^m w=(L-\lambda^2 I)^{m-1}w^{(1)}=2\lambda(L-\lambda^2 I)^{m-1}\tilde{w}\neq 0.$$

Hence, (w,λ^2) is an eigenpair of rank $m+1$ of L. To express v in terms of w, we deduce from (4.8) that

$$w^{(l)}:=(L-\lambda^2 I)^l w=2\lambda\tilde{w}^{(l-1)}+\sum_{j=1}^{m-l}\gamma_j\,\tilde{w}^{(l-1+j)},\quad l=1,2,\cdots,m,$$

which form the $m\times m$ linear system of equations $\widetilde{W}=\Pi_\lambda\,\widetilde{W}'$, where $\widetilde{W}:=(w^{(1)},\cdots,w^{(m)})^T$, $\widetilde{W}':=(\tilde{w},\tilde{w}^{(1)}\cdots,\tilde{w}^{(m-1)})^T$ and

$$\Pi_\lambda=\begin{pmatrix}2\lambda & \gamma_1 & \gamma_2 & \cdots & \gamma_{m-1}\\ 0 & 2\lambda & \gamma_1 & \cdots & \gamma_{m-2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \gamma_1\\ 0 & 0 & 0 & \cdots & 2\lambda\end{pmatrix}.$$

By the definition (4.4) of γ_n, the inverse of Π_λ is given by

$$\Pi_\lambda^{-1}=\begin{pmatrix}\gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_m\\ 0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{m-1}\\ 0 & 0 & \gamma_1 & \cdots & \gamma_{m-2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \gamma_1\end{pmatrix}.$$

This implies that the first component of $\widetilde{W}'$ is given by

$$\tilde{w} = \sum_{j=1}^{m} \gamma_j \, w^{(j)}.$$

Together with the first relation in (4.5) we obtain

$$v = \lambda w + \tilde{w} = \lambda w + \sum_{j=1}^{m} \gamma_j \, w^{(j)}.$$

$\Leftarrow$: Suppose that (w, λ^2) is an eigenpair of rank $m+1$ of L and $v = \lambda w + \sum_{j=1}^{m} \gamma_j \, w^{(j)}$. We have to prove that $W = (w, v)^T$ is an eigenfunction of rank $m+1$ of M. It suffices to show that $\widetilde{W} = (M - \lambda \mathbf{I}) W = (\tilde{w}, \tilde{v})^T$has the rank m. By the definition of M and the expression of v from our supposition,

$$\tilde{w} = -\lambda w + v = \sum_{j=1}^{m} \gamma_j \, w^{(j)}, \tag{4.9}$$

$$\tilde{v} = Lw - \lambda v = (1 - \lambda\gamma_1) w^{(1)} - \lambda \sum_{j=2}^{m} \gamma_j \, w^{(j)}. \tag{4.10}$$

Recalling the induction hypotheses, we only need to verify the relation

$$\tilde{v} = \lambda \tilde{w} + \sum_{j=1}^{m-1} \gamma_j \; \tilde{w}^{(j)}. \tag{4.11}$$

Using (4.9) and the definition (4.4) of γ_n, straightforward calculations show that

$$\begin{aligned}
\lambda \tilde{w} + \sum_{j=1}^{m-1} \gamma_j \; \tilde{w}^{(j)} &= \lambda \sum_{j=1}^{m} \gamma_j \, w^{(j)} + \sum_{j=1}^{m-1} \gamma_j \left(\sum_{l=1}^{m-j} \gamma_l \, w^{(j+l)} \right) \\
&= \lambda \sum_{j=1}^{m} \gamma_j \, w^{(j)} + \sum_{j=2}^{m} w^{(j)} \left(\sum_{l=1}^{j-1} \gamma_l \gamma_{j-l} \right) \\
&= \lambda \sum_{j=1}^{m} \gamma_j \, w^{(j)} + \sum_{j=2}^{m} w^{(j)} (-2\lambda\gamma_j) \\
&= \lambda \gamma_1 w^{(1)} - \lambda \sum_{j=2}^{m} \gamma_j \, w^{(j)}.
\end{aligned}$$

Since $2\lambda\gamma_1 = 1$ and (4.10), the previous identity confirms the relation (4.11). The proof is completed. □

The chain $W^{(j)}$ generated by W is given in Def. 4.1, the chain $h^{(j)}$ generated by h in Def. 4.2. As a consequence of the proof to Lemma 4.3, we obtain

Lemma 4.4. *(i) Suppose (h, λ^2) is an eigenpair of rank $m \geq 1$ of L. Then the vector functions*

$$\begin{pmatrix} h^{(l)} \\ \lambda h^{(l)} + \sum_{j=1}^{m-1-l} \gamma_j \, h^{(j+l)} \end{pmatrix}, \qquad l = 0, 1, 2, \cdots, m-1,$$

are the associated eigenfunctions of rank $m-l$ of operator M corresponding to the eigenvalue λ.

(ii) Suppose (W, λ) is an eigenpair of rank $m \geq 1$ of operator M. Write $W^{(l)} = (W_1^{(l)}, W_2^{(l)})^T$ for $l = 0, 1, \cdots, m-1$. Then $(W_1^{(l)}, \lambda^2)$ is an eigenpair of rank $m-l$ of L and

$$W_2^{(l)} = \lambda W_1^{(l)} + \sum_{j=1}^{m-l-1} \gamma_l \, \left(L - \lambda^2\right)^j \, W_1^{(l)}.$$

Proof. Lemma 4.4 follows from Lemma 4.3 and the fact that $(W^{(l)}, \lambda)$, $(h^{(l)}, \lambda^2)$ are eigenpairs of rank $m-l$ corresponding to M and L, respectively. Note that, in the case of $l = 0$, the assertions of Lemma 4.4 coincide with those in Lemma 4.3. □

By Lemma 4.3, in order to get the spectrum of M, it suffices to investigate the spectrum of the quasiperiodic differential operator L. We collect properties of the nonself-adjoint operator L in the subsequent two lemmas.

Lemma 4.5. *(i) The spectrum $\sigma_p(L)$ of L is a discrete set of eigenvalues and the only accumulation point is infinity.*

(ii) The geometric multiplicity of each eigenvalue $\mu \in \sigma_p(L)$ is finite, i.e., $\dim(\ker(L - \mu I)) < \infty$.

(iii) The algebraic multiplicity of each eigenvalue $\mu \in \sigma_p(L)$ is finite, i.e., $\dim(A_L(\mu)) < \infty$, where

$$\begin{aligned} A_L(\mu) \quad := \quad \Big\{ & h \in \mathcal{D} : \text{there is an } m \in \mathbb{N} \text{ s.t.} \\ & L^j h \in \mathcal{D}, \ j = 1, \ldots, m-1 \text{ and } (L - \mu I)^m \, h = 0 \Big\}. \end{aligned} \tag{4.12}$$

(iv) The eigenvalues can be denoted as $\mu_n = \mu_n(\alpha) \in \sigma_p(L)$ for index n running in $\mathbb{Z}$.

If $\alpha \neq 0, 1/2$, then the algebraic multiplicity of the μ_n is equal to one for sufficiently large $|n|$. Choosing a suitable scaling factor for the rank-one eigenfunction h_n corresponding to μ_n, we get $h_n(0) = 1$ and the asymptotics

$$\mu_n(\alpha) \quad = \quad (n + \alpha)^2 - \frac{k^2}{2\pi} \int_0^{2\pi} q(t) \, dt + \mathcal{O}\left(\frac{1}{|n|}\right), \tag{4.13}$$

$$h_n(x_1) \quad = \quad \exp\Big(i(n + \alpha)x_1\Big) + \mathcal{O}\left(\frac{1}{|n|}\right), \quad n \in \mathbb{Z}, \tag{4.14}$$

as $|n| \to \infty$, where the term $\mathcal{O}(1/|n|)$ is uniform with respect to $x_1 \in [0, 2\pi]$.

If $\alpha = 0, 1/2$, then the algebraic multiplicity of the μ_n is one or two for sufficiently large $|n|$. The eigenvalue asymptotics (4.13) holds with $\mathcal{O}(1/|n|)$ replaced by $\mathcal{O}(1/|n|^{1/2})$. Instead of (4.14), the eigenfunctions of rank one admit the asymptotic expansion

$$h_n(x_1) \;=\; C_+(n)\exp\left[i(n+\alpha)x_1\right] + C_-(n)\exp\left[-i(n+\alpha)x_1\right] + \mathcal{O}\left(\frac{1}{|n|}\right), \quad (4.15)$$

where $C_\pm(n) \in \mathbb{C}$ and $n \in \mathbb{Z}$ with $|n| \to \infty$. For normalization, in (4.15) we may suppose $h_n(0) \in \mathbb{R}$ and $|C_+(n)|^2 + |C_-(n)|^2 = 1$. Furthermore, for sufficiently large $|n|$ and for eigenvalues $\mu_n(\alpha) = \mu_{-n-2\alpha}(\alpha)$ with two linearly independent eigenfunctions of rank one, a pair of eigenfunctions h_n and $h_{-n-2\alpha}$ can be found satisfying (4.14) with n set to n and $-n-2\alpha$, respectively.

The assertions (i)-(iii) follow from the spectral theory of nonself-adjoint differential equations (see e.g., [9, 15, 16] and references therein). The asymptotic behavior of the spectrum of L was studied, e.g., in [38] for $\alpha \neq 0, 1/2$, in [15] for $\alpha = 0, 1/2$ and in [29] for the general case. The results in the last assertion were used in the proof of [36, Thm. 4.12] to derive uniqueness for the identification of a periodic medium, which depends only on x_2, from near-field measurement data of infinitely many incoming waves.

Obviously, one has

$$\dim\left(\ker\,(L-\mu)\right) \le \dim\left(A_L(\mu)\right)$$

for each $\mu \in \sigma_p(L)$, $\mu \neq 0$. The set of all eigenfunctions and associated eigenfunctions of $\mu \in \sigma_p(L)$ form the eigenspace corresponding μ, which is a closed linear subspace of $L^2(0, 2\pi)$ with dimension equal to the algebraic multiplicity of μ. For $q = 0$ and $n \in \mathbb{Z}$, we have $\mu_n = (n+\alpha)^2$ and all associated eigenfunctions $h_n(x_1) = \exp\left(i(n+\alpha)x_1\right)$ are of rank one. For $q \neq 0$, the eigenvalues as well as the eigenfunctions and associated eigenfunctions are obtained by perturbation arguments. Therefore, we have the same general indices $n \in \mathbb{Z}$ for the set of all eigenfunctions and associated eigenfunctions. So this covers the case of associated eigenfunctions of rank greater than one. Indeed, in this case the values μ_n might coincide for several $n \in \mathbb{Z}$ and the corresponding h_n span the space of all eigenfunctions and associated eigenfunctions.

Since the α-quasiperiodic boundary conditions are non-degenerate, we infer from [29, Thm. 1.3.1], [15, Thm. 2.1] and [38, Thm. 3] that

Lemma 4.6. *The system of eigenfunctions and associated eigenfunctions h_n, $n \in \mathbb{Z}$ of the α-quasiperiodic operator L is complete over $L^2(0, 2\pi)$. Further, they form a Riesz basis of $L^2(0, 2\pi)$ if $\alpha \neq 0, 1/2$.*

Let us comment on the choice of eigenfunctions for a basis. Note that, for $\alpha \neq 0, 1/2$, each eigenvalue μ_n with sufficiently large $|n|$ has an eigenfunction of rank one, which is unique by the normalization $h_n(0) = 1$. A basis transform for the general eigenfunctions with n in a finite set does not change the Riesz property. For $\alpha = 0, 1/2$, the eigenvalues of multiplicity two have a non-unique basis. If the two eigenfunctions are both of rank one, then the basis can be fixed by $h_n(0) = 1$ and (4.14) without changing the Riesz property. However, if there is a generalized eigenfunction of rank two, then the Riesz property might depend on a good choice of generalized eigenfunctions for the basis. In particular, it might be necessary to choose two eigenfunctions of rank two for some of the eigenvalues in order to form a Riesz basis. Choosing a chain of generalized eigenfunctions might lead to a system without Riesz property. We suppose that the system of generalized eigenfunctions h_n is chosen such that the Riesz property is fulfilled whenever this is possible. Moreover, we assume a special choice of rank-two eigenfunctions. By I_d we denote the set of indices n such that $\mu_n=\mu_{-n-2\alpha}$ has exactly two rank-two eigenfunctions h_n and $h_{-n-2\alpha}$ in the Riesz system. Then, for $n\in I_d$,

$$\begin{aligned} \big([-\partial^2-k^2qI]-\mu_n\big)h_n &= c_{n,1,1}h_n+c_{n,1,2}h_{-n-2\alpha}, \\ \big([-\partial^2-k^2qI]-\mu_n\big)h_{-n-2\alpha} &= c_{n,2,1}h_n+c_{n,2,2}h_{-n-2\alpha}. \end{aligned} \tag{4.16}$$

For a linear combination $f_nh_n+f_{-n-2\alpha}h_{-n-2\alpha}$ with $f_n,f_{-n-2\alpha}\in\mathbb{C}$, we get

$$\begin{aligned} \big\|\partial^2\big(f_nh_n+f_{-n-2\alpha}h_{-n-2\alpha}\big)\big\|^2 &\sim \left\langle B_n^*B_n(f_n,f_{-n-2\alpha})^\top,(f_n,f_{-n-2\alpha})^\top\right\rangle, \\ B_n &:= \begin{pmatrix} \mu_n+c_{n,1,1} & c_{n,2,1} \\ c_{n,1,2} & \mu_n+c_{n,2,2} \end{pmatrix}. \end{aligned}$$

By the eigenvalue decomposition of self-adjoint matrices there exists a unitary matrix U_n and non-negative eigenvalues $\kappa_n,\kappa_{-n-2\alpha}$ such that $B_n^*B_n=U_n^*\mathrm{diag}(\kappa_n,\kappa_{-n-2\alpha})U_n$. In other words, applying a basis transform for the basis functions h_n and $h_{-n-2\alpha}$, we may suppose $U_n=I$ and arrive at

$$\big\|\partial^2\big(f_nh_n+f_{-n-2\alpha}h_{-n-2\alpha}\big)\big\|^2 \sim \kappa_n|f_n|^2+\kappa_{-n-2\alpha}|f_{-n-2\alpha}|^2. \tag{4.17}$$

This normalization of pairs of basis functions for $\alpha=0,1/2$ will always be supposed in the following. If $\alpha\neq 0,1/2$, then we set $I_d=\emptyset$, since, for large $|n|$, all eigenvalues μ_n have algebraic multiplicity one.

The adjoint operator of L over the quasiperiodic functions is the operator L^* over quasiperiodic functions, which is defined as L in (4.1) but with q replaced by the complex conjugate function $\overline{q}$. Since the eigenfunctions and the associated eigenfunctions of L^* corresponding to $\overline{\mu}_n$ are L^2 orthogonal to the eigenfunctions and associated eigenfunctions of L corresponding to μ_m for $\mu_m\neq\mu_n$ (cf. the proof of [38, Thm. 3]), we conclude that there exists a dual system $h_n^*,\ n\in\mathbb{Z}$ such that $\langle h_m^*,h_n\rangle=\delta_{m,n}$ and $\langle h_m^*,h_n\rangle=\delta_{m,n}$. The existence of a complete dual system implies that the system $h_n,\ n\in\mathbb{Z}$ is total and minimal. Of course, the scaling for the dual system is different than that in Lemma 4.5, (iv). In particular, if the algebraic multiplicity of an eigenvalue is greater than one, then the scaling is difficult to estimate and the Riesz property might get lost.

If $\alpha=0,1/2$, then the α-quasiperiodic boundary conditions reduce to the periodic boundary conditions and the antiperiodic boundary conditions $h(0)=-h(2\pi),\ h'(0)=-h'(2\pi)$, respectively. Unfortunately, the modified asymptotics (4.13) does not exclude the identity $\mu_n(\alpha)=\mu_{-n+2\alpha}(\alpha)$ for large $|n|$, which might lead to troubles in estimating the norms of the dual basis. We refer to [15, Thm. 1.2, Cor. 1.5] for necessary and sufficient conditions, under which the eigenfunctions form a Riesz or Schauder basis over $L^2(0,2\pi)$ in the case of $\alpha=0,1/2$.

For general α but real-valued q, the operator L over quasiperiodic functions is self-adjoint and the system $h_n,\ n\in\mathbb{Z}$ forms an orthogonal basis in the Hilbert space L^2. In this paper we suppose that either $\alpha \neq 0, 1/2$, or q is real-valued, or the conditions in [15, Thm. 1.2, Cor. 1.5] hold for $\alpha = 0, 1/2$, so that the $h_n,\ n \in \mathbb{Z}$ always form a Riesz basis. Note that, for the main result in Thm. 5.7, the Riesz basis assumption can be replaced by assuming a subexponential bound for the norms of the dual basis. However, this leads to more involved definitions and proofs, since the convergence of an expansion with respect to a Riesz basis is to be replaced by density arguments for finite linear combinations of the $h_n,\ n \in \mathbb{Z}$. With the Riesz basis assumption, for each α we obtain the following equivalence of the Sobolev norms with weighted ℓ^2 norms of the coefficients with respect to the Riesz basis $h_n,\ n \in \mathbb{Z}$.

Lemma 4.7. *Suppose $h_n,\ n \in \mathbb{Z}$ is a Riesz basis in $L^2(0, 2\pi)$. For each s fixed with $-2 \leq s \leq 2$, there exists a constant $c_s > 0$ such that, for all sequences $f_n \in \mathbb{C}$ and for the κ_n from (4.17),*

$$\frac{1}{c_s}\Big\| \sum_{n\in\mathbb{Z}} f_n h_n \Big\|^2_{H^s_\alpha(0,2\pi)} \leq \sum_{n\in\mathbb{Z}\setminus I_d} (1+|n|)^{2s}|f_n|^2 + \sum_{n\in I_d} (1+\kappa_n)^s |f_n|^2 \leq c_s \Big\| \sum_{n\in\mathbb{Z}} f_n h_n \Big\|^2_{H^s_\alpha(0,2\pi)}.$$

Note, by a coarse estimate, we have $\kappa_n \leq \mathcal{O}(|n|^4)$.

Proof. For $s = 0$ the norm equivalence is a well-known fact for any kind of Riesz basis. If $s = 2$ and all eigenfunctions with eigenvalue $\mu_n \geq n_0$ are of rank one, then

$$\begin{aligned} \Big\| \sum_{n\in\mathbb{Z}} f_n h_n \Big\|^2_{H^2_\alpha(0,2\pi)} &\sim \Big\| \sum_{n\in\mathbb{Z}} f_n\, h_n^{''} \Big\|^2_{L^2_\alpha(0,2\pi)} + \Big\| \sum_{n\in\mathbb{Z}} f_n h_n \Big\|^2_{L^2_\alpha(0,2\pi)} \\ &\sim \Big\| \sum_{n\in\mathbb{Z}:|n|\geq n_0} f_n(\mu_n + k^2 q) h_n \Big\|^2_{L^2_\alpha(0,2\pi)} + \sum_{n\in\mathbb{Z}} |f_n|^2. \end{aligned} \tag{4.18}$$

Using $q \in L^\infty(0, 2\pi)$ and the fact that $\mu_n \sim |n|^2$ for $n \to \pm\infty$ (see Lemma 4.5, (iv)) we continue

$$\begin{aligned} \Big\| \sum_{n\in\mathbb{Z}} f_n h_n \Big\|^2_{H^2_\alpha(0,2\pi)} &\sim \Big\| \sum_{n\in\mathbb{Z}:|n|\geq n_0} (\mu_n f_n) h_n \Big\|^2_{L^2_\alpha(0,2\pi)} + \sum_{n\in\mathbb{Z}} |f_n|^2 \\ &\sim \sum_{n\in\mathbb{Z}} |\mu_n|^2 |f_n|^2 + \sum_{n\in\mathbb{Z}} |f_n|^2 \\ &\sim \sum_{n\in\mathbb{Z}} (1 + |n|)^4 |f_n|^2. \end{aligned}$$

Hence, the assertion holds for $s = 2$, and the norm of the dual space $H^{-2}_\alpha(0, 2\pi)$ is equivalent to dual of the weighted ℓ^2 space, i.e., the assertion is true for $s = -2$. By interpolating the spaces, we obtain the assertion for any s with $-2 \leq s \leq 2$.

The proof in the general case follows analogously, if we apply $h_n'' = (\mu_n + k^2 q) h_n + g_n$ instead of $h_n'' = (\mu_n + k^2 q) h_n$ to (4.18) and if we use (4.17). It remains to show the estimate of the κ_n. If $n \in I_d$, then we get (4.16). We denote the rank-one eigenfunction on the right-hand side of (4.16) by g_n. Fixing a suitable $c_0 > 0$, the operator $[(-\partial^2 - k^2 q I) + c_0 I]$ is invertible and its inverse is the compact resolvent operator $B := [(-\partial^2 - k^2 qI) + c_0 I]^{-1}$. Hence, the property $(-\partial^2 - k^2 qI)g_n = \mu_n g_n$ of the rank-one eigenfunction g_n leads us to

$$\begin{aligned} [(-\partial^2 - k^2 qI) + c_0 I]h_n - (\mu_n + c_0)h_n &= g_n, \\ (\mu_n + c_0)^{-1} h_n - B h_n &= (\mu_n + c_0)^{-2} g_n, \\ g_n &= (\mu_n + c_0)h_n - (\mu_n + c_0)^2 B h_n. \end{aligned}$$

Here $\|(\mu_n + c_0)h_n\| = \mathcal{O}(|n|^2)$, and B is a bounded operator in L^2. Thus $\|g_n\| = \mathcal{O}(|n|^4)$ such that $c_{n,1,j} = \mathcal{O}(|n|^4),\ j=1,2$. Similarly, $c_{n,2,j} = \mathcal{O}(|n|^4),\ j=1,2$, and the non-negative singular value κ_n is at most $\mathcal{O}(|n|^4)$. □

By Lemma 4.6, the set of eigenfunctions and associated eigenfunctions of L is complete over $L^2(0,2\pi)$ for any $\alpha \in [0,1)$. To consider eigenfunctions of higher ranks, we denote by $(h_{n,m}, \mu_n)$ with $h_{n,m} \in A_L(\mu_n)$ an eigenpair of rank $m \geq 1$ of L. However, we should always keep in mind that the system $(h_{n,m}, \mu_n)$ coincides with the previously used notation (h_n, μ_n). By Lemma 4.3 we may construct eigenpairs $(W^{\pm}_{n,m}, \lambda^{\pm}_n)$ of rank $m \geq 1$ of M as follows:

$$\lambda^{\pm}_n = \pm\sqrt{\mu_n}, \quad W^{\pm}_{n,m}(x_1) = \begin{pmatrix} h_{n,m}(x_1) \\ \lambda^{\pm}_n\, h_{n,m}(x_1) + \sum_{j=1}^{m-1} \gamma^{\pm}_{j,n}\, h^{(j)}_{n,m}(x_1) \end{pmatrix} \in A_M(\lambda^{\pm}_n), \tag{4.19}$$

where the $\gamma^{\pm}_{j,n}$ are defined the same way as γ_j with λ replaced by $\lambda^{\pm}_n$ (see (4.4)). Here, the functions $h^{(j)}_{n,m} = (L - \mu_n I)^j h_{n,m}$ represent the chain generated by $h_{n,m}$ and the set $A_M(\lambda)$ denotes the eigenspace of the operator M corresponding to the eigenvalue λ, that is (cf. (4.12)),

$$\begin{aligned} A_M(\mu) \;:=\; \Big\{ g \in \mathcal{D}^2 &: \text{there is an } m \in \mathbb{N} \text{ s.t.} \\ & M^j g \in \mathcal{D}^2,\ j = 1, \ldots, m-1 \text{ and } (M - \lambda\mathbf{I})^m g = 0 \Big\}. \end{aligned}$$

As will be seen later, we shall switch between the indices $+$ and $-$ to define upward and downward radiating wave modes for $x_2 \geq b$ and $x_2 \leq b$, respectively.

Lemma 4.8. *Suppose (g, λ) with $g = (g_1, g_2)^\top \in A_M(\lambda)$ is an eigenpair of rank $m \geq 1$ of M. Then the unique solution $W(x_1, x_2) = (u(x_1, x_2), v(x_1, x_2))^\top$ to the quasiperiodic initial boundary value problem*

$$\partial_2 W = M\, W, \quad W(\cdot, b) = g, \tag{4.20}$$

is given by

$$W(x_1, x_2) = e^{\lambda(x_2 - b)} \sum_{n=0}^{m-1} \frac{g^{(n)}(x_1)\ (x_2 - b)^n}{n\,!},$$

where $\{g^{(n)} : n = 1, \cdots, m\}$ denotes the chain generated by g as defined for generator h in Def. 4.2.

Proof. Without loss of generality we suppose that $b=0$. Obviously, $W(x_1,x_2):=\exp(Mx_2)g(x_1)$ is the unique solution to (4.20). For $m=1$, we have $(M-\lambda\boldsymbol{I})g=0$, implying that $M^j g=\lambda^j g$ for any $j\in\mathbb{N}$. Hence, by the definition of the exponential function of a matrix we obtain

$$W(x_1,x_2)=\exp(Mx_2)g(x_1)=\sum_{j=0}^{\infty}\frac{x_2^j}{j!}M^j g=\sum_{j=0}^{\infty}\frac{x_2^j\lambda^j}{j!}g=e^{\lambda x_2}\,g.$$

Next we will verify the lemma in the general case of $m\geq 1$. From the definition of $g^{(n)}$, using an induction argument we see

$$M^j\,g=\sum_{n=0}^{\min\{j,m-1\}}\lambda^{j-n}\,g^{(n)}\binom{j}{n},\quad \binom{j}{n}:=\frac{j\,!}{(j-n)\,!\,n\,!}\;. \tag{4.21}$$

Note that in deriving (4.21), we have used the relation $Mg^{(n)}=\lambda g^{(n)}+g^{(n+1)}$. We split the function $e^{Mx_2}g$ into the sum of

$$\exp(Mx_2)g(x_1)=\sum_{j=0}^{m-1}\frac{x_2^j}{j!}M^j g+\sum_{j=m}^{\infty}\frac{x_2^j}{j!}M^j g\;. \tag{4.22}$$

The first sum on the right-hand side of the previous identity can be rewritten using (4.21) as

$$\begin{aligned}\sum_{j=0}^{m-1}\frac{x_2^j}{j!}M^j g &= \sum_{j=0}^{m-1}\frac{x_2^j}{j!}\sum_{n=0}^{j}\lambda^{j-n}\,g^{(n)}\binom{j}{n} \quad = \sum_{j=0}^{m-1}x_2^j\sum_{n=0}^{j}\frac{\lambda^{j-n}}{(j-n)\,!\,n\,!}\,g^{(n)}\\ &= \sum_{n=0}^{m-1}\frac{x_2^n}{n\,!}\,g^{(n)}\sum_{j=n}^{m-1}\frac{x_2^{j-n}\,\lambda^{j-n}}{(j-n)\,!},\end{aligned}$$

where the summation over the indices j and m has been interchanged in the last step. Analogously,

$$\sum_{j=m}^{\infty}\frac{x_2^j}{j!}M^j g = \sum_{n=1}^{m-1}\frac{x_2^n}{n\,!}\,g^{(n)}\sum_{j=m}^{\infty}\frac{x_2^{j-n}\,\lambda^{j-n}}{(j-n)\,!}.$$

The previous two identities together with (4.22) imply

$$\exp(Mx_2)g=\sum_{n=0}^{m-1}\frac{x_2^n}{n\,!}g^{(n)}\left(\sum_{j=0}^{\infty}\frac{x_2^j\,\lambda^j}{j\,!}\right)=e^{\lambda x_2}\sum_{n=0}^{m-1}\frac{x_2^n}{n\,!}\,g^{(n)}.$$

□

Theorem 4.9. *Suppose $(h_{n,m},\mu_n)$ with $h_{n,m}\in A_L(\mu_n)$ is an eigenpair of rank $m\geq 1$ of L and define $\lambda_n^{\pm}$ and $W_{n,m}^{\pm}$ as in (4.19). Consider the boundary value problem for α-quasiperiodic solutions u:*

$$\Delta u+k^2qu=0 \quad \textit{in} \quad \mathbb{R}^2, \qquad u=h_{n,m} \quad \textit{on} \quad \Gamma_b, \tag{4.23}$$

DOI 10.20347/WIAS.PREPRINT.2726

(i) The solution $u = u_{n,m} \in H^2_{loc}(\mathbb{R}^2)$ *can be represented by* $u_{n,m} = C^+ u^+_{n,m} + C^- u^-_{n,m}$, *where* $C^\pm \in \mathbb{C}$ *and*

$$u^\pm_{n,m}(x_1, x_2) = e^{\lambda^\pm_n (x_2 - b)} \sum_{j=0}^{m-1} (W^\pm_{n,m})^{(j)}_1(x_1) \, \frac{(x_2 - b)^j}{j\,!}. \tag{4.24}$$

Here $(W^\pm_{n,m})^{(j)}_1$ *denotes the first component of the chain* $(W^\pm_{n,m})^{(j)}$ *generated by* $W^\pm_{n,m}$. *Furthermore, for* $0 \le j \le m-1$, *the associated eigenfunction* $(W^\pm_{n,m})^{(j)}$ *of the operator* M *with the corresponding eigenvalue* $\lambda^\pm_n$ *is of rank* $m-j$ *and can be represented as*

$$(W^\pm_{n,m})^{(j)}_1 = \sum_{l=0}^{m-1} A^{(j)}_l \, h^{(l)}_{n,m}, \quad (W^\pm_{n,m})^{(j)}_2 = \sum_{l=0}^{m-1} B^{(j)}_l \, h^{(l)}_{n,m}, \quad 0 \le j \le m-1, \tag{4.25}$$

with the coefficients $A^{(j)}_l = A^{\pm,(j)}_l$, $0 \le l \le m-1$ *and* $B^{(j)}_l = B^{\pm,(j)}_l$, $0 \le l \le m-1$ *given by the recursion*

$$A^{(0)}_0 := 1,\ B^{(0)}_0 := \lambda^\pm_n,\ A^{(0)}_l := 0,\ B^{(0)}_l := \gamma^\pm_{l,n},\ 0 < l \le m-1, \tag{4.26}$$

$$A^{(j+1)}_l = -\lambda^\pm_n A^{(j)}_l + B^{(j)}_l, B^{(j+1)}_l = A^{(j)}_{l-1} + \mu_n A^{(j)}_l - \lambda^\pm_n B^{(j)}_l,\ 0 \le l \le m-1, \tag{4.27}$$

where $\mu_n = [\lambda^\pm_n]^2$ *and* $A^{(j)}_{-1} := 0$.

(ii) It holds that

$$\partial_2 u^\pm_{n,m}(x_1, b) = \lambda^\pm_n h_{n,m}(x_1) + \sum_{j=1}^{m-1} \gamma^\pm_{j,n} \, h^{(j)}_{n,m}(x_1).$$

Proof. Suppose $\lambda^\pm_n$ and $W^\pm_{n,m}$ are defined by (4.19). By Lemma 4.3, the eigenpairs $(W^\pm_{n,m}, \lambda^\pm_n)$ of M are of rank m. Hence, the $u^\pm_{n,m}$ are solutions of the α-quasiperiodic boundary value problem (4.23) if and only if $W^\pm = (u^\pm_{n,m}, \partial_2 u^\pm_{n,m})$ satisfy the α-quasiperiodic ODE systems

$$\partial_2 W^\pm = M W^\pm \quad \text{in} \quad \mathbb{R}^2, \qquad\qquad W^\pm = W^\pm_{n,m} \quad \text{on} \quad \Gamma_b,$$

By Lemma 4.8, we get the solutions

$$W^\pm(x_1, x_2) = e^{\lambda^\pm_n (x_2 - b)} \sum_{j=0}^{m-1} (W^\pm_{n,m})^{(j)}(x_1) \frac{(x_2 - b)^j}{j\,!}. \tag{4.28}$$

Recall from (4.19) that

$$\begin{aligned} (W^\pm_{n,m})^{(0)}_1 &= (W^\pm_{n,m})_1 = h_{n,m}, \\ (W^\pm_{n,m})^{(0)}_2 &= \lambda^\pm_n \, h_{n,m} + \sum_{j=1}^{m-1} \gamma^\pm_{j,n} \, h^{(j)}_{n,m}. \end{aligned} \tag{4.29}$$

DOI 10.20347/WIAS.PREPRINT.2726

The expression of $u^{\pm}_{n,m}$ follows from the first component of (4.28), and consequently, $\partial_2 u^{\pm}_{n,m}|_{\Gamma_b}$ coincides with the second component of $W^{\pm}_{n,m}|_{\Gamma_b}$. Finally, the initial condition (4.26) follows from (4.19) and the recursion (4.27) for the coefficients in (4.25) from

$$(M-\lambda_n^{\pm}\mathbf{I}) \;=\; \begin{pmatrix} -\lambda_n^{\pm} I & I \\ (L-\mu_n I)+\mu_n I & -\lambda_n^{\pm} I \end{pmatrix}.$$

□

As a consequence of Thm. 4.9, we present the solutions for eigenvalues of rank two.

Corollary 4.10. *Suppose (h, λ^2) with $h \in A_L(\lambda^2)$ is an eigenpair of L of rank two. Then the solution $u \in H^2_{loc}(\mathbb{R}^2)$ of the boundary value problem (4.23) can be represented by $u = C^+u^+ + C^-u^-$, where $C^{\pm} \in \mathbb{C}$ and*

$$u^{\pm}(x_1, x_2) = e^{\pm\lambda(x_2-b)} \left[h(x_1) \pm \frac{1}{2\lambda}(x_2 - b)\, h^{(1)}(x_1) \right], \quad x \in \mathbb{R}^2,$$

where $h^{(1)} = (L - \lambda^2 I)h \neq 0$. In particular, we have

$$\partial_2 u^{\pm}(x_1, b) = \pm\lambda h(x_1) \pm \frac{1}{2\lambda} h^{(1)}(x_1) \qquad \text{for} \quad x_2 = b.$$

Proof. The assertion follows from Theorem 4.9 with the following replacement

$$m = 2,\ \lambda_n^{\pm} = \pm\lambda,\ \gamma_{1,n}^{\pm} = \frac{1}{2\lambda_n^{\pm}} = \pm\frac{1}{2\lambda}, \quad u^{\pm}_{n,2} = u^{\pm},\ h_{n,2} = h.$$

□

4.3 The eigenvalue zero

In this subsection we suppose that $\mu = 0$ is an eigenvalue of L with the eigenfunction h. If $(h, 0)^{\top}$ is an eigenpair of rank one, by Thm. 4.9 the solution u to the quasiperiodic boundary value problem (4.23) takes the form

$$u(x) = h(x_1), \quad x \in \mathbb{R}^2, \tag{4.30}$$

implying that $\partial_2 u(x_1, x_2) = 0$ for any (x_1, x_2). For higher ranks $m \geq 2$, however, Thm. 4.9 is not meaningful because the coefficients γ_j, $j \geq 1$ (cf. (4.4)) are not well defined for eigenvalue zero.

Lemma 4.11. *Suppose $\lambda = 0$ is an eigenvalue for M of rank $2m-1$ or $2m$ with $m \geq 1$. Then the corresponding eigenspace of rank $2m-1$ consists of vector functions of the form $(u_m, v_{m-1})^T$, while the eigenspace of rank $2m$ consists of functions of the form $(u_m, v_m)^T$. Here the u_m, v_m and v_{m-1} ($v_0 \equiv 0$) are eigenfunctions of L with respect to the eigenvalue $\mu = 0$ of rank m and $m-1$, respectively.*

DOI 10.20347/WIAS.PREPRINT.2726

Proof. Denote by $W=(u,v)^T$ the eigenfunction of M that corresponds to the eigenvalue $\lambda=0$. It is easy to see

$$MW = \begin{pmatrix} 0 & 1 \\ L & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ Lu \end{pmatrix}. \tag{4.31}$$

Hence, $(W,0)$ is an eigenpair of rank one if and only if $v=0$ and $Lu=0$, that is $W=(u,0)$, where the eigenvector u of L corresponding to the eigenvalue zero is of rank one. Analogously, $(W,0)$ is an eigenpair of rank two if and only $(v,Lu)^T$ is an eigenfunction of rank one, which implies that v is of rank one and $Lu=0$, that is both v and u are of rank one. This proves Lemma 4.11 in the cases $m=1$ and $m=2$. The general case $m\geq 3$ can be proved easily via induction and using (4.31). □

Theorem 4.12. *Suppose $(h_{0,m},0)$, $h_{0,m}\in A_L(0)$ is an eigenpair of operator L with rank $m\geq 1$. Then the solution $u\in H^2_{loc}(\mathbb{R}^2)$ to the quasiperiodic boundary value problem (4.23) takes the form $u=C^+u_m^+ + C^-u_m^-$, where $C^\pm\in\mathbb{C}$ and*

$$u_m^\pm(x_1,x_2) = \sum_{j=0}^{2m-1} w_m^{j,\pm}(x_1)\,(x_2-b)^j/j\,!, \qquad x_2\in\mathbb{R},$$

where, for $n=0,1,\cdots,m-1$,

$$w_m^{2n,\pm}(x_1) = h_{0,m}^{(n)}(x_1), \quad w_m^{2n+1,+}(x_1) = v_m^{(n)}(x_1), \quad w_m^{2n+1,-}(x_1) = v_{m-1}^{(n)}(x_1). \tag{4.32}$$

Here v_m, v_{m-1} ($v_0\equiv 0$) are arbitrary eigenfunctions of L of rank m and $m-1$, respectively, and $v_m^{(n)} := L^n v_m$ denotes the chain generated by v_m corresponding to operator L and eigenvalue zero. In particular, it holds that

$$\partial_2 u_m^+(x_1,b) = v_m(x_1), \quad \partial_2 u_m^-(x_1,b) = v_{m-1}(x_1),$$

Proof. By Lemma 4.11, the vector functions $W_m^+ := (h_m, v_m)^T$, $W_m^- := (h_m, v_{m-1})^T$ are of rank $2m$ and $2m-1$, respectively. Now, consider the quasiperiodic boundary value problems

$$\partial_2 W^\pm = MW^\pm, \qquad W^\pm(\cdot,b) = W_m^\pm,$$

where $W^\pm = (u_m^\pm, \partial_2 u_m^\pm)^T$. By Lemma 4.8, we have the solution

$$W^\pm(x) = \sum_{j=0}^{2m-1} (W_m^\pm)^{(j)}(x_1)\,(x_2-b)^j/j\,!,$$

where $(W_m^\pm)^{(j)} = M^j W_m^\pm$ denotes the chain generated by $W_m^\pm$. By the definition of M, we get

$$(W_m^+)^{(2n)} = \begin{pmatrix} h_m^{(n)} \\ v_m^{(n)} \end{pmatrix}, \quad (W_m^+)^{(2n+1)} = \begin{pmatrix} v_m^{(n)} \\ h_m^{(n+1)} \end{pmatrix}, \quad n=0,1,\cdots,m-1.$$

The first component of $(W_m^+)^{(j)}$, $j=0,1,\cdots,2m-1$ gives the definition of $w_m^{j,+}$ in (4.32). Analogously, we can get

$$(W_m^-)^{(2n)} = \begin{pmatrix} h_m^{(n)} \\ v_{m-1}^{(n)} \end{pmatrix}, \quad (W_m^-)^{(2n+1)} = \begin{pmatrix} v_{m-1}^{(n)} \\ h_m^{(n+1)} \end{pmatrix}, \quad n=0,1,\cdots,m-1,$$

which imply the expressions of $w_m^{j,-}$. The representation of $\partial_2 u_m^\pm$ on $x_2=b$ follows from the expression of $u_m^\pm$ and definition of $w_m^{1,\pm}$. □

In the case of $m=1$, we have

$$u_1^+(x) = h_1(x_1) + (x_2-b)v_1(x_1), \quad u_1^-(x_1) = h_1(x_1).$$

For $m\geq 1$, the functions $u_m^\pm(x_1,x_2)$ are polynomials with respect to x_2 of order $2m-1$ and $2m-2$, respectively. Since u_1^+ and $u_m^\pm$ ($m\geq 2$) are unbounded as $x_2\to\pm\infty$, these wave modes are physically not meaningful. Hence, in this paper we make the assumption that the rank of $\mu=0$ of L is one and the corresponding eigenfunction is given by $u=u_1^-=h_1(x_1)$, which coincides with the solution obtained by Thm. 4.9 by formally setting $\mu_n=0$ and $m=1$ (cf. (4.30)). Note that for complex-valued periodic potential $q\in L^\infty(\mathbb{R})$, one cannot exclude, in general, that zero has an associated eigenfunction of rank $m\geq 2$.

4.4 Upward and downward radiation conditions

Suppose the operator L in (4.1) is defined with a function $q\in L^\infty(\mathbb{R})$. We introduce the following assumption on L.

Definition 4.13. *We shall say that Assumption RC(q) is fulfilled if the system of eigenfunctions corresponding to L (cf. Lemma 4.6) forms a Riesz basis and if either there is no eigenvalue zero of L or any eigenfunction u of eigenvalue zero is of rank one, i.e., $L^2u=0$ implies $Lu=0$.*

We suppose the space is filled with material, the refractive index $\tilde{q}(x)$ of which is equal to $q^+(x_1)$ and to $q^-(x_1)$ in an upper and lower half space, respectively. Denote the operator L of (4.1) with $q=q^\pm$ by $L^\pm$. In this and the following sections we shall assume the Assumptions RC($q^\pm$). For $q=q^\pm$ and $L=L^\pm$, the Riesz basis $\{h_n : n\in\mathbb{Z}\}$ can be denoted by $\{h_{n,m} : \tilde{\mu}_n\in\sigma_p(L), h_{n,m}\in A_L^F(\tilde{\mu}_n)\}$ with a finite subset $A_L^F(\tilde{\mu}_n)\subset A_L(\tilde{\mu}_n)$. Whereas the eigenvalues μ_n, $n\in\mathbb{Z}$ in Lemma 4.5, point (iv) need not to be different for different indices n, the eigenvalues $\tilde{\mu}_n$, $n\in\mathbb{N}$ in the new notation satisfy $\tilde{\mu}_1<\tilde{\mu}_2<\tilde{\mu}_3<\cdots$. Setting $\mathcal{I}:=\{(n,m):\ n\in\mathbb{N},\ m\in A_L^F(\mu_n)\}$, we can even write the system as $\{h_{n,m} : (n,m)\in\mathcal{I}\}$. The subscript $m\geq 1$ indicates the rank m of eigenfunction $h_{n,m}$, and the corresponding set of eigenpairs is $\{(h_{n,m},\tilde{\mu}_n) : (n,m)\in\mathcal{I}\}$. To simplify notation we even write μ_n for the new $\tilde{\mu}_n$. Furthermore, suppose $u_{n,m}^\pm$ is given by (4.24) and let $\lambda_n^\pm$ and $W_{n,m}^\pm$ be defined as in (4.19). Set

$$\hat{\lambda}_n := \begin{cases} \sqrt{\mu_n} & \text{if } \operatorname{Re}\sqrt{\mu_n}<0 \text{ or } \operatorname{Re}\sqrt{\mu_n}=0,\ \operatorname{Im}\sqrt{\mu_n}\geq 0, \\ -\sqrt{\mu_n} & \text{otherwise.} \end{cases} \tag{4.33}$$

It is clear that we always have either $\mathrm{Re}\,(\hat{\lambda}_n) < 0$ or $\mathrm{Re}\,(\hat{\lambda}_n) = 0$ and $\mathrm{Im}\,(\hat{\lambda}_n) \geq 0$. Similarly, define

$$\widehat{W}_{n,m} := \begin{cases} W^+_{n,m} & \text{if} \quad \mathrm{Re}\,\sqrt{\mu_n} < 0 \quad \text{or} \quad \mathrm{Re}\,\sqrt{\mu_n} = 0,\, \mathrm{Im}\,\sqrt{\mu_n} \geq 0, \\ W^-_{n,m} & \text{otherwise.} \end{cases}$$

Note that, for $\hat{\lambda}_n = 0$, we have $m = 1$ and $\widehat{W}_{n,m} = \widehat{W}_{n,1} = (h_{n,1}, 0)^T$, where $h_{n,1} = h_1$ denotes the eigenfunction of rank one that corresponds to the eigenvalue zero and operator L.

Definition 4.14. *An upward (resp. downward) radiating mode $u^{(U)}_{n,m}$ (resp. $u^{(D)}_{n,m}$) is defined as*

$$u^{(U)}_{n,m} = e^{\hat{\lambda}_n(x_2-b)} \sum_{j=0}^{m-1} (\widehat{W}_{n,m})^{(j)}_1(x_1) \frac{(x_2-b)^j}{j\,!}, \quad x_2 \geq b,$$
$$u^{(D)}_{n,m} = e^{-\hat{\lambda}_n(x_2-b)} \sum_{j=0}^{m-1} (\widehat{W}_{n,m})^{(j)}_1(x_1) \frac{(x_2-b)^j}{j\,!}, \quad x_2 \leq b.$$

We shall call the modes $u^{(U)}_{n,m}$ and $u^{(D)}_{n,m}$ propagating wave mode if $\mathrm{Re}\,\hat{\lambda}_n = 0$, i.e., if it is not decaying exponentially for $x_2 \to \infty$ and $x_2 \to -\infty$, respectively.

Remark 4.15. *Each upward and downward radiating mode belongs to $H^2_{loc}(\mathbb{R}^2)$. For $\alpha \neq 0, 1/2$ and for $|n|$ sufficiently large, by Lemma 4.5 (iv) the eigenpair (h_n, μ_n) of L has the rank one. Together with Theorem 4.9, this implies that*

$$u^{(U)}_{n,m} = u^{(U)}_n = e^{\hat{\lambda}_n(x_2-b)} h_n, \quad u^{(D)}_{n,m} = u^{(D)}_n = e^{-\hat{\lambda}_n(x_2-b)} h_n.$$

Independent on whether the rank is one or two, for large $|n|$ the function $u^{(U)}_n$ (resp. $u^{(D)}_n$) decays exponentially as $x_2 \to +\infty$ (resp. $x_2 \to -\infty$), due to the definition of $\hat{\lambda}_n$ and the asymptotics of $\hat{\lambda}_n$ shown in Lemma 4.5 (iv).

Definition 4.16. *The α-quasiperiodic function $u \in H^1_{loc}(\Omega^+_b)$ (resp. $u \in H^1_{loc}(\Omega^+_b)$) is called an upward (resp. downward) radiating solution if u is a linear combination of the upward (resp. downward) radiating modes, that is,*

$$u(x) = \sum_{(n,m)\in\mathcal{I}} C^+_{n,m}\, u^{(U)}_{n,m}(x),$$
$$\text{(resp.)} \quad u(x) = \sum_{(n,m)\in\mathcal{I}} C^-_{n,m}\, u^{(D)}_{n,m}(x),$$

for some sequence of coefficients $C^{\pm}_{n,m} \in \mathbb{C}$. The sums converge in $H^1_{loc}(\Omega^+_b)$ (resp. $H^1_{loc}(\Omega^+_b)$).

Recall our definition of $H^1_{loc}(\Omega^{\pm}_b)$ as the space of all functions v over $\Omega^{\pm}_b$ such that, for any radius $r > 0$, the restriction of v to $\Omega^{\pm}_{b,r} := \{x \in \Omega^{\pm}_b :\ |x| < r\}$ is in $H^1(\Omega^{\pm}_{b,r})$. Note that the functions $u \in H^1_{loc}(\Omega^{\pm}_b)$ of Def. 4.16 satisfy the Helmholtz equation $\Delta u(x_1, x_2) + k^2 q(x_1) u(x_1, x_2) = 0$ for $(x_1, x_2) \in \Omega^{\pm}_b$.

If $q(x)\equiv q_0\in\mathbb{C}$, the upward and downward propagating modes defined in Definitions 4.16 and 4.14 are exactly the Rayleigh modes occurring in a homogeneous periodic medium. In fact, the spectrum (μ_n, h_n) of the differential operator L is given by

$$\mu_n = \alpha_n^2 - k^2 q_0 \in \mathbb{C}, \qquad h_n(x_1) = \exp(i\alpha_n x_1), \quad n \in \mathbb{Z}.$$

In particular, each eigenvalue μ_n is of rank one and there is no associated eigenfunctions of rank $m\geq 2$ (see the arguments below). Correspondingly, the spectrum (λ_n, W_n) of the matrix differential operator M can be represented as (see Lemma 4.3)

$$\lambda_n^{\pm} = \pm\sqrt{\alpha_n^2 - k^2 q_0}, \quad W_n^{\pm} = \exp(i\alpha_n x_1)\begin{pmatrix} 1 \\ \pm\sqrt{\alpha_n^2 - k^2 q_0}\end{pmatrix}.$$

Note that the branch of $\sqrt{a}$ is taken such that $\operatorname{Im}\sqrt{a}\geq 0$ for $a\in\mathbb{C}$. By the definition (4.33), the parameter $\hat{\lambda}_n\in\mathbb{C}$ turns out to be

$$\hat{\lambda}_n := \begin{cases} -\sqrt{\alpha_n^2 - k^2 q_0} & \text{if } \ |\alpha_n|^2 > |k^2 q_0|, \\ \sqrt{k^2 q_0 - \alpha_n^2} & \text{if } \ |\alpha_n|^2 \leq |k^2 q_0|. \end{cases}$$

Hence, the upward and downward going modes take the form

$$\begin{aligned} u_n^{(U)}(x) &= e^{i\alpha_n x_1 + \hat{\lambda}_n (x_2 - b)}, \quad x_2 \geq b, \\ u_n^{(D)}(x) &= e^{i\alpha_n x_1 - \hat{\lambda}_n (x_2 - b)}, \quad x_2 \leq b. \end{aligned}$$

In the special case $q(x)\equiv 1$, it holds that

$$\hat{\lambda}_n := \begin{cases} -\sqrt{\alpha_n^2 - k^2} & \text{if } \ |\alpha_n| > k, \\ i\sqrt{k^2 - \alpha_n^2} & \text{if } \ |\alpha_n| \leq k, \end{cases}$$

which coincides with $i\beta_n$ for any $n\in\mathbb{Z}$ (cf. (2.3)). If $\mu_n=0$ is an eigenvalue of L, we have either $\alpha_n=k$ or $\alpha_n=-k$, that is, the dimension of the eigenspace $\sigma_L(0)$ is at most two, with the eigenfunctions $e^{\pm ikx_1}$. These eigenmodes can be regarded as both upward and downward going modes. When $\alpha_n=0$ for some $n\in\mathbb{Z}$, it holds that $u_n^{(U)}(x)=e^{ikx_2}$ and $u_n^{(U)}(x)=e^{-ikx_2}$, which are 2π-periodic wave modes in the x_2-direction.

Next we show that the rank of the eigenvalue μ_n of the operator $L=-(\partial_1^2+k^2q_0 I)$ with $q_0\in\mathbb{C}$ is at most one. For this purpose, it suffices to prove that, for any given $n\in\mathbb{N}$, there do not exist α-quasiperiodic solutions to the ordinary differential equation

$$w''(x_1) + \alpha_n^2 w(x_1) = e^{i\alpha_n x_1}, \quad x_1 \in \mathbb{R}. \tag{4.34}$$

If $\alpha_n\neq 0$, a general solution to (4.34) takes the form

$$\begin{aligned} w(x_1) &= c^+ e^{i\alpha_n x_1} + c^- e^{-i\alpha_n x_1} + v(x_1), \quad c^{\pm} \in \mathbb{C}, \\ v(x_1) &= \frac{1}{\alpha_n}\int_0^{x_1} \sin\big(\alpha_n(x_1 - y_1)\big) e^{i\alpha_n y_1}\, dy_1 \\ &= \frac{-e^{i\alpha_n x_1}}{4\alpha_n^2}\left(e^{-i2\alpha_n x_1} - 1 + i2\alpha_n x_1\right). \end{aligned} \tag{4.35}$$

DOI 10.20347/WIAS.PREPRINT.2726

It is easy to see

$$v'(x_1) = \int_0^{x_1} \cos\left(\alpha_n(x_1 - y_1)\right) e^{i\alpha_n y_1}\, dy_1 = \frac{i\, e^{i\alpha_n x_1}}{4\alpha_n} \left(e^{-i2\alpha_n x_1} - 1 - i2\alpha_n x_1\right) \qquad (4.36)$$

and $v(0)=v'(0)=0$. The function w is α-quasiperiodic in x_1 if $w(0) = w(2\pi)e^{-i2\pi\alpha}$ and $w'(0) = w'(2\pi)e^{-i2\pi\alpha}$. Since $e^{i\alpha_n x_1}$ is α-quasiperiodic, we get conditions on c^- and can assume $c^+=0$. The first condition together with $\alpha_n = \alpha + n$ leads us to $(c^- + v(0))e^{i\alpha 2\pi} = (c^- e^{-i\alpha 2\pi} + v(2\pi))$ i.e., to the formula $2i\sin(\alpha 2\pi)c^- = v(2\pi)$. Similarly, the second condition for the derivatives implies $2i\sin(\alpha 2\pi)c^- = \frac{i}{\alpha_n} v'(2\pi)$. In other words, an existence of a quasiperiodic solution (4.34) requires $v(2\pi) = \frac{i}{\alpha_n} v'(2\pi)$. Substituting $x_1 = 2\pi$ into the formulas (4.35) and (4.36), we get $\alpha_n = 0$, which is a contradiction to the assumption $\alpha_n \neq 0$ for our case. If $\alpha_n = 0$, it holds that $\alpha = -n$ for some $n \in \mathbb{Z}$, implying that the solution w to the ordinary equation $w'' = 1$ must be 2π-periodic. A general solution of (4.34) is given by $w(x_1) = 1/2\ x_1^2 + ax_1 + b$ with $a, b \in \mathbb{C}$. However, such general solutions cannot be 2π-periodic. In summary, eigenvalues for constant potentials cannot be of rank $m \geq 2$.

5 Solvability of grating diffraction problems in an inhomogeneous periodic medium

The results on the solvability of the boundary value problem, modeling the scattering of an incoming wave by the grating structure between inhomogeneous media, goes along the same lines as in the case of homogeneous cover and substrate materials. In Subsect. 5.1, we shall define Dirichlet-to-Neumann (DtN) mappings over the lower boundary line of the cover material and over the upper boundary of the substrate. Mapping properties of these DtN operators will be investigated in Lemmata 5.3, 5.4 and 5.5. In particular, definiteness and strong ellipticity of the quadratic forms corresponding to the two Dirichlet-to-Neumann mappings are presented. In Section 5.2, we formulate the scattering problem as a quasiperiodic boundary value problem. An equivalent variational formation is given by enforcing the Dirichlet-to-Neumann mappings on an artificial boundary inside the inhomogeneous material, and the strong ellipticity of the corresponding sesqui-linear form is proved. The definiteness of the quadratic forms imply the uniqueness of the scattered far-field, namely the reflected and transmitted propagating wave modes. By Fredholm's alternative, we obtain unique solvability of the scattering problem for absorbing materials and also existence of solutions in non-absorbing materials for special incoming waves.

5.1 Dirichlet-to-Neumann mappings

Again (cf. Subsect. 4.4), in contrast to the notation h_n, $n \in \mathbb{N}_0$ for the system of eigenfunctions and associated eigenfunctions used in Subsect. 4.2 (cf. Lemma 4.6), we denote the system by $h_{n,m}$, $(n,m) \in \mathcal{I}$ with the new index set $\mathcal{I} := \{(n,m) \colon n \in \mathbb{N},\ m \in A_L^F(\mu_n)\}$. The index m denotes the rank of the associate eigenfunction $h_{n,m} \in A_L(\mu_n)$ for the eigenvalue μ_n introduced after Lemma 4.7. In the subsequent sections we identify the straight line Γ_b with the finite section over a single period $\{(x_1, b) \colon x_1 \in (0, 2\pi)\}$. For $d > b$, we define the rectangular domain

$R_{b,d}:=\{x\in\mathbb{R}^2:\ b<x_2<d,\ 0<x_1<2\pi\}$. Hence, $\Gamma_b\cup\Gamma_d$ is a subset of the boundary of $R_{b,d}$.

Lemma 5.1. *The system $h_{n,m},\ (n,m)\in\mathcal{I}$ is complete in $H_\alpha^{1/2}(\Gamma_b)$. If it is a Riesz basis in $L^2(\Gamma_b)$, then a scaled version of the system is a Riesz basis in $H_\alpha^{1/2}(\Gamma_b)$.*

Proof. In accordance with Lemma 4.6 the linear span of the system $h_{n,m},\ (n,m)\in\mathcal{I}$ is dense in $L^2(\Gamma_b)$. Using that $L^2(\Gamma_b)$ is a dense subspace in $H_\alpha^{-1}(\Gamma_b)$, we conclude that the span of system $h_{n,m},\ (n,m)\in\mathcal{I}$ is dense in $H_\alpha^{-1}(\Gamma_b)$ as well. Now, knowing that $q\in L^\infty$, we can choose a real number κ such that $A:=L+\kappa I:\ H_\alpha^1(\Gamma_b)\to H_\alpha^{-1}(\Gamma_b)$ is invertible. Then the span of system $A^{-1}h_{n,m},\ (n,m)\in\mathcal{I}$ is dense in $H_\alpha^1(\Gamma_b)$. However, the $h_{n,m}$ are eigenfunctions or associate eigenfunctions of operator A. Consequently, the span of system $A^{-1}h_{n,m},\ (n,m)\in\mathcal{I}$ coincides with the span of the system $h_{n,m},\ (n,m)\in\mathcal{I}$. In other words, the span of system $h_{n,m},\ (n,m)\in\mathcal{I}$ is dense in $H_\alpha^1(\Gamma_b)$. Since $H_\alpha^1(\Gamma_b)$ is dense in $H_\alpha^{1/2}(\Gamma_b)$, the span of system $h_{n,m},\ (n,m)\in\mathcal{I}$ is dense in $H_\alpha^{1/2}(\Gamma_b)$. The Riesz basis property follows from Lemma 4.7. □

In the following definition, we suppose Assumption RC(q) of Def. 4.13 and extend q from Ω_b^- to $\mathbb{R}^2$ by setting $q(x)=q(x_1)$ for all $x\in\mathbb{R}^2$.

Definition 5.2. *The Dirichlet-to-Neumann maps $\mathcal{T}_b^\pm$ for upward and downward radiating solutions are defined as*

$$\mathcal{T}_b^\pm(f):=\pm(\partial_2 u_\pm^{sc})|_{\Gamma_b},\qquad f\in H_\alpha^{1/2}(\Gamma_b),$$

where $u_\pm^{sc}$ are the upward and downward radiating solutions to the Dirichlet boundary value problem

$$\Delta u_\pm^{sc}+k^2 q u_\pm^{sc}=0\quad\text{for}\quad x_2\geq b\ (x_2\leq b),\qquad u_\pm^{sc}|_{\Gamma_b}=f. \tag{5.1}$$

Given $f\in H_\alpha^{1/2}(\Gamma_b)\subset L_\alpha^2(\Gamma_b)$, by Lemmas 4.6 and 4.7 we may expand f into the series

$$f=\sum_{(n,m)\in\mathcal{I}} f_{n,m}\,h_{n,m},\quad f_{n,m}:=\langle f,h_{n,m}^*\rangle\in\mathbb{C}, \tag{5.2}$$

where $\{h_{n,m}^*\}$ is the dual system of $\{h_{n,m}\}$. Recall the equivalent norm (cf. Lemma 4.7 valid for the Riesz basis $h_{n,m},\ (n,m)\in\mathcal{I}$)

$$||f||^2_{H_\alpha^{1/2}(\Gamma_b)}\sim\sum_{(n,m)\in\mathcal{I}}(1+|n|)\,|f_{n,m}|^2+\sum_{(n,m)\in\mathcal{I}_d}(1+\kappa_{n,m})^{1/2}\,|f_{n,m}|^2. \tag{5.3}$$

Using Theorem 4.9, the solution $u_\pm^{sc}\in H^1_{loc}(\Omega_b^\pm)$ to the boundary value problem (5.1) takes the form

$$u_+^{sc}=\sum_{(n,m)\in\mathcal{I}} f_{n,m}\,u_{n,m}^{(U)},\quad x_2\geq b, \tag{5.4}$$

$$u_-^{sc}=\sum_{(n,m)\in\mathcal{I}} f_{n,m}\,u_{n,m}^{(D)},\quad x_2\leq b. \tag{5.5}$$

Lemma 5.3. *Suppose Assumption RC(q) given in Def. 4.13. Then the sums in (5.4) and (5.5) converge in $H^1_{loc}(\Omega_b^+)$, and the mappings $\mathcal{T}_b^\pm$ are continuous from $H_\alpha^{1/2}(\Gamma_b)$ to $H_\alpha^{-1/2}(\Gamma_b)$.*

Proof. Without loss of generality we consider the case of $+$ and upgoing waves. Any approximation of $\mathcal{T}_b^+$, defined by a finite section of the index set, is obviously continuous. Thus, due to Lemma 4.5 (iv), we may suppose that all $h_{n,m}$ are eigenfunctions of rank one or two for eigenvalues μ_n with $\operatorname{Re}\mu_n > 0$. First we assume that all these eigenfunctions are of rank one. We fix a small $\varepsilon_D > 0$. If $h^*_{n,1}$ is a function in the dual system, then

$$T_{co}f(x_1) \;:=\; u_+^{sc}(x_1, b+\varepsilon_D) \;=\; \sum_n \langle f, h^*_{n,1}\rangle u^{(U)}_{n,1}(x_1, b+\varepsilon_D).$$

We assume that the sum contains only a finite number of terms. From Lemma 4.5, (iv) and $u^{(U)}_{n,1}(x_1,x_2) = \exp\big(-\sqrt{\mu_n}(x_2-b)\big)h_{n,1}(x_1)$, we obtain

$$|u_+^{sc}(x_1, b+\varepsilon_D)| \;\le\; c\sum_n \|f\|_{L^2(\Gamma_b)} \exp[-\operatorname{Re}\sqrt{\mu_n}\,\varepsilon_D\,] \;\le\; c\|f\|_{H_\alpha^{1/2}(\Gamma_b)}.$$

Similarly, we can estimate $|\partial^2_{x_1} u_+^{sc}(x_1, b+\varepsilon_D)|$ if we use that $h_{n,1}$ is an eigenfunction of L. We arrive at

$$\|T_{co}f\|_{H_\alpha^{1/2}(\Gamma_b)} = \|u_+^{sc}|_{\Gamma_{b+\varepsilon_D}}\|_{H_\alpha^{1/2}(\Gamma_b)} \;\le\; c\|f\|_{H_\alpha^{1/2}(\Gamma_b)}.$$

Now we use the continuity of the Dirichlet problem for α-quasiperiodic Helmholtz solutions in the rectangle $R_{b,b+\varepsilon_D}$. For sufficiently small ε_D, the variational form $(u,v) \mapsto -\int \nabla u \cdot \nabla \hat{v} + k^2 \int q u \hat{v}$ of the quasiperiodic Dirichlet problem

$$\Delta u(x) + k^2 q(x_1) u(x) = 0,\; x \in R_{b,b+\varepsilon_D}, \quad u|_{\Gamma_b} = f,\; u|_{\Gamma_{b+\varepsilon_D}} = f_2 \tag{5.6}$$

is coercive over the space of functions $u \in H^1_\alpha(\mathbb{R}_{b,b+\varepsilon_D})$ with $u|_{\Gamma_b} = 0$ and $u|_{\Gamma_{b+\varepsilon_D}} = 0$. We denote the solution of (5.6) by $U[f, f_2]$ and get

$$\|U[f,f_2]\|_{H^1_\alpha(R_{b,b+\varepsilon_D})} \;\le\; c\|f\|_{H_\alpha^{1/2}(\Gamma_b)} + c\|f_2\|_{H_\alpha^{1/2}(\Gamma_{b+\varepsilon_D})}.$$

as well as $U[f,f_2] = u_+^{sc}|_{R_{b,b+\varepsilon_D}}$. We conclude

$$\begin{aligned}\|\mathcal{T}_b^+ f\|_{H^{-1/2}(\Gamma_b)} \;&\le\; c\|U[f, T_{co}f]\|_{H^1(R_{b,b+\varepsilon_D})} \;\le\; c\left\{\|f\|_{H_\alpha^{1/2}(\Gamma_b)} + \|T_{co}f\|_{H_\alpha^{1/2}(\Gamma_{b+\varepsilon_D})}\right\}\\ &\le\; c\|f\|_{H_\alpha^{1/2}(\Gamma_b)}.\end{aligned}$$

Consequently, we can extend $\mathcal{T}_b^+$ to a continuous operator over $H_\alpha^{1/2}(\Gamma_b)$, and the sum (5.4) converges in $H^1_\alpha(R_{b,b+\varepsilon_D})$. Similarly, we get convergence and boundedness in $H^1_\alpha(R_{b+\varepsilon_D,b+2\varepsilon_D})$, in $H^1_\alpha(R_{b+2\varepsilon_D,b+3\varepsilon_D})$, and so on. In other words, we get convergence in $H^1_{loc}(\Omega_b^+)$.

If there exist rank-two eigenfunctions in the sum, then we can proceed similarly. We only have to use Cor. 4.10 together with (4.16) and $c_{n,k,j} = \mathcal{O}(|n|^4),\; k,j = 1,2$, which has been shown at the end of the proof to Lemma 4.7. □

Below we investigate other properties of $\mathcal{T}_b^{\pm}$. In contrast to the orthogonal basis $e^{i\alpha_n x_1}$ (identical with its dual system) for a homogeneous medium, the Riesz bases $h_{n,m}$ in our case may not be orthogonal. The following two lemmas for the homogeneous case were justified in a straightforward manner by the definition of DtN mappings. As we shall show, their generalization to media with non-constant but real-valued q is easy. In this paper we shall make use of variational arguments to prove them even for complex-valued q.

Lemma 5.4. *Suppose Assumption RC(q) given in Def. 4.13 and let $f \in H_\alpha^{1/2}(\Gamma_b)$ be given by (5.2) with coefficients $f_{n,m} \in \mathbb{C}$.*

(i) For real-valued q, each mode $u_{n,m}^{(U)}$ (resp. $u_{n,m}^{(D)}$) corresponds to associate eigenfunctions of rank one, i.e., $m=1$. Furthermore, we have

$$\operatorname{Im} \int_{\Gamma_b} \mathcal{T}_b^{\pm} f\, \bar{f} \quad \geq \quad 0 \quad \textit{for all} \quad f \in H_\alpha^{1/2}(\Gamma_b). \tag{5.7}$$

If the equality sign in (5.7) holds, then we have $f_{n,1}=0$ for all n with $\operatorname{Im}\hat{\lambda}_n > 0$, that is, the solution to the boundary value problem (2.4) has no propagating wave mode with $\operatorname{Re}\hat{\lambda}_n = 0$ and $\operatorname{Im}\hat{\lambda}_n > 0$.

(ii) If $\operatorname{Im} q \geq c_q > 0$ on a subdomain, then there is no propagating mode. Moreover, the inequality (5.7) still holds, and, in the case of equality sign, we have $f_{n,m}=0$ for all $(n,m) \in \mathcal{I}$.

Proof. We consider $\mathcal{T}_b^{+}$ and the upward radiating modes only. The case of $\mathcal{T}_b^{-}$ can be treated analogously.

(i) For real-valued q, we have a self-adjoint operator, and there is no $h_{n,m}$ with rank m greater than one. Moreover, the eigenfunctions are orthogonal. Choosing a sufficiently large n_0 and substituting

$$(\mathcal{T}_b^{+} f)(x_1) \quad = \quad \sum_{n \in \mathbb{Z}} \hat{\lambda}_n\, f_{n,1}\, h_{n,1}(x_1)$$

into (5.7), the assertion follows from $\operatorname{Im}\hat{\lambda}_n \geq 0$ and the identity

$$\operatorname{Im} \int_{\Gamma_b} \mathcal{T}_b^{+} f\, \bar{f}\, ds = \sum_{n \in \mathbb{Z}} (\operatorname{Im}\hat{\lambda}_n)\, |f_{n,1}|^2 \int_0^{2\pi} |h_{n,1}(x_1)|^2\, dx_1 \geq 0.$$

(ii) Now consider the boundary value problem (2.4) in $x_2 \geq b$ and suppose $\operatorname{Im} q > 0$ on a set of positive measure. Equ. (5.4) together with Green's formula leads us to

$$\int_{\Gamma_b} \mathcal{T}_b^{+} f\, \bar{f}\, ds \quad = \quad \int_{\Gamma_d} \partial_{x_2} u_+^{sc} \bar{u}_+^{sc}\, ds + \int_{R_{b,d}} \left\{ k^2 q\, |u_+^{sc}|^2 - |\nabla u_+^{sc}|^2 \right\} dx. \tag{5.8}$$

To prove that there is no propagating mode, we only need to consider a propagating mode of rank one. Taking $f := \tilde{h}_n$ with $\operatorname{Re}\hat{\lambda}_n = 0$, we get $\operatorname{Im}\hat{\lambda}_n \geq 0$ and

$$\begin{aligned} u_+^{sc}(x) \quad &= \quad e^{\hat{\lambda}_n (x_2 - b)}\, \tilde{h}_n(x_1), \qquad \text{in} \quad x_2 \geq b, \\ \partial_2 u_+^{sc}(x) \quad &= \quad \hat{\lambda}_n\, e^{\hat{\lambda}_n (d - b)}\, \tilde{h}_n(x_1), \quad \text{on} \quad x_2 = d. \end{aligned}$$

Taking the imaginary part of (5.8) and using $q=q(x_1)$ we get

$$\begin{aligned}\mathrm{Im}\int_{\Gamma_b}\mathcal{T}_b^+\tilde{h}_n\,\overline{\tilde{h}_n}ds &= k^2\int_{R_{b,d}}\mathrm{Im}\,(q)|u_+^{sc}|^2\,dx+\mathrm{Im}\,(\hat{\lambda}_n)\int_{\Gamma_d}|\tilde{h}_n|^2\,ds\\ &= k^2(d-b)\int_0^{2\pi}\mathrm{Im}\,(q)|\tilde{h}_n|^2\,dx_1+\mathrm{Im}\,(\hat{\lambda}_n)\int_0^{2\pi}|\tilde{h}_n|^2\,dx_1,\end{aligned}$$

for any $d>b$. Since the right-hand side should be independent of $d>b$, we conclude that $\int_0^{2\pi}\mathrm{Im}\,(q)|\tilde{h}_n|^2\,dx_1=0$. Hence, $\tilde{h}_n(x_1)=0$ over the subdomain where $\mathrm{Im}\,q(x_1)\geq c_q$. This further yields $u_+^{sc}\equiv 0$ in $x_2\geq b$ by unique continuation of the elliptic equation (see e.g., [20, Theorem 17.2.6, Chapter XVII]) and thus $\tilde{h}_n\equiv 0$.

Next we shall prove the inequality (5.7) for complex-valued $q(x_1)$. For $f=\sum_{n,m}f_{n,m}h_{n,m}$, the solution u_+^{sc} is given by (5.4). As $d\to\infty$, the exponentially decaying terms $u_{n,m}^{(U)}(x_1,d)$ with $\mathrm{Re}\,\hat{\lambda}_n=0$ tend to zero, and only the propagating modes remain. Hence

$$u_+^{sc}(x_1,d)\to\sum_{(n,m)\in\mathcal{I}:\mathrm{Re}\,\hat{\lambda}_n=0}\hat{\lambda}_n\;f_{n,m}\;u_{n,m}^{(U)}(x_1,d)=0,\quad \text{as}\quad d\to\infty.$$

In the last step, we have used the vanishing of the propagating modes, that is, $u_{n,m}^{(U)}\equiv 0$ if $\mathrm{Re}\,\hat{\lambda}_n=0$. Similarly, one can prove that $\partial_2 u_+^{sc}(x_1,d)\to 0$ as $d\to\infty$. Taking the imaginary part of (5.8) and letting $d\to\infty$, we obtain

$$\mathrm{Im}\int_{\Gamma_b}\mathcal{T}_b^+ f\,\bar{f}\,ds=\lim_{d\to\infty}\left\{\int_{R_{b,d}}k^2[\mathrm{Im}\,q]\,|u_+^{sc}|^2\,dx\right\}\geq 0.$$

In the case of equality sign, we must have $u_+^{sc}\equiv 0$ and thus $f_{n,m}=0$ for all $(n,m)\in\mathcal{I}$. □

Lemma 5.5. *Suppose there holds Assumption RC(q) given in Def. 4.13. Then there exists a compact operator* $\mathcal{T}_{b,0}^{\pm}:\ H_\alpha^{1/2}(\Gamma_b)\to H_\alpha^{-1/2}(\Gamma_b)$ *such that*

$$\int_0^{2\pi}\left[-\mathcal{T}_b^{\pm}+\mathcal{T}_{b,0}^{\pm}\right]f\bar{f}\,ds\quad\geq\quad c_0\,||f||^2_{H_\alpha^{1/2}(\Gamma_b)},\quad c_0>0.$$

In other words, $-\mathcal{T}_b^{\pm}$ *can be decomposed into the sum of a coercive operator and a compact operator.*

Proof. The assertions for $\mathcal{T}_b^+$ and $\mathcal{T}_b^-$ follow analogously. So we only consider the case of $\mathcal{T}_b^+$. For $d>b$, the identity (5.8) can be decomposed into two parts:

$$-\int_{\Gamma_b}\mathcal{T}_b^+ f\,\bar{f}\,ds\;=\;\int_{R_{b,d}}\left\{|\nabla u_+^{sc}|^2+|u_+^{sc}|^2\right\}\,dx-\int_{\Gamma_b}\mathcal{T}_{b,0}^+ f\,\bar{f}\,ds \tag{5.9}$$

where $\mathcal{T}_{b,0}^+:\ H_\alpha^{1/2}(\Gamma_b)\to H_\alpha^{-1/2}(\Gamma_b)$ is defined as

$$\int_{\Gamma_b}\mathcal{T}_{b,0}^+ f\;\bar{g}\,ds:=\int_{R_{b,d}}\left\{(1+k^2q)\,u_+^{sc}\,\bar{w}_+^{sc}\right\}dx+\int_{\Gamma_d}\partial_2 u_+^{sc}\;\bar{w}_+^{sc}\,ds,\quad g\in H_\alpha^{1/2}(\Gamma_b).$$

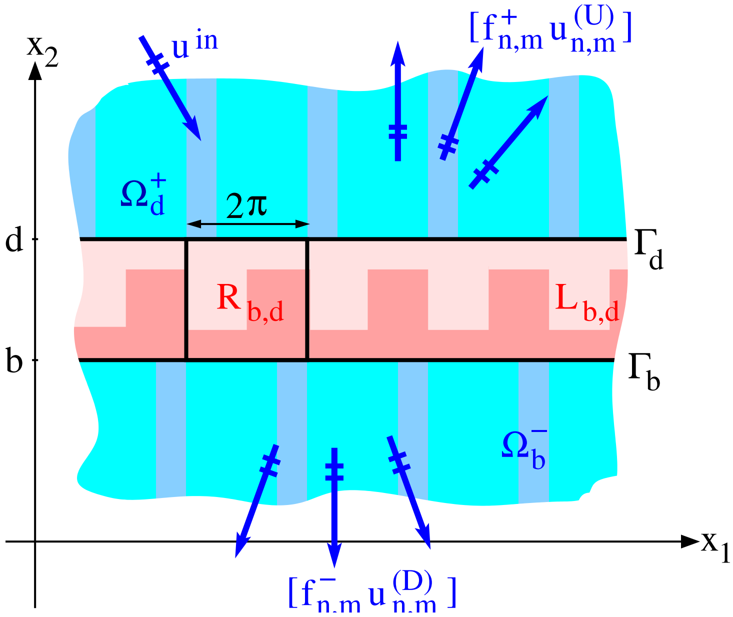

Figure 2: The geometry settings for the boundary value problem.

Here $w_+^{sc} \in H^1(R_{b,d})$ is the unique radiating solution to the boundary value problem (5.1) with the Dirichlet data $w_+^{sc} = g$ on Γ_b. The operator $\mathcal{T}_{b,0}^+$ is compact, because the mappings

$$G_1 : H_\alpha^{1/2}(\Gamma_b) \to H_\alpha^{1/2}(\Gamma_d), \qquad G_1(g) := w_+^{sc}|_{\Gamma_d},$$
$$G_2 : H_\alpha^{1/2}(\Gamma_b) \to L_\alpha^2(R_{b,d}), \qquad G_2(g) := w_+^{sc}|_{R_{b,d}},$$

are both compact. On the other hand, by (5.9) it is clear that $-\mathcal{T}_b^+ + \mathcal{T}_{b,0}^+$ is a coercive operator on $H_\alpha^{1/2}(\Gamma_b)$. □

5.2 Well-posedness of the transmission problem

Next we consider the boundary value problem for the simulation of waves scattered at a grating located between the two inhomogeneous half spaces Ω_d^+ and Ω_b^- with $b < d$ (cf. Fig. 2). In particular, we assume $\tilde{q} \in L^\infty(\mathbb{R}^2)$ such that $\tilde{q}(x) = q^+(x_1)$ for $x_2 \geq d$ and $\tilde{q}(x) = q^-(x_1)$ for $x_2 \leq b$. In other words, the univariate function previously denoted by q is now changed to $q^\pm$. Of course, for the refractive index, we suppose there is a constant $c_q > 0$ such that either $\tilde{q}(x) > c_q$ or $\operatorname{Im} \tilde{q}(x) > c_q$. By $L_{b,d}$ we denote the layer $\{x \in \mathbb{R}^2 : b < x_2 < d\}$ and, as before, by $R_{b,d}$ the rectangle $\{x \in \mathbb{R}^2 : b < x_2 < d, 0 < x_1 < 2\pi\}$. For any given functions $f_D^d \in H_\alpha^{1/2}(\Gamma_d)$, $f_N^d \in H_\alpha^{-1/2}(\Gamma_d)$, $f_D^b \in H_\alpha^{1/2}(\Gamma_b)$, and $f_N^b \in H_\alpha^{-1/2}(\Gamma_b)$, we look for a triple of α-quasiperiodic field solutions $u\in H^1_\alpha(L_{b,d})$, $u^+\in H^1_{\alpha,\,loc}(\Omega_d^+)$, and $u\in H^1_{\alpha,\,loc}(\Omega_b^-)$ of

$$\begin{array}{rcl} \Delta u(x) \;\; + k^2\tilde{q}(x) \quad u(x) &=& 0, \quad x\in L_{b,d},\\ \Delta u^+(x) + k^2 q^+(x_1) u^+(x) &=& 0, \quad x\in \Omega_d^+,\\ \Delta u^-(x) + k^2 q^-(x_1) u^-(x) &=& 0, \quad x\in \Omega_b^-,\\ u|_{\Gamma_d} \;=\; u^+|_{\Gamma_d} + f_D^d, \quad \partial_2 u|_{\Gamma_d} &=& \partial_2 u^+|_{\Gamma_d} + f_N^d,\\ u|_{\Gamma_b} \;=\; u^-|_{\Gamma_b} + f_D^b, \quad \partial_2 u|_{\Gamma_b} &=& \partial_2 u^-|_{\Gamma_b} + f_N^b,\\ u^+ \text{ is an upward radiating wave in } \Omega_d^+, && \\ u^- \text{ is a downward radiating wave in } \Omega_b^-. && \end{array} \tag{5.10}$$

Suppose that $u^{in}\in H^1_{\alpha,\,loc}(\Omega_d^+)$ is a downward incoming wave satisfying the Helmholtz equation $(\Delta+k^2q^+I)u^{in}=0$ in Ω_d^+. Then the wave solution of (5.10) with $f_D^d = u^{in}|_{\Gamma_d}$, $f_N^d = \partial_{x_2}u^{in}|_{\Gamma_d}$, $f_D^b = 0$, and $f_N^b = 0$ is the wave scattered by the grating, i.e., u^+ is the reflected wave, u^- the transmitted wave, and u the wave induced inside the grating.

Clearly, the weak formulation of (5.10) is the variational equation

$$\begin{aligned} a(u,v) &= F(v),\ \forall v\in H^1_\alpha(R_{b,d}), &(5.11)\\ a(u,v) &:= \int_{R_{b,d}} \left\{-\nabla u\cdot\nabla\bar v + k^2\tilde q\, u\,\bar v\right\}\, dx + \int_{\Gamma_d} \mathcal{T}_d^+ u\,\bar v\, ds + \int_{\Gamma_b} \mathcal{T}_b^- u\,\bar v\, ds,\\ F(v) &:= \int_{\Gamma_d} \left[\mathcal{T}_d^+ f_D^d - f_N^d\right]\bar v\, ds + \int_{\Gamma_b} \left[\mathcal{T}_b^- f_D^b + f_N^b\right]\bar v\, ds. \end{aligned}$$

The variational solution $u\in H^1_\alpha(R_{b,d})$ can be extended to Ω_d^+ and Ω_b^- as follows. If u is the weak solution, then we get $u|_{\Gamma_d} - f_D^d = \sum_{n,m} f^+_{n,m} h_{n.m}$ with coefficients $f^+_{n,m}\in\mathbb{C}$ and the eigenfunction $h_{n,m}=h_{n,m}(\Omega_d^+)$ for the domain Ω_d^+. We get the solution for $x_2>d$ by the extension $u^+=\sum_{n,m} f^+_{n,m} u^{(U)}_{n.m}$. For $x_2<b$, we get $u|_{\Gamma_b} - f_D^b = \sum_{n,m} f^-_{n,m} h_{n.m}$ with $f^-_{n,m}\in\mathbb{C}$ and the eigenfunction $h_{n,m}=h_{n,m}(\Omega_b^-)$ for Ω_b^-. The solution for $x_2<b$ is the extension $u^-=\sum_{n,m} f^-_{n,m} u^{(D)}_{n.m}$.

Now we prepare the solvability theorem by

Lemma 5.6. *Suppose the Assumptions RC($q^\pm$) introduced in Def. 4.13 hold. The sesqui-linear form $a\colon H^1_\alpha(R_{b,d})\times H^1_\alpha(R_{b,d})\to\mathbb{R}$ is bounded. Moreover, it is strongly elliptic, i.e., there exists a compact operator $T_{se}\colon H^1_\alpha(R_{b,d})\to H^{-1}_\alpha(R_{b,d})$ and a constant $c_{se}>0$ such that, for all $u\in H^1_\alpha(R_{b,d})$,*

$$|a(u,u) + \langle T_{se}u, u\rangle| \;\;\geq\;\; c_{se}\|u\|^2_{H^1_\alpha(R_{b,d})},$$

where $\langle v,u\rangle$ denotes the duality pairing between $H^{-1}_\alpha(R_{b,d})$ and $H^1_\alpha(R_{b,d})$, which is equal to the L^2 scalar product for $v\in L^2(R_{b,d})$. The right-hand side functional $F\colon H^1_\alpha(R_{b,d})\to 0$ is continuous.

Proof. The boundedness follows from Lemma 5.3, the strong ellipticity from Lemma 5.4. The continuity of F is a consequence of Lemma 5.3. □

Theorem 5.7. *Suppose the Assumptions RC($q^\pm$) introduced in Def. 4.13 hold.*

(i) The space of all weak solutions to the homogeneous boundary value problem (5.10) with $f_D^d = f_D^b = f_N^d = f_N^b = 0$ *has a finite dimension. The space of homogeneous solutions of the adjoint differential operator, i.e.,*

$$\ker \;:=\; \left\{ v \in H^1_\alpha(R_{b,d}) \colon\, a(w,v) = 0,\, \forall w \in H^1_\alpha(R_{b,d}) \right\}$$

has the same finite dimension. There exists a weak solution of (5.10) if and only if, for any $v \in \ker$*, the condition* $F(v) = 0$ *holds. If this solvability condition is satisfied and if* u_p *is a particular solution of (5.10), then the general weak solution is* $u = u_p + u_h$ *with* u_h *a weak solution of the homogeneous boundary value problem (5.10).*

(ii) Assume the function q^+ *is real-valued and let* $\hat{\lambda}_{n_0} = \hat{\lambda}_{n_0}(\Omega_d^+)$ *be defined as in (4.33) such that* $\operatorname{Re}\hat{\lambda}_{n_0} = 0,\ \operatorname{Im}\hat{\lambda}_{n_0} > 0$*. Suppose that the incoming wave* u^{in} *in* Ω_d^+ *is the propagating downward radiating mode* $u^{in} = u^{(D)}_{n_0,1}(\Omega_d^+)$*. Then there exists a weak solution of (5.10) with* $f_D^d = u^{in}|_{\Gamma_d}$*,* $f_N^d = \partial_2 u^{in}|_{\Gamma_d}$ *and* $f_D^b = f_N^b = 0$*.*

(iii) For real-valued squared refractive index $q^\pm$*, the propagating upward (resp. downward) radiating modes in* Ω_d^+ *(resp.* Ω_b^-*) with* $\operatorname{Re}\hat{\lambda}_n = 0$ *and* $\operatorname{Im}\hat{\lambda}_n > 0$ *for the general boundary value problem (5.10) are uniquely determined.*

(iv) Suppose that $\operatorname{Im}\tilde{q}(x) \ge c_{\tilde{q}} > 0$ *over a subdomain* $D_0 \subset R_{b,d}$ *or that* $\operatorname{Im} q^\pm(x_1) \ge c_{q^\pm} > 0$ *over a subinterval of* $[0, 2\pi]$*. Then there exists a unique weak solution* u *of (5.10), and for a constant* $C_s > 0$ *independent of the boundary data* f_D^d*,* f_N^d*,* f_D^b *and* f_N^b *, we get*

$$\begin{aligned} &\|u\|_{H^1_\alpha(R_{b,d})} + \|u^+|_{\Gamma_d}\|_{H^{1/2}_\alpha(\Gamma_d)} + \|u^-|_{\Gamma_b}\|_{H^{1/2}_\alpha(\Gamma_b)} \\ &\quad \le C_s \Big\{ \|f_D^d\|_{H^{1/2}_\alpha(\Gamma_d)} + \|f_N^d\|_{H^{-1/2}_\alpha(\Gamma_d)} + \|f_D^b\|_{H^{1/2}_\alpha(\Gamma_b)} + \|f_N^b\|_{H^{-1/2}_\alpha(\Gamma_b)} \Big\} \end{aligned}$$

Proof. (i) Clearly, part (i) is a simple consequence of Fredholm's alternative applied to the variational equation (5.11), the sesqui-linear form of which is strongly elliptic due to Lemma 5.6.

(ii) We apply (i). Suppose $v \in \ker$ is a solution of the homogeneous adjoint equation. Then we get $\operatorname{Im} a(v,v) = 0$. Using $\operatorname{Im}\int k^2 \tilde{q} v \bar{v} \ge 0$ and Lemma 5.4 over Ω_b^-, we get $\operatorname{Im}\int_{\Gamma_d} \mathcal{T}_b^+ v\, \bar{v} = 0$. In the case of real-valued q^+, the eigenfunctions have rank one and form an orthogonal basis. There is a finite number of eigenvalues $\hat{\lambda}_n$ with $\operatorname{Re}\hat{\lambda}_n = 0$, and the remaining eigenvalues satisfy $\operatorname{Re}\hat{\lambda}_n > 0$. Thus, for $v = \sum_n f_{n,1} h_{n,1}$ it follows from Lemma 5.4 (i) that all propagating modes must vanish, i.e., $f_{n,1} = 0$ for $\operatorname{Im}\hat{\lambda}_n > 0$. In particular, we have $f_{n_0,1} = 0$. Hence, by the choice of the f_D^d, f_N^d, f_D^b,f_N^b and the orthogonality of $h_{n,m}$ we obtain

$$F(v) \;=\; \int_{\Gamma_d} \left[\mathcal{T}_d^+ h_{n_0,1} - h_{n_0,1} \right] \bar{v}\, ds \;=\; (\hat{\lambda}_{n_0} - 1) \bar{f}_{n_0,1} \int_0^{2\pi} |h_{n_0,1}|^2 dx_1 \;=\; 0.$$

The solution exists by Fredholm's alternative in part (i) of the lemma.

(iii) As shown in the proof of (ii), it follows from the variational formulation for the homogeneous boundary value problem that

$$\operatorname{Im}\int_{\Gamma_d} \mathcal{T}_d^+ u^+\, \overline{u^+}\, ds + \operatorname{Im}\int_{\Gamma_b} \mathcal{T}_b^- u^-\, \overline{u^-}\, ds = 0,$$

which together with Lemma 5.4 (i) proves the assertion.

(iv) We have to show that any weak solution u of the homogeneous problem is identically zero. From the variational equation (5.11) we conclude $\mathrm{Im}\, a(u,u)=0$ and thus

$$0 = \mathrm{Im}\, a(u,u) \geq \int_{D_0} k^2\, \mathrm{Im}\, q\, |u|^2\, dx + \mathrm{Im} \int_{\Gamma_d} \mathcal{T}_d^+ u\, \bar{u}\, ds + \mathrm{Im} \int_{\Gamma_b} \mathcal{T}_b^- u\, \bar{u}\, ds \geq 0.$$

Applying Lemma 5.4 gives $u\equiv 0$ over D_0 if $\mathrm{Im}\, q(x)\geq c_{\tilde{q}}>0$ in D_0. Hence, by unique continuation we get $u\equiv 0$ over $R_{b,d}$ (see [20, Theorem 17.2.6, Chapter XVII]). The case of $\mathrm{Im}\, q^{\pm}(x_1)\geq c_{q^\pm}>0$ over a subinterval of $[0,2\pi]$ can be proved analogously by applying Lemma 5.4 (ii). □

Remark 5.8. *Equivalently, we could have formulated the theorem with the data $f_D^d,\, f_N^d$ and $f_D^b,\, f_N^b$ restricted to the subspace of traces $v^-|_{\Gamma_d},\, \partial_2 v^-|_{\Gamma_d}$ of downward radiating waves v^- and to the subspace of traces $v^+|_{\Gamma_b},\, -\partial_2 v^+|_{\Gamma_b}$ of upward radiating waves v^-, respectively (cf. the subsequent Lemma 6.1). Indeed, the problem is linear such that the solution for general data is the superposition of solutions corresponding to the data given as traces of upward and downward radiating waves. However, the solution for $f_D^b=0=f_N^b$ and $f_D^d=v^+|_{\Gamma_d},\, f_N^d=\partial_2 v^+|_{\Gamma_d}$ with v^+ an upward radiating wave is simply $u=0=u^-$ and $u^+=v^+$. Similarly, the solution for $f_D^d=0=f_N^d$ and $f_D^b=v^-|_{\Gamma_d},\, f_N^b=\partial_2 v^-|_{\Gamma_d}$ with v^- a downward radiating wave is simply $u=0=u^+$ and $u^-=v^-$.*

6 Scattering matrix algorithm without discretization

6.1 Splitting into upward and downward radiating functions

In this section, we shall introduce the scattering matrix algorithm on a continuous level, i.e., without discretization by truncated Fourier and wave-mode expansion. We shall consider the boundary value problem (5.10) and introduce the slicing, which is a partition into horizontal layers. Over each boundary line between two such slices we shall define a splitting of the wave functions into upgoing and downgoing parts in this subsection. Using this splitting, in Subsect. 6.2 we shall define a simple integration algorithm for the function valued ODE equivalent to the Helmholtz equation. Of course, this T-matrix algorithm is unstable. However, based on the T-matrix algorithm, we shall define the stable scattering matrix algorithm, the S-matrix algorithm. Note that the S-matrix on the continuous level, used for the algorithm, is nothing else than a solution operator of Thm. 5.7 (cf. Rem. 5.8), i.e., it maps the incoming waves modes to the reflected and transmitted wave solutions. In the classical case of the RCWA method, the material in each slice is supposed to have a refractive index independent of the vertical coordinate x_2. For this case, we shall look at the operator entries in the T- and S-matrix in Subsect. 6.3. Unfortunately, the S-matrix algorithm relies on the inversion of entries in the T-matrix. As we shall see in Subsect. 6.3, the existence of the inverse is not known. Therefore, in Subsect. 6.4 we shall introduce a modification, where the invertibility of a corresponding matrix can be shown under natural conditions. We shall not analyze the discretization of the S-matrix algorithm, though the analysis of the continuous method is the right starting point for a numerical analysis.

Now consider the boundary value problem (5.10) with $q^{\pm}>0$ and $f_D^d=u^{in}|_{\Gamma_d},\, f_N^d=\partial_{x_2} u^{in}|_{\Gamma_d},\, f_D^b=0$, and $f_N^b=0$, where u^{in} is a propagating downward radiating wave mode $u_{n,m}^{(D)}$. We choose a slicing

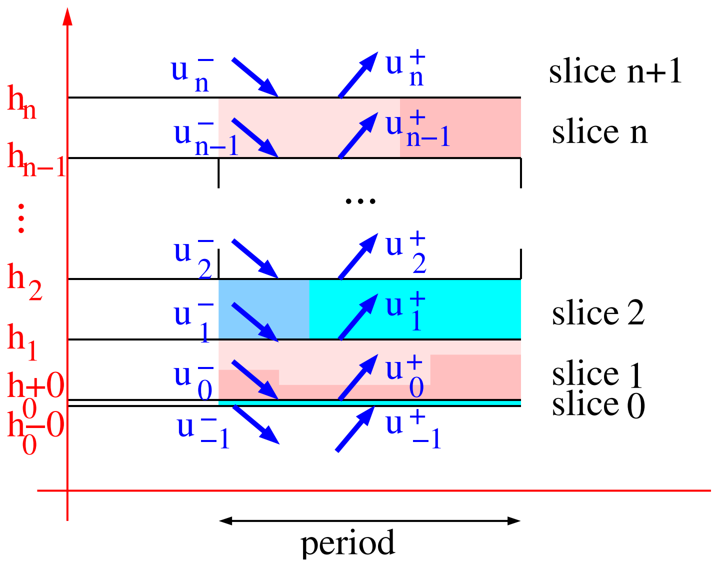

Figure 3: The geometric settings of the scattering matrix algorithm.

of the underlying domain $R_{b,d}$ (cf. Fig. 3), i.e., we fix a partition $h_0 := b < h_1 < \cdots < h_{n-1} < h_n := d$ and write R_{h_{j-1},h_j} for the jth slice of the partition of $R_{b,d}$. Formally, the zeroth slice is defined as the infinitesimally thin slice R_{h_0-0,h_0+0} filled with the material of the squared refractive index $q := q^-$, and the $(n+1)$th slice is $R_{h_n,\infty}$.

At the lower boundary $\Gamma_{j-1} := \Gamma_{h_{j-1}}$ of the jth slice, we consider a splitting of the space of Helmholtz solutions in the space B^+_{j-1} of upward radiating solutions $\sum_{n,m} f^+_{n,m} u^{(U)}_{nm}$ and the space B^-_{j-1} of downward radiating solutions $\sum_{n,m} f^-_{n,m} u^{(D)}_{nm}$. Here the $u^{(U)}_{nm}$ and $u^{(D)}_{nm}$ are the wave modes defined on $\Omega^+_{h_{j-1}}$ and with the univariate q replaced by $x_1 \mapsto \tilde{q}(x_1, h_{j-1}+0)$. More precisely, over the lower boundary line of the slice Γ_{j-1} we split the space for the boundary values $(u|_{\Gamma_{j-1}}, \partial_{x_2} u|_{\Gamma_{j-1}})$ for $u \in H^1_\alpha(R_{h_{j-1},h_j})$ in $B_{j-1} := H^{1/2}_\alpha(\Gamma_{j-1}) \times H^{-1/2}_\alpha(\Gamma_{j-1})$. We split this space as $B_{j-1} = B^+_{j-1} \oplus B^-_{j-1}$ (cf. the subsequent Lemma 6.1), where

$$B^\pm_{j-1} \;:=\; \left\{ (f_D, \pm\mathcal{T}^\pm_{h_{j-1}} f_D) \colon\, f_D \in H^{1/2}_\alpha(\Gamma_{j-1}) \right\},$$

i.e., the space $B^\pm_{j-1}$ contains all boundary data of Helmholtz solutions bounded over the half space $\Omega^\pm_{h_{j-1}}$ satisfying the radiation condition. However, if there is an eigenvalue $\hat{\lambda}_{n_0} = 0$, then a slight modification is needed. For $\hat{\lambda}_{n_0}$, we define

$$\begin{aligned} u^{(U)}_{n_0,1}(x_1,x_2) \;&:=\; \begin{cases} h_{n_0,1}(x_1)(1+[x_2-h_{j-1}]) & \text{if } 0 \le j \le n \\ h_{n_0,1}(x_1) & \text{if } j = n+1 \end{cases}, \\ u^{(D)}_{n_0,1}(x_1,x_2) \;&:=\; \begin{cases} h_{n_0,1}(x_1)(1-[x_2-h_{j-1}]) & \text{if } 0 < j \le n+1 \\ h_{n_0,1}(x_1) & \text{if } j = 0 \end{cases}. \end{aligned} \tag{6.1}$$

These functions are bounded wave modes in the slices, and the wave modes radiating into the half spaces are bounded and physically meaningful.

Lemma 6.1. *Suppose, for function q defined as $q(x_1):=\tilde{q}(x_1,h_{j-1}+0)$, there holds Assumption RC(q) introduced in Def. 4.13. Then the Hilbert space B_{j-1} is the direct sum of the subspaces B^+_{j-1} and B^-_{j-1}.*

Proof. First we show that the intersection $B^+_{j-1}\cap B^-_{j-1}$ is the trivial space $\{(0,0)\}$. If there is a pair of boundary data $(u_D,u_N)\in B^+_{j-1}\cap B^-_{j-1}$ over Γ_{j-1}, then we can extend function u_D to a Helmholtz solution u over $\Omega^\pm_{h_{j-1}}$ (cf. the extensions in Def. 5.2). Thus u is a uniformly bounded Helmholtz solution with refractive index $\tilde{q}_{j-1}(x_1,x_2)=\tilde{q}(x_1,h_{j-1}+0)$ defined over $\mathbb{R}^2$. Suppose $h_{n,m},\ (n,m)\in\mathcal{I}$ is the corresponding system of eigenfunctions and $h^*_{n,m},\ (n,m)\in\mathcal{I}$ the dual system. Then we can show that the functions $x_2\mapsto f_{n,m}(x_2):=\int u(x_1,x_2)\overline{h^*_{n,1}(x_1)}\mathrm{d}x_1$ with rank $m=1$ take the form $f_{n,m}(x_2)=c^+_{n,m}e^{\hat{\lambda}_n x_2}+c^-_{n,m}e^{-\hat{\lambda}_n x_2}$ with constants $c^\pm_{n,m}\in\mathbb{C}$. Indeed, for a smooth function $\varphi(x_2)$ with bounded support, the Helmholtz equation for u and the eigenfunction property $L^*h^*_{n,1}=[\,\overline{\hat{\lambda}_n}\,]^2h^*_{n,1}$ imply

$$
\begin{aligned}
0 &= \Big\langle\nabla u,\nabla(h^*_{n,1}\varphi)\Big\rangle-k^2\Big\langle\tilde{q}_{j-1}u,h^*_{n,1}\varphi\Big\rangle\\
&= \int\Big\{\int\partial_2 u(x)\overline{h^*_{n,1}(x_1)\partial_2\varphi(x_2)}\mathrm{d}x_1\\
&\qquad+\int\Big[\partial_1 u(x)\overline{\partial_1 h^*_{n,1}(x_1)\varphi(x_2)}-k^2q_{j-1}u(x)\overline{h^*_{n,1}(x_1)\varphi(x_2)}\Big]\mathrm{d}x_1\Big\}\mathrm{d}x_2\\
&= \int\Big\{\partial_2\int u(x)\overline{h^*_{n,1}(x_1)}\mathrm{d}x_1\overline{\partial_2\varphi(x_2)}+[\hat{\lambda}_n]^2\int u(x)\overline{h^*_{n,1}(x_1)}\mathrm{d}x_1\overline{\varphi(x_2)}\Big\}\mathrm{d}x_2,
\end{aligned}
$$

which is the weak formulation of $-\partial_2^2 f_{n,1}+[\hat{\lambda}_n]^2 f_{n,1}=0$. Consequently, the well-known formula for the general ODE solution yields $f_{n,1}(x_2)=c^+_{n,1}e^{\hat{\lambda}_n x_2}+c^-_{n,1}e^{-\hat{\lambda}_n x_2}$. For $x_2>h_{j-1}$, Def. 4.14 and (5.4) imply $c^-_{n,1}=0$ and, for $x_2<h_{j-1}$, we similarly get $c^+_{n,1}=0$. Hence $f_{n,1}=0$. Using this fact and the same arguments as above, we get $f_{n,2}=0$, and by induction $f_{n,m}=0$ for any rank m. In other words, $u_D=u|_{\Gamma_{j-1}}$ is orthogonal to the system $h^*_{n,m},\ (n,m)\in\mathcal{I}$, and Lemma 4.6 leads us to $u=0$. Since the extension of u_D under the radiation condition is unique (cf. Def. 5.2), we get $u_N=0$.

It remains to prove that any boundary data (u_D,u_N) with $u_D\in H^{1/2}_\alpha(\Gamma_{j-1})$ and $u_N\in H^{-1/2}_\alpha(\Gamma_{j-1})$ can be represented as the sum of data from B^+_{j-1} and B^-_{j-1}. Here B^+_{j-1} and B^-_{j-1} are closed disjoint subspaces of the Hilbert space B_{j-1}. Clearly, it suffices to prove that data in the dense subset of finite linear combinations of the system functions $h_{n,m}$ admits such a splitting. Equivalently, we have to give the splitting for the boundary data $(h_{n,m},0)$ and $(0,h_{n,m})$. If $\lambda_{n_0}=0$, then restricting (6.1) to Γ_{j-1} implies the representations

$$
\begin{aligned}
(h_{n_0,1},0) &= \frac{1}{2}(h_{n_0,1},h_{n_0,1})+\frac{1}{2}(h_{n_0,1},-h_{n_0,1}) \;=\; \frac{1}{2}\Big(u^{(U)}_{n_0,1},\partial_2 u^{(U)}_{n_0,1}\Big)+\frac{1}{2}\Big(u^{(D)}_{n_0,1},\partial_2 u^{(D)}_{n_0,1}\Big),\\
(0,h_{n_0,1}) &= \frac{1}{2}(h_{n_0,1},h_{n_0,1})-\frac{1}{2}(h_{n_0,1},-h_{n_0,1}).
\end{aligned}
$$

Similarly, if $\lambda_n \neq 0$, then we arrive at

$$\begin{aligned}(h_{n,1},0) &= \frac{1}{2}(h_{n,1},\hat{\lambda}_n h_{n,1}) + \frac{1}{2}(h_{n,1},-\hat{\lambda}_n h_{n,1}),\\ (0,h_{n,1}) &= \frac{1}{2\hat{\lambda}_n}(h_{n,1},\hat{\lambda}_n h_{n,1}) - \frac{1}{2\hat{\lambda}_n}(h_{n,1},-\hat{\lambda}_n h_{n,1}).\end{aligned}$$

For rank $m>1$, we can reduce the rank recursively by (cf. Def. 4.14 and (4.25), and observe that $A_0^{(j)}=0$ for $j\geq 1$)

$$\begin{aligned}(h_{n,m},0) &= \frac{1}{2}(h_{n,m},\hat{\lambda}_n h_{n,m}) + \frac{1}{2}(h_{n,m},-\hat{\lambda}_n h_{n,m}) + \text{rank } (m-1) \text{ terms},\\ (0,h_{n,m}) &= \frac{1}{2\hat{\lambda}_n}(h_{n,m},\hat{\lambda}_n h_{n,m}) - \frac{1}{2\hat{\lambda}_n}(h_{n,m},-\hat{\lambda}_n h_{n,m}) + \text{rank } (m-1) \text{ terms}.\end{aligned}$$

Altogether, any finite linear combination of the $(h_{n,m},0)$ and $(0,h_{n,m})$ can be split by explicit formulas. The resulting parts in $B_{j-1}^{\pm}$ are again such finite linear combinations. □

Of course, there exists a continuous projection P_{j-1}^{+} of B_{j-1} onto B_{j-1}^{+} along B_{j-1}^{-}, and $P_{j-1}^{-}:=I-P_{j-1}^{+}$ is the continuous projection of B_{j-1} onto B_{j-1}^{-} along B_{j-1}^{+}. Note that a boundary value pair $(f_D^{\pm},f_N^{\pm})\in B_{j-1}^{\pm}$ is usually given by the coefficients $f_{n,m}^{\pm}\in\mathbb{C}$ of $f_D^{\pm}=\sum_{n,m} f_{n,m}^{\pm}h_{n,m}$, since $f_N^{+}=\sum_{n,m} f_{n,m}^{+}\partial_{x_2}u_{n,m}^{(U)}$ and $f_N^{-}=\sum_{n,m} f_{n,m}^{-}\partial_{x_2}u_{n,m}^{(D)}$. Splitting into the finite sum of eigenfunctions with rank $m>1$ and the remaining infinite sum, we get

$$\begin{aligned}f_N^{+} &= \sum_{n,m:m>1} f_{n,m}^{+}\partial_{x_2}u_{n,m}^{(U)} + \sum_{n} f_{n,1}^{+}\hat{\lambda}_n h_{n,1}\\ f_N^{-} &= \sum_{n,m:m>1} f_{n,m}^{-}\partial_{x_2}u_{n,m}^{(D)} - \sum_{n} f_{n,1}^{-}\hat{\lambda}_n h_{n,1}.\end{aligned} \tag{6.2}$$

In other words, we identify

$$(f_D^{\pm},f_N^{\pm})\in B_{j-1}^{\pm} \quad\leftrightarrow\quad f_D^{\pm}\in B_{j-1}^{\pm}. \tag{6.3}$$

With this identification we get $B_{j-1}^{\pm}=H_\alpha^{1/2}(\Gamma_{j-1})$. Using the identification (6.3), we even shall write $P_{j-1}^{+}f_D\in H_\alpha^{1/2}(\Gamma_{j-1})$.

The identification (6.3) and the projection $P_{j-1}^{\pm}$ applied to the Dirichlet data f_D, however, is only meaningful if there is a rule to determine f_N for f_D. E.g., for $f_D=f_D^{+}+f_D^{-}$, the $f_N^{\pm}$ and $f_N=f_N^{+}+f_N^{-}$ might be given by (6.2) or, better, by $f_N^{\pm}=\pm\mathcal{T}_{h_{j-1}}^{\pm}f_D^{\pm}$. Then $P_{j-1}^{+}f_D=f_D^{+}$. However, in the case, $f_D=f_{j,D}^{+}+f_{j,D}^{-}$ with $f_{j,N}^{\pm}=\pm\mathcal{T}_{h_j}^{\pm}f_{j,D}^{\pm}$, we have $P_j^{+}f_D=f_{j,D}^{+}$ but, generally, $P_{j-1}^{+}f_D\neq f_{j,D}^{+}$. To get $P_{j-1}^{+}f_D$, we really have to form $(f_D,f_N=f_{j,D}^{+}+f_{j,D}^{-})$, to apply the splitting $(f_D,f_N)=(f_{j-1,D}^{+},f_{j-1,N}^{+})+(f_{j-1,D}^{-},f_{j-1,N}^{-})$, and then to restrict to the Dirichlet part $P_{j-1}^{+}f_D=f_{j-1,D}^{+}$. More precisely, this means $f_D^{\pm}=\sum f_{j,n,m}^{\pm}h_{n,m}$ might be given for the jth basis $\{h_{n,m}=h_{j,n,m}\}$ defined by the eigenfunctions of L based on $q(x_1):=\tilde{q}(x_1,h_j+0)$. We form

$f_N^+ = \sum f_{j,n,m}^+ \partial_{x_2} u_{j,n,m}^{(U)}$ and $f_N^- = \sum f_{j,n,m}^- \partial_{x_2} u_{j,n,m}^{(D)}$ with respect to the jth basis. Thus $f_N = f_N^+ + f_N^-$. Applying a basis transform from the jth basis to the $(j-1)$th basis, we expand

$$\begin{aligned}(f_D, f_N) &= \sum_{n,m} f_{j,n,m}^+ \Big(h_{j,n,m}, \partial_{x_2} u_{j,n,m}^{(U)}\Big) + \sum_{n,m} f_{j,n,m}^- \Big(h_{j,n,m}, \partial_{x_2} u_{j,n,m}^{(D)}\Big) \\ &= \sum_{n,m} f_{j-1,n,m}^+ \Big(h_{j-1,n,m}, \partial_{x_2} u_{j-1,n,m}^{(U)}\Big) + \sum_{n,m} f_{j-1,n,m}^- \Big(h_{j-1,n,m}, \partial_{x_2} u_{j-1,n,m}^{(D)}\Big)\end{aligned}$$

with respect to the $(j-1)$th basis. Finally, we get $P_{j-1}^+ f_D = \sum f_{j-1,n,m}^+ h_{n,m}$ for the $(j-1)$th basis $\{h_{n,m} = h_{j-1,n,m}\}$ defined by the eigenfunctions of L based on $q(x_1) := \tilde{q}(x_1, h_{j-1}+0)$.

6.2 The T- and S-matrix algorithms

Now we are in the position to introduce iterative algorithms for the solution of the boundary value problem (5.10). The Helmholtz equation can be looked at as an ordinary differential equation with respect to x_2, but defined for functions with values, which are functions with respect to x_1. So it is natural to solve the equation like an initial value problem of (4.2). Given the boundary data $u_{j-1} = (u_{j-1,D}, u_{j-1,N})$ over Γ_{j-1}, the solution at $\Gamma_j := \Gamma_{h_j}$ is $u_j = (u_{j,D}, u_{j,N})$. For functions u_{j-1} on Γ_{j-1}, we use the identification (6.3) based on (6.2). For functions u_j on Γ_j, we use the identification (6.3) with $j-1$ replaced by j based on (6.2) with $j-1$ replaced by j. Using the splitting of Lemma 6.1, we get $u_j = u_j^+ + u_j^-$ and write the corresponding operator $\mathbf{T}_j$, $j = 0, 1, \cdots, n$ of integration of the Helmholtz equation as a matrix (cf. Fig. 3).

$$\begin{pmatrix} u_j^+ \\ u_j^- \end{pmatrix} = \mathbf{T}_j \begin{pmatrix} u_{j-1}^+ \\ u_{j-1}^- \end{pmatrix}, \quad \mathbf{T}_j = \begin{pmatrix} \mathbf{T}_j^{++} & \mathbf{T}_j^{+-} \\ \mathbf{T}_j^{-+} & \mathbf{T}_j^{--} \end{pmatrix}, \tag{6.4}$$

Similarly, we introduce the accumulated T-matrices.

$$\begin{pmatrix} u_j^+ \\ u_j^- \end{pmatrix} = \mathcal{T}_j \begin{pmatrix} u_{-1}^+ \\ u_{-1}^- \end{pmatrix}, \quad \mathcal{T}_j = \mathbf{T}_j \, \mathbf{T}_{j-1} \ldots \mathbf{T}_0 = \mathbf{T}_j \mathcal{T}_{j-1}. \tag{6.5}$$

We assume that the local operators $\mathbf{T}_j$ are available. For instance, if $\tilde{q}$ is independent of x_2, then $\mathbf{T}_j$ can be represented with an exponential function of an operator acting on x_1 dependent functions. Equivalently, an expansion of the boundary functions with respect to the wave modes $h_{n,m} = h_{n.m}(\Omega_{h_{j-1}}^{\pm})$ can be computed. Then the solution of the Helmholtz equation is given by the corresponding expansion with respect to the wave modes $u_{n,m}^{(D)}$ and $u_{n,m}^{(U)}$. Alternatively, an ODE solver like the Runge-Kutta method can be employed.

T-matrix algorithm: If the local operators $\mathbf{T}_j$ are available, then we can compute the matrices $\mathcal{T}_j$ recursively for $j = 0, 1, \cdots, n$ by the second equation in (6.5). We arrive at the matrix equation (6.5) for $j = n$. In this system, u_n^- is the given incoming wave and $u_{-1}^+ = 0$ since no wave is arriving from below. The unknowns are the reflected wave u_n^+ and the transmitted wave u_{-1}^-. We get these diffracted waves solving the system. Knowing these functions, even the solution for $h_0 < x_2 < h_n$ can be computed. We start from $j = -1$ and compute the u_j^+ and u_j^- recursively for $j = 0, 1, \cdots, n-1$ using (6.4). Even the values for $h_{j-1} < x_2 < h_j$ can be computed by the above mentioned integration method leading to $\mathbf{T}_j$.

Unfortunately, this T-matrix algorithm is unstable similarly to other ODE integration methods. For instance, the wave-mode expansion with the $u_{n,m}^{(D)}$ contains exponentials which blow up. To overcome this trouble, a stable S-matrix algorithm has been designed. Looking at Thm. 5.7 and Rem. 5.8, we rather have the input of downward radiating waves from above and upward radiating waves from below, and the solution of (5.10) provides us with the resulting upward radiating reflected wave above and with the resulting downward radiating transmitted wave below. In other words, we work with the matrices defined by (cf. Fig. 3)

$$\begin{pmatrix} u_j^+ \\ u_{j-1}^- \end{pmatrix} = \mathbf{S}_j \begin{pmatrix} u_{j-1}^+ \\ u_j^- \end{pmatrix}, \quad \mathbf{S}_j = \begin{pmatrix} \mathbf{S}_j^{++} & \mathbf{S}_j^{+-} \\ \mathbf{S}_j^{-+} & \mathbf{S}_j^{--} \end{pmatrix} \tag{6.6}$$

$$= \begin{pmatrix} \mathbf{T}_j^{++} - \mathbf{T}_j^{+-}[\mathbf{T}_j^{--}]^{-1}\mathbf{T}_j^{-+} & \mathbf{T}_j^{+-}[\mathbf{T}_j^{--}]^{-1} \\ -[\mathbf{T}_j^{--}]^{-1}\mathbf{T}_j^{-+} & [\mathbf{T}_j^{--}]^{-1} \end{pmatrix},$$

$$\begin{pmatrix} u_j^+ \\ u_{-1}^- \end{pmatrix} = \mathcal{S}_j \begin{pmatrix} u_{-1}^+ \\ u_j^- \end{pmatrix}. \tag{6.7}$$

Clearly, for the existence and the boundedness of the S-matrices Thm. 5.7 is useful. To get a recursion for the matrices $\mathcal{S}_j$, we form a system of four equations by joining (6.7) and (6.4) with j replaced by $j+1$. We eliminate $u_j^\pm$ and solve the remaining system with respect to u_{j+1}^+ and u_{-1}^-. Comparing this with (6.7), we obtain

$$\mathcal{S}_{j+1} = \begin{pmatrix} \{\mathbf{T}_{j+1}^{++} - [\mathbf{T}_{j+1}^{++}\mathcal{S}_j^{+-} + \mathbf{T}_j^{+-}]A_j\mathbf{T}_{j+1}^{-+}\}\mathcal{S}_j^{++} & [\mathbf{T}_{j+1}^{++}\mathcal{S}_j^{+-} + \mathbf{T}_j^{+-}]A_j \\ \mathcal{S}_j^{-+} - \mathcal{S}_j^{--}A_j\mathbf{T}_{j+1}^{-+}\mathcal{S}_j^{++} & \mathcal{S}_j^{--}A_j \end{pmatrix}, \tag{6.8}$$

$$A_j := [\mathbf{T}_{j+1}^{--} + \mathbf{T}_{j+1}^{-+}\mathcal{S}_j^{+-}]^{-1}.$$

S-matrix algorithm: The recursion starts with $\mathcal{S}_0 = \mathbf{S}_0$ given by the second equation in (6.6), and then the matrix $\mathcal{S}_j$ is computed recursively for $j = 1, 2, \cdots, n$ by (6.8). If $\mathcal{S}_n$ is computed, then u_n^+ and u_{-1}^- can be computed by (6.7) with j replaced by n. If the intermediate values at $x_2 = h_j$ are of interest, one can utilize the systems (6.7) with respect to u_j^+ and u_j^- for $j = 0, \cdots, n-1$. Even the values for $h_{j-1} < x_2 < h_j$ can be computed by the above mentioned integration method leading to the $\mathbf{T}_j$.

Finally, we note that, for $u_n^- = 0$, the recursion over j of the four matrices $\mathcal{S}_j^{\pm\pm}$ and $\mathcal{S}_j^{\pm\mp}$ can be reduced to a recursion of two matrices and two vectors (compare the subsequent (6.18) of the modified algorithm in Subsect. 6.4). A similar recursion can be derived for accumulated S-matrices defined by $(u_n^+, u_j^-)^\top = \mathcal{S}_j(u_j^+, u_n^-)^\top$ (compare (6.7)). In this case, we get a reduced recursion of two matrices and two vectors for the case $u_{-1}^- = 0$.

6.3 The structure of the T- and S-matrix for $\tilde{q}$ independent of x_2

Now we look at the structure of the matrices $\mathbf{T}_j$ and $\mathbf{S}_j$ over the jth slice, for which we assume $\tilde{q}(x_1, x_2) = \tilde{q}(x_1)$ is independent of x_2 over the slice. We suppose that the boundary value problem

(5.10) over the slice admits a unique solution such that the S-matrix is well defined. We denote the projections of B_j onto $B_j^{\pm}$ along $B_j^{\mp}$ by $P_j^{\pm}$ (cf. the identification (6.3)). Furthermore, we denote the transition operator mapping the boundary data from $u_{j-1}^{\pm}\in B_{j-1}^{\pm}$ to the restriction of the Helmholtz solution to Γ_j by $Tr_j^{\pm}$. In other words, if u satisfies the Helmholtz equation $(\Delta+k^2\tilde{q}I)u=0$ and $u|_{\Gamma_{j-1}}=u_D^{\pm}$ as well as $-\partial_{x_2}u|_{\Gamma_{j-1}}=u_N^{\pm}$ (cf. (6.2)), then $Tr_j^{\pm}(u_D^{\pm},u_N^{\pm}):=(u|_{\Gamma_j},\partial_{x_2}u|_{\Gamma_j})$. Clearly, we write $Tr_j^{+}[u_{n,m}^{(U)}|_{\Gamma_{j-1}}]=u_{n,m}^{(U)}|_{\Gamma_j}$ and $Tr_j^{-}[u_{n,m}^{(D)}|_{\Gamma_{j-1}}]=u_{n,m}^{(D)}|_{\Gamma_j}$. For the eigenfunction $h_{n,1}$ of rank $m=1$, we get $Tr_j^{\pm}h_{n,1}=e^{\pm\hat{\lambda}_n(h_j-h_{j-1})}h_{n,1}$. In general, we can form blocks of all basis functions $h_{n,m}$ with the same eigenvalue μ_n, and the transition operator over such a block is $e^{\pm\hat{\lambda}_n(h_j-h_{j-1})}$ multiplied by a matrix polynomial in (h_j-h_{j-1}) with constant coefficients (cf. Def. 4.14 and (4.25)). Thus the matrix of $Tr_j^{\pm}$ with respect to the system $h_{n,m},\ (n,m)\in\mathcal{I}$ is block diagonal with exponential-polynomial entries. Obviously, we get

$$\mathbf{T}_j^{\pm+}=P_j^{\pm}Tr_j^{+},\quad \mathbf{T}_j^{\pm-}=P_j^{\pm}Tr_j^{-}.$$

On the other hand, any incoming wave $u_{j-1}^{+}=\sum f_{n,m}^{+}h_{n,m}\in B_{j-1}^{+}$ leads to a Helmholtz solution $u=\sum f_{n,m}^{+}u_{n,m}^{(U)}$ over the jth slice such that the downward radiating part at Γ_{j-1} is $u_{j-1}^{-}=0$, and the upward and downward radiating parts at the line Γ_j are $u_j^{\pm}=P_j^{\pm}Tr_j^{+}u_{j-1}^{+}$. We arrive at $\mathbf{S}_j^{++}u_{j-1}^{+}+\mathbf{S}_j^{+-}u_j^{-}=u_j^{+}$ and $\mathbf{S}_j^{-+}u_{j-1}^{+}+\mathbf{S}_j^{--}u_j^{-}=u_{j-1}^{-}$, i.e.,

$$\mathbf{S}_j^{++} = -\mathbf{S}_j^{+-}P_j^{-}Tr_j^{+}+P_j^{+}Tr_j^{+} = -\mathbf{S}_j^{+-}\mathbf{T}_j^{-+}+\mathbf{T}_j^{++}, \tag{6.9}$$

$$\mathbf{S}_j^{-+} = -\mathbf{S}_j^{--}P_j^{-}Tr_j^{+} = -\mathbf{S}_j^{--}\mathbf{T}_j^{-+}. \tag{6.10}$$

In view of the diagonal structure of Tr_j^{+} and the exponential decay of the diagonal entries (cf. point (iv) of Lemma 4.5), we see that $\mathbf{S}_j^{-+}$ is a compact operator. Similarly to the derivation of (6.9) and (6.10), starting with an outgoing vector u_{j-1}^{-} such that $Tr_j^{-}u_{j-1}^{-}\in H_\alpha^{1/2}(\Gamma_j)$ and with $u_{j-1}^{+}=0$, we get $u_j^{\pm}=P_j^{\pm}Tr_j^{-}u_{j-1}^{-}$, i.e.,

$$\mathbf{S}_j^{--}\mathbf{T}_j^{--} = \mathbf{S}_j^{--}P_j^{-}Tr_j^{-} = I. \tag{6.11}$$

Hence the operator entry $\mathbf{T}_j^{--}$, defined over a natural domain of definition, is invertible from the left, and the matrix entry $\mathbf{S}_j^{--}$ is a one-sided inverse for $\mathbf{T}_j^{--}$. However, using the inverse of $\mathbf{T}_j^{--}$, we do not know the value of $\mathbf{S}_j^{--}$ for functions not in the image space of $\mathbf{T}_j^{--}$. So the last equation in (6.6) is correct only if the image of $\mathbf{T}_j^{--}$ is the full space. Therefore, we start with an arbitrary function $u_j^{-}\in B_j^{-}$ and form the projections $P_{s,j-1}^{+}u_j^{-}$ and $P_{s,j-1}^{-}u_j^{-}$ with $P_{s,j-1}^{\pm}$ the projector $P_{j-1}^{\pm}$ shifted from Γ_{j-1} to Γ_j (i.e. $P_{s,j-1}^{-}$ is defined over Γ_j as P_j^{-} but with $\tilde{q}$ from the $(j+1)$th slice replaced by $\tilde{q}$ from the jth slice.). Solving the Helmholtz equation over the jth slice with boundary data $P_{s,j-1}^{\pm}u_j^{-}$ and restricting to Γ_{j-1}, we get the boundary values $[Tr_j^{-}]^{-1}P_{j-1}^{-}u_j^{-}$ and $[Tr_j^{+}]^{-1}P_{j-1}^{+}u_j^{-}$. We arrive at

$$\mathbf{S}_j^{--} = [Tr_j^{-}]^{-1}P_{s,j-1}^{-}|_{B_j^{-}}, \tag{6.12}$$

$$\mathbf{S}_j^{+-} = [Tr_j^{+}]^{-1}P_{s,j-1}^{+}|_{B_j^{-}}. \tag{6.13}$$

Note that the operators $[Tr_j^{-}]^{-1}$ and Tr_j^{+} correspond to a stable integration from above to below and from below to above, respectively. The operators $[Tr_j^{+}]^{-1}$ and Tr_j^{-} are unbounded. Nevertheless, the boundedness of $\mathbf{S}_j^{+-}$ implies the boundedness of $[Tr_j^+]^{-1}$ over the image of $P_{s,j-1}^+|_{B_j^-}$. Collecting (6.9), (6.10), (6.12), and (6.13), we get

$$\begin{aligned}\mathbf{S}_j &= \begin{pmatrix}\mathbf{S}_j^{++} & \mathbf{S}_j^{+-}\\ \mathbf{S}_j^{-+} & \mathbf{S}_j^{--}\end{pmatrix} = \begin{pmatrix}-[Tr_j^+]^{-1}P_{s,j-1}^+\mathbf{T}_j^{-+} + \mathbf{T}_j^{++} & [Tr_j^+]^{-1}P_{s,j-1}^+|_{B_j^-}\\ -[Tr_j^-]^{-1}P_{s,j-1}^-\mathbf{T}_j^{-+} & [Tr_j^-]^{-1}P_{s,j-1}^-|_{B_j^-}\end{pmatrix}\\ &= \begin{pmatrix}-[Tr_j^+]^{-1}P_{s,j-1}^+P_j^-Tr_j^+ + P_j^+Tr_j^+ & [Tr_j^+]^{-1}P_{s,j-1}^+|_{B_j^-}\\ -[Tr_j^-]^{-1}P_{s,j-1}^-P_j^-Tr_j^+ & [Tr_j^-]^{-1}P_{s,j-1}^-|_{B_j^-}\end{pmatrix}, \end{aligned} \tag{6.14}$$

which provides a stable alternative to a computation by the last equation in (6.6).

6.4 Additional assumptions and an alternative recursion

In this subsection we assume that $\tilde q(x_1,x_2) = \tilde q(x_1)$ is independent of x_2 over each slice. The theoretical problem of the S-matrix method of Subsect. 6.2 is the use of the inverse operators $[\mathbf{T}_j^{--}]^{-1}$ and A_j, which appear in (6.6) and (6.8). **Suppose** that the image space of $\mathbf{T}_j^{--}$ is B_j^-, i.e.,

$$B_j^- \;=\; \left\{\mathbf{T}_j^{--}u_j^- :\; u_j \in B_{j-1}^- \text{ s.t. } Tr_j^- u_j \in H_\alpha^{1/2}\right\}.$$

Then there exists $[\mathbf{T}_j^{--}]^{-1}$ and, due to (6.11), $\mathbf{S}_j^{--} = [\mathbf{T}_j^{--}]^{-1}$. The operator A_j is the inverse of $\mathbf{T}_{j+1}^{--}\{I+\mathbf{S}_{j+1}^{--}\mathbf{T}_{j+1}^{-+}\mathcal{S}_j^{+-}\}$. Here $\mathbf{S}_{j+1}^{--}$ is a compact operator (cf. the arguments in the proof of Lemma 5.3) and $\{I+\mathbf{S}_{j+1}^{--}\mathbf{T}_{j+1}^{-+}\mathcal{S}_j^{+-}\}$ is a Fredholm operator of index zero. If we, additionally, **suppose** that its null space is trivial, then we arrive at $A_j = \{I+\mathbf{S}_{j+1}^{--}\mathbf{T}_{j+1}^{-+}\mathcal{S}_j^{+-}\}^{-1}[\mathbf{T}_{j+1}^{--}]^{-1}$, and the inverse A_j exists.

Finally, we look for an **alternative recursion** without additional assumptions. Replacing (6.6) by (6.14), we do not need $[\mathbf{T}_j^{--}]^{-1}$. It remains to circumvent the troubles with the inverse A_j. Recall that $\tilde q$ is independent of x_2 in all slices. The left equations in (6.6) with j replaced by $j+1$ and Equ. (6.7) imply

$$\begin{pmatrix} I & -\mathbf{S}_{j+1}^{++} & 0 & 0\\ 0 & I & -\mathcal{S}_j^{+-} & 0\\ 0 & -\mathbf{S}_{j+1}^{-+} & I & 0\\ 0 & 0 & -\mathcal{S}_j^{--} & I\end{pmatrix}\begin{pmatrix}u_{j+1}^+\\ u_j^+\\ u_j^-\\ u_{-1}^-\end{pmatrix} \;=\; \begin{pmatrix}\mathbf{S}_{j+1}^{+-}u_{j+1}^-\\ \mathcal{S}_j^{++}u_{-1}^+\\ \mathbf{S}_{j+1}^{--}u_{j+1}^-\\ \mathcal{S}_j^{-+}u_{-1}^+\end{pmatrix}. \tag{6.15}$$

Clearly, there is a solution of the Helmholtz equation over the grating for $h_0 < x_2 < h_{j+1}$ with boundary data u_{j+1}^- and u_{-1}^+ if and only if there are solutions on the gratings for $h_0<x_2<h_j$ and $h_j<x_2<h_{j+1}$ with boundary data u_j^-, u_{-1}^+ and u_{j+1}^-, u_j^+, respectively. Using S-matrices, it is natural to assume the unique solvability of (5.10) for these three gratings (cf. Thm. 5.7 and Rem.5.8). In other words, Equ. (6.7) with j replaced by $j+1$ holds if (6.15) is satisfied. The $B_j^\pm$ part of the restrictions to Γ_j of the grating solution corresponding to (6.7) are $u_j^\pm$. Vice versa, if Equ. (6.7) with j replaced by $j+1$ is satisfied and if the $u_j^\pm$ are the restrictions to Γ_j of the grating solution, then (6.15) holds. In other words, (6.15) has a unique solution, and we get

$$\begin{pmatrix} I & -\mathcal{S}_j^{+-}\\ -\mathbf{S}_{j+1}^{-+} & I\end{pmatrix}\begin{pmatrix}u_j^+\\ u_j^-\end{pmatrix} \;=\; \begin{pmatrix}\mathcal{S}_j^{++}u_{-1}^+\\ \mathbf{S}_{j+1}^{--}u_{j+1}^-\end{pmatrix}, \tag{6.16}$$

which has a unique solution too. From (6.10) we infer that $\mathbf{S}_{j+1}^{-+}$ is compact and that the matrix operator on the left-hand side is Fredholm with index zero. Consequently, the matrix in (6.16) is invertible, and its determinant operator $D_j := (I - \mathbf{S}_{j+1}^{-+}\mathcal{S}_j^{+-})$ is invertible too. We get

$$\begin{pmatrix} I & -\mathcal{S}_j^{+-} \\ -\mathbf{S}_{j+1}^{-+} & I \end{pmatrix}^{-1} = \begin{pmatrix} I + \mathcal{S}_j^{+-} D_j^{-1}\mathbf{S}_{j+1}^{-+} & \mathcal{S}_j^{+-} D_j^{-1} \\ D_j^{-1}\mathbf{S}_{j+1}^{-+} & D_j^{-1} \end{pmatrix}$$

Using this formula to solve (6.16) and (6.15), we finally obtain the recurrence relation

$$\begin{aligned} \mathcal{S}_{j+1} &= \begin{pmatrix} 0 & \mathbf{S}_{j+1}^{+-} \\ \mathcal{S}_j^{-+} & 0 \end{pmatrix} + \begin{pmatrix} \mathbf{S}_{j+1}^{++} & 0 \\ 0 & \mathcal{S}_j^{--} \end{pmatrix} \begin{pmatrix} I & -\mathcal{S}_j^{+-} \\ -\mathbf{S}_{j+1}^{-+} & I \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{S}_j^{++} & 0 \\ 0 & \mathbf{S}_{j+1}^{--} \end{pmatrix}. \\ &= \begin{pmatrix} \mathbf{S}_{j+1}^{++}\left[I + \mathcal{S}_j^{+-} D_j^{-1}\mathbf{S}_{j+1}^{-+}\right]\mathcal{S}_j^{++} & \mathbf{S}_{j+1}^{+-} + \mathbf{S}_{j+1}^{++}\mathcal{S}_j^{+-} D_j^{-1}\mathbf{S}_{j+1}^{--} \\ \mathcal{S}_j^{-+} + \mathcal{S}_j^{--} D_j^{-1}\mathbf{S}_{j+1}^{-+}\mathcal{S}_j^{++} & \mathcal{S}_j^{--} D_j^{-1}\mathbf{S}_{j+1}^{--} \end{pmatrix} \end{aligned} \qquad (6.17)$$

So the S-matrix algorithm can be used with (6.8) replaced by (6.17) and with (6.6) by (6.14). Note that, if the S-matrices are computed directly by FEM (cf. the variational formulation in (5.11)), then the S-matrix algorithm is a clever version of a non-overlapping domain decomposition method. From all these arguments we infer

Theorem 6.2. *Suppose the grating admits a slicing such that the refractive index function is independent of x_2 over each slice. Suppose, for the q defined as $q(x_1) := \tilde{q}(x_1, h_{j-1} \pm 0)$, there hold the Assumptions RC(q) introduced in Def. 4.13. In order to have well-defined S-matrices $\mathbf{S}_j$, we suppose that the boundary value problems (5.10) over the slices $\{x \in \mathbb{R}^2 : h_{j-1} < x_2 < h_j\}$ with indices $j = 1, \cdots, n$ have unique weak solutions for all right-hand sides (cf. Thm. 5.7). Finally, to have well-defined accumulated S-matrices $\mathcal{S}_j$, we suppose that the boundary value problems (5.10) over the accumulated slices $\{x \in \mathbb{R}^2 : h_0 < x_2 < h_j\}$ with indices $j = 1, \cdots, n$ have unique weak solutions for all right-hand sides (cf. Thm. 5.7). Then the recursion of the S-matrix algorithm based on (6.17) and (6.14) yields the operators $\mathcal{S}_n^{\pm+}$ and $\mathcal{S}_n^{\pm-}$ of the full grating, i.e., over the union of all slices. For given incoming waves $u_{-1}^+ \in B_{-1}^+$ and $u_n^- \in B_n^-$, the reflected and transmitted waves $u_n^+ \in B_n^+$ and $u_{-1}^- \in B_{-1}^-$ are given by $u_n^{\pm} = \mathcal{S}_n^{\pm+} u_{-1}^+ + \mathcal{S}_n^{\pm-} u_n^-$.*

In the case that $u_n^- = 0$, we **reduce the scattering matrix algorithm** to a recursion over the two matrices $\mathcal{S}_j^{\pm-}$ and the two vectors $v_j^{\pm} := \mathcal{S}_j^{\pm+} u_{-1}^+$. From the recursion (6.17), we easily obtain

$$\begin{aligned} \mathcal{S}_{j+1}^{--} &= \mathcal{S}_j^{--} D_j^{-1}\mathbf{S}_{j+1}^{--}, \\ \mathcal{S}_{j+1}^{+-} &= \mathbf{S}_{j+1}^{+-} + \mathbf{S}_{j+1}^{++}\mathcal{S}_j^{+-} D_j^{-1}\mathbf{S}_{j+1}^{--}, \\ v_{j+1}^- &= v_j^- + \mathcal{S}_j^{--} D_j^{-1}\mathbf{S}_{j+1}^{-+} v_j^+, \\ v_{j+1}^+ &= \mathbf{S}_{j+1}^{++}\left[I + \mathcal{S}_j^{+-} D_j^{-1}\mathbf{S}_{j+1}^{-+}\right] v_j^+. \end{aligned}$$

In other words, in each iteration step of index $j = 0, \cdots, n-1$ we have to perform the elementary steps:

$$\begin{array}{lrcl} i) & D_j^{-1} & = & (I-\mathbf{S}_{j-1}^{-+}\mathcal{S}_j^{+-})^{-1}, \\ ii) & E_j & := & D_j^{-1}\mathbf{S}_{j+1}^{--}, \\ iii) & G_j & := & \mathcal{S}_j^{+-}E_j, \\ iv) & \mathcal{S}_{j+1}^{--} & = & \mathcal{S}_j^{--}E_j, \\ v) & \mathcal{S}_{j+1}^{+-} & = & \mathbf{S}_{j+1}^{+-}+\mathbf{S}_{j+1}^{++}G_j, \\ vi) & w_j & := & D_j^{-1}[\mathbf{S}_{j+1}^{-+}v_j^{+}], \\ vii) & v_{j+1}^{-} & = & v_j^{-}+\mathcal{S}_j^{--}w_j, \\ viii) & v_{j+1}^{+} & = & \mathbf{S}_{j+1}^{++}v_j^{+}+\mathcal{S}_j^{+-}w_j. \end{array} \tag{6.18}$$

Starting from $v_{-1}^{\pm} := \mathbf{S}_0^{\pm+}u_{-1}^{+}$, in the last iteration step we arrive at $u_n^{+}=v_n^{+}$ and $u_{-1}^{-}=v_n^{-}$.

There remain several **open questions** to be answered by future work. For a numerical analysis the discretization must be investigated. In particular, a finite-section method reducing Fourier series expansions into finite sums must be applied to the S- and T-matrices. In this step, the possible existence of associated eigenfunctions must be taken into account. Note that the eigenvalue decomposition is not stable if eigenfunctions of rank higher than one appear. Additionally, the S-matrix method might need a modification if the underlying boundary value problems defining the S-matrices satisfy Fredholm's alternative, but are not uniquely solvable. For the FMM, the case of $\tilde{q}$ depending on x_2 must be analyzed.

References

[1] T. Abboud, Formulation variationnelle des équations de Maxwell dans un réseau bipériodique de $\mathbb{R}^3$, *C.R. Acad. Sci. Pairs,* **317** (1993), pp. 245–248.

[2] H. Ammari, N. Béreux and E. Bonnetier, Analysis of the radiation properties of a planar antenna on a photonic crystal substrate, *Math. Methods Appl. Sci.*, **24** (2001), pp. 1021–1042.

[3] A.-S. Bonnet-Ben Dhia, G. Dakhia, C. Hazard and L. Chorfi, Diffraction by a defect in an open waveguide: a mathematical analysis based on a modal radiation condition, *SIAM J. Appl. Math.*, **70** (2009), pp. 677–693.

[4] T. Arens, The scattering of plane elastic waves by a one-dimensional periodic surface, *Math. Methods Appl. Sci.*, **22** (1999), pp. 55–72.

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

[6] G. Bao, L. Cowsar and W. Masters (eds.), *Mathematical Modeling in Optical Science*, SIAM, 2001.

[7] 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** (1994), pp. 305–338.

[8] D. Dobson and A. Friedman, The time-harmonic Maxwell equations in a doubly periodic structure, *J. Math. Anal. Appl.*, **166** (1992), pp. 507–528.

[9] M.S.P. Eastham, *The spectral theory of periodic differential equations*, Scottish Academic Press, Edinburgh, 1973.

[10] J. Elschner and G. Hu, Variational approach to scattering of plane elastic waves by diffraction gratings, *Math. Methods Appl. Sci.*, **33** (2010), pp. 1924–1941.

[11] J. Elschner and G. Hu, Scattering of plane elastic waves by three-dimensional diffraction gratings, *Mathematical Models and Methods in Applied Sciences*, **22** (2012), pp. 1150019.

[12] J. Elschner and G. Schmidt, Diffraction in periodic structures and optimal design of binary gratings I. Direct problems and gradient formulas, *Math. Meth. Appl. Sci.*, **21** (1998), pp. 1297–1342.

[13] P. Joly, J.R. Li and S. Fliss, Exact boundary conditions for periodic waveguides containing a local perturbation, *Commun. Comput. Phys.*, **1** (2006), pp. 945–973.

[14] S. Fliss and P. Joly, Solutions of the time-harmonic wave equation in periodic waveguides: asymptotic behaviour and radiation condition, *Arch. Ration. Mech. Anal.*, **219** (2016), pp. 349–386.

[15] F. Gesztesy and V. Tkachenko, A Schauder and Riesz basis criterion for non-self-adjoint Schrödinger operators with periodic and antiperiodic boundary conditions, *Journal of Differential Equations*, **253** (2012), pp. 400–437.

[16] I.C. Gohberg and M.G. Krein, *Introduction to the theory of linear nonself-adjoint operators*, AMS, Providence, 1969.

[17] G. Granet and J. Chandezon, The method of curvilinear coordinates applied to the problem of scattering from surface-relief gratings defined by parametric equations: application to scattering from cycloidal grating, *Pure Appl. Opt.*, **6** (1997), pp. 727–740.

[18] J.J. Hench and Z. Strakoš, The RCWA method - A case study with open questions and perspectives of algebraic computations, *Electronic Transactions on Numerical Analysis*, **31** (2008), pp. 331–357.

[19] V. Hoang, The Limiting Absorption Principle in a semi-infinite periodic waveguide, *SIAM J. Appl. Math.*, **71** (2011), pp. 791–810.

[20] L. Hörmander, *The Analysis of Linear Partial Differential Operators III*, Springer, Berlin, 1985.

[21] G. Hu and A. Rathsfeld, Scattering of time-harmonic electromagnetic plane waves by perfectly conducting diffraction gratings, *IMA Appl. Math.*, **80** (2015), pp. 508–532.

[22] G. Hu and A. Rathsfeld, Convergence analysis of the FEM coupled with Fourier-mode expansion for the electromagnetic scattering by biperiodic structures, *Electronic Transactions on Numerical Analysis*, **41** (2014), pp. 350–375.

[23] A. Kirsch, Diffraction by periodic structures, In: *Proc. Lapland Conf. Inverse Problems*, L. Päivärinta et al, editors, (1993), Berlin, Springer, pp. 87–102.

[24] A. Kirsch and A. Lechleiter, The limiting absorption principle and a radiation condition for the scattering by a periodic layer, *SIAM J. Math. Anal.*, **50** (2018), pp. 2536–2565.

[25] A. Kirsch, Scattering by a periodic tube in $\mathbb{R}^3$: part i. The limiting absorption principle, *Inverse Problems*, **35** (2019), pp. 104004.

[26] A. Lamacz and B. Schweizer, Outgoing wave conditions in photonic crystals and transmission properties at interfaces, *ESAIM: Mathematical Modelling and Numerical Analysis*, **52** (2018), pp. 1913–1945.

[27] L. Li, Justification of matrix truncation in the modal methods of diffraction gratings, *J. Opt. A: Pure Appl. Opt.*, **1** (1999), pp. 531–536.

[28] J.W.S. Lord Rayleigh, On the dynamical theory of gratings, *Proc. Roy. Soc. Lond. A*, **79** (1907), pp. 399–416.

[29] V.A. Marchenko, *Sturm-Liouville Operators and Applications*, Birkhäuser, Basel, 1986.

[30] M.G. Moharam and T.K. Gaylord, Rigorous coupled wave analysis of planar grating diffraction, *J. Opt. Soc. Amer.*, **71** (1981), pp. 811–818.

[31] J.C. Nedelec and F. Starling, Integral equation methods in a quasi-periodic diffraction problem for the time-harmonic Maxwell's equations, *SIAM J. Math. Anal.*, **22** (1991), pp. 1679–1701.

[32] M. Nevière and E. Popov, *Light propagation in periodic media*, Marcel Dekker, Inc., New York, Basel, 2003.

[33] E. Popov, ed., *Gratings: Theory and numerical applications*, Presses universitaires de Provence (PUP), www.fresnel.fr/numerical-grating-book-2, 2012.

[34] R. Petit, *Electromagnetic theory of gratings*, Topics in Current Physics, Vol. **22**, Springer, Berlin, 1980.

[35] S.P. Shipman, *Wave propagation in periodic media: Analysis, numerical techniques and practical applications* , Bentham Science Publishers, 2010, chapter: Resonant scattering by open periodic waveguides.

[36] B. Strycharz, Uniqueness in the inverse transmission scattering problem for periodic media, *Mathematical Methods in the Applied Sciences*, **22** (1999), pp. 753–772.

[37] H.P. Urbach, Convergence of the Galerkin method for two-dimensional electromagnetic problems. *SIAM J. Numer. Anal.*, **28** (1991), pp. 697–710.

[38] O.A. Veliev and M.T. Duman, The spectral expansion for a nonself-adjoint Hill operator with a locally integrable potential, *Journal of Mathematical Analysis and Applications*, **265** (2002), pp. 76–90.