Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Convergence of the method of rigorous coupled-wave analysis for the diffraction by two-dimensional periodic surface structures

Andreas Rathsfeld

submitted: December 22, 2023

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

No. 3081
Berlin 2023

2020 *Mathematics Subject Classification.* 35P25, 74J20, 76B15, 78A45, 78A46.

Key words and phrases. Scattering problem.

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/

Convergence of the method of rigorous coupled-wave analysis for the diffraction by two-dimensional periodic surface structures

Andreas Rathsfeld

Abstract

The scattering matrix algorithm is a popular numerical method to simulate the diffraction of optical waves by periodic surfaces. The computational domain is divided into horizontal slices and, by a domain decomposition method coupling neighbour slices over the common interface via scattering data, a clever recursion is set up to compute an approximate operator, mapping incoming waves into outgoing. Combining this scattering matrix algorithm with numerical schemes inside the slices, methods like rigorous coupled wave analysis and Fourier modal methods were designed. The key for the analysis is the scattering problem over the slices. These are scattering problems with a radiation condition generalized for inhomogeneous cover and substrate materials and were first analyzed in [7]. In contrast to [7], where the scattering matrix algorithm for transverse electric polarization was treated without full discretization (no approximation by truncated Fourier series), we discuss the more challenging case of transverse magnetic polarization and look at the convergence of the fully-discretized scheme, i.e., at the rigorous coupled wave analysis for a fixed slicing into layers with vertically invariant optical index.

1 Introduction

We start with the question of what a Scattering Matrix Algorithm (SMA), a Rigorous Coupled-Wave Analysis (RCWA), and a Fourier Modal Method (FMM) is. These names are used differently by different authors. Inspired by [13, 15] and by personal taste, we stick to the following naming.

- SMA is a general iterative solver and RCWA/FMM are special realization of the SMA. To simulate scattering problems for the Helmholtz or the Maxwell's equations over periodic and biperiodic surface structures, SMA is probably the most popular algorithm in the engineering community. Its first version was described by Moharam and Gaylord [11], and good introductions with many details can be found e.g. in the books [13, 15].
- Mathematically speaking or in the language of specialists for Finite-Element Methods (FEM), SMA is a non-overlapping Domain Decomposition Method (DDM), leading to a recursive algorithm for the computation of the global solution. The iterative recursion algorithm results from the partition into the union of subdomains (slices), where each subdomain has a common boundary with at most two other subdomains. The coupling of the data over the common interface of two subdomains is realized not by equating Dirichlet, Neumann, and/or Robin data, but by equating scattering data, i.e., in- and outgoing parts of the wave.
- Discretizing this, various realizations of the SMA are possible.
  a) In principle, one could use FEM and would arrive at a special DDM for the FEM. However, we have not seen reports on this. Maybe, the reason is that splitting in in- and outgoing waves is not natural for the FEM, though this splitting relies on Dirichlet-to-Neumann operators (cf. Equ. (5.6)), which could be computed e.g. by an FEM solution of a Dirichlet problem over a small strip with perfectly matched layer to replace the radiation condition.

DOI 10.20347/WIAS.PREPRINT.3081 Berlin 2023

b) In the engineering community, the wave solution is discretized by truncated Fourier-series expansions w.r.t. the horizontal coordinates. The Fourier coefficients are functions of the vertical coordinate. This way, the numerical solution of the Boundary Value Problem (BVP) for our Partial Differential Equation (PDE) is reduced to the numerical solution of a system of Ordinary Differential Equations (ODE) w.r.t. the vertical coordinate. For the RCWA, we suppose that the PDE coefficients (wavenumber function) are equal to or, at least, approximated by coefficients, which, in each subdomain, are independent of the vertical coordinate. Hence, the matrix coefficients of the ODE are independent of the vertical coordinate, and an explicit formula of the solution based on an Eigenvalue Decomposition (EVD) can be used. For the FMM, the ODE is solved by a Finite Difference Method (FDM).

A huge number of authors contributed to the development and improvement of the RCWA and FMM and reported on their successful use. Here we only mention a few, cf. e.g. [1–3, 5, 7, 9, 10, 12, 16] and see the comparison to other methods in [8]. A first step of the analysis was provided by Hench, Strakoš [6], by Civiletti, Lakhtakia, Monk [4], and by [7]. For more comments on these, we refer to the beginning of Sect. 8. So far, to our knowledge, there is no full convergence analysis.

Of course, the most interesting version of the RCWA/FMM is that for the scattering by periodic and biperiodic surface structures modeled by the Three-Dimensional (3D) time-harmonic Maxwell's equations. However, to start the analysis, we shall restrict our consideration to the simplest case. The current paper is concerned with the Two-Dimensional (2D) Helmholtz equation.

In other words, we consider the 3D time-harmonic Maxwell's equations for the scattering by a surface around a flat plane. We suppose the surface is invariant in one of the two directions of the plane and periodic into the other. For the classical diffraction, the propagation direction of the plane wave incident to the surface is orthogonal to the direction of invariance. Then the incident wave and the resulting scattered waves are superpositions of a wave of Transverse Electric (TE) polarization and a wave of Transverse Magnetic (TM) polarization. So we can separately simulate the waves of TE or TM polarization. For these two polarizations, the time-harmonic Maxwell's equations reduces to the 2D Helmholtz equation, i.e., to $\Delta u + k^2 u = 0$ for TE and to (3.1) for TM. The scalar wave function u is a component of the electric and the magnetic field, respectively. Indeed, it is the component in the directions of invariance of the surface (cf. [14]). Most of the results will be presented for the case of TM polarization. For the easier case of TE polarization, we shall give a few hints.

Suppose the surface structure is a finite union of horizontal slices s.t. the wavenumber k is independent of the vertical point coordinates over each slice. Then the DDM can be based on a fixed finite number of subdomains, where each subdomain is such a slice with wavenumber independent of the vertical direction. In such a case, for the analysis of the method, we suggest two steps. First we consider the DDM with its SMA iteration on the continuous level, i.e., without the approximation by truncates Fourier series. As shown for the TE case in [7], the iteration leads to the true solution provided the S-matrices exists, i.e. if the problems over the subdomains are uniquely solvable (cf. Sect. 5 and Thm. 8.1). These subproblems are scattering problems but with a radiation condition for special inhomogeneous cover and substrate materials treated in [7, Thm. 5.7] and in Thm. 4.2 for TE and TM polarization, respectively. Unique solvability over the subdomains means to exclude eigenmodes (trapped modes), which may occur in exceptional cases. In the case of unique solvability, there exists a solution operator mapping the given incoming waves into the unknown outgoing waves. This is called S-matrix. Since the wavenumbers in the subdomain are independent of the vertical direction, the representation formula, which in its discretized form is the basis for the RCWA, can be used to set up the S-matrix (cf. Sect. 6).

The second step is to discretize all the operators appearing in the representation formula of the S-matrices and in the recursive SMA. The analysis of this is new even for TE polarization. Note that, roughly speaking, all the operators of the SMA on the continuous level can be expressed as infinite

matrices w.r.t. the eigenfunctions of special ODE systems. The RCWA on the discretized level is nothing else than replacing these infinite matrices by the corresponding finite matrices w.r.t. the discretized eigenfunctions, i.e., to eigenfunctions of the special ODE systems discretized by a Galerkin method based on truncated Fourier series. We get the convergence of the RCWA for the truncation index tending to infinity (cf. Thm. 8.8 and the remarks following it), showing that the operators defined by the discretized EVDs converge strongly to the operators of the continuous level and, for the inverse operators involved in the formulas, by showing that the inverse discretized operators converge strongly to the inverse. So far, we can prove this only for real-valued wavenumber functions k, where, for any x_2, the section $x_1 \mapsto k(x_1, x_2)$ must be Hölder continuous and piecewise twice continuous for the TM case and piecewise continuous and for TE polarization. We believe the convergence probably holds in much more cases. So there remain many open problems around the assumptions imposed in Sect. 8 (cf. the end of Sect. 8).

For the case of wavefunctions depending on the vertical coordinate, the wavefunction can be replaced by approximate wavefunctions, which are slice-wise constant in vertical direction. The smaller the maximal width of the slices, the closer is the approximate wavefunction to the true one. Under special non-trapping conditions, the error of this approximation was estimated in [4]. If a good wavenumber approximation is fixed, then the above mentioned analysis for a fixed slicing applies. However, a general convergence analysis for maximal width tending to zero and truncation index tending to infinity is still open. The problem of stable convergence of the SMA iteration with finer and finer slicing reminds on the stability analysis of FDM, especially if the FMM is employed. The update by the coupling over the slice interfaces reminds on implicit time steps. So there remain many open problems for a complete analysis of the SMA recursion in the RCWA.

The plane of the paper is as follows. In Sect. 2 we shall introduce the BVP for the scattering by gratings under TM polarization. To prepare the generalization of the radiation condition, in Sect. 3 we shall discuss an EVD of a one-dimensional ODE derived from the elliptic PDE. We shall present the asymptotics of eigenvalues and functions together with proofs s.t. similar asymptotics can be derived for the discretized ODE. In Sect. 4 we shall define the generalized radiation condition for special inhomogeneous cover and substrate materials and present a theorem on the unique solvability of wave scattering by periodic surfaces. For a fixed slicing of the grating structure, we shall derive the SMA in Sect. 5 on the continuous level, i.e., without any discretization in horizontal direction. The full discretization will follow in Sects. 6 – 7, where we shall give a formula to compute the solution operator over the slice with vertically constant wavenumber function and introduce the discretization by truncated Fourier series expansions. In Sect. 8, we shall present Thm. 8.1 on the SMA over the continuous level and the main result Thm. 8.8 on the convergence of the RCWA for the truncation index tending to infinity. We shall comment on the area of application and on the open problems in Sect. 9.

2 Preliminaries

We define the Two-Dimensional (2D) scattering Problem for TM polarization (cf. e.g. [14]). Here an incoming plane wave is scattered by a surface structure in $\{(x_1, y, x_2)^\top \in \mathbb{R}^3 : \ a \le x_2 \le b\}$ (cf. Fig. 1), which is periodic in x_1 direction and invariant w.r.t. shifts in y-direction. The incident plane wave is defined as $u_b^{\mathrm{inc}}(x_1, y, x_2) := e^{\mathbf{i}\vec{\alpha}\cdot(x_1,x_2)}$ with $\vec{\alpha} = (\alpha, -\sqrt{[k^+]^2 - \alpha^2}\,)^\top$, $0 < \alpha \le k^+$. Note that $|\vec{\alpha}\,| = k^+$ and $k^+ = \omega\sqrt{\mu_0\varepsilon_0}\,\mathbf{n}_+$ is the constant wavenumber for the half space with $x_2 > b$, where $\omega := 2\pi/\lambda_{\mathrm{inc}} > 0$ is the frequency of the incoming light of wavelength λ_{inc}, where ε_0 and μ_0, respectively, are the electric permittivity and the magnetic permeability in vacuum, and where $\mathbf{n}_+$ is the refractive index of the material. Similarly, there is a constant wavenumber $k^- = \omega\sqrt{\mu_0\varepsilon_0}\,\mathbf{n}_-$ for the

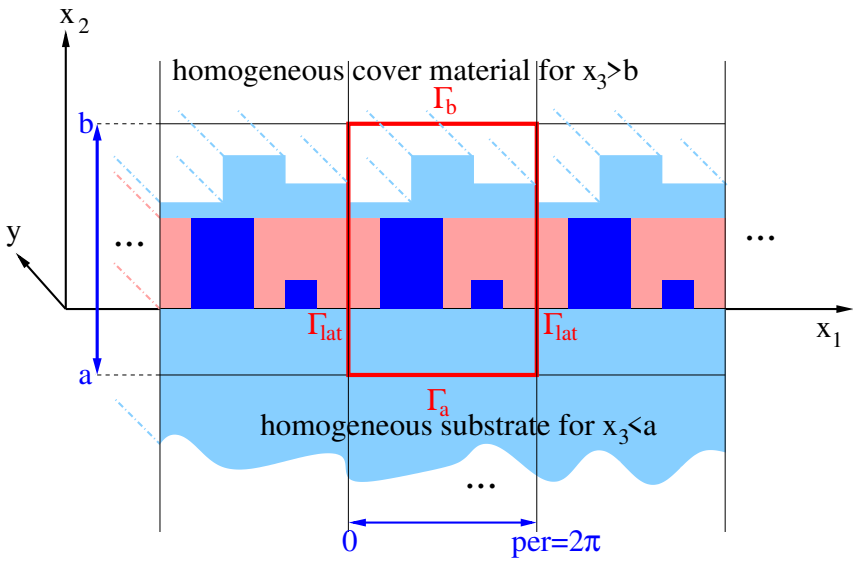

Figure 1: Geometry settings for homogeneous cover material and substrate.

half space with $x_2 < a$.

The function $u^{\rm inc}$ is α-quasiperiodic, i.e. the function $e^{-{\rm i}\alpha x_1}u^{\rm inc}(x_1,x_2)$ is 2π-periodic, and we get $u_b^{\rm inc}(x_1+2\pi,x_2) = e^{{\rm i}2\pi\alpha}u_b^{\rm inc}(x_1,x_2)$. Consequently, all the waves and their boundary values on $\Gamma_c := \{(x_1,c)\colon 0\le x_1\le 2\pi\}$, $c=a,b$ are in the 2D Sobolev spaces $H^s_\alpha(\Omega)$ and 1D Sobolev spaces $H^1_\alpha(\Gamma_c) = H^s_\alpha(0,2\pi)$, respectively, i.e. in spaces of α-quasiperiodic H^s-functions. We can even admit a general incident field $u^{\rm inc}(x_1,x_2)$ for $x_2 \ge b$ (cf. the subsequent (2.3)) if only the restriction $u^{\rm inc}|_{\Gamma_b}$ is α-quasiperiodic. Clearly, we can change the α in the definition of quasiperiodicity by subtracting an integer, i.e. we can assume w.l.o.g. that $0 \le \alpha < 1$. Besides the wave incoming from above, we even can admit an incoming wave $u_b^{\rm inc}(x_1,x_2)$ from below, i.e. from $x_2 \le a$. However, we have to assume that the restriction $u_a^{\rm inc}|_{\Gamma_a}$ is α-quasiperiodic with the same α.

In the case of TM polarization we look for the y-component of the magnetic field $u(x_1,y,x_2)$, which is independent of y. So the 3D Helmholtz equations turns into the 2D equation for the function $u(x_1,x_2)=u(x_1,y,,x_2)$. Altogether, the wave u is the solution over the domain $\Omega := [0,2\pi]\times[a,b]$ satisfying

$$\begin{aligned}
&a) \quad \text{“Helmholtz” equation: } \nabla\cdot k(x_1,x_2)^{-2}\nabla u(x_1,x_2)+u(x_1,x_2)=0,\ (x_1,x_2)^\top\in\Omega,\\
&b) \quad \alpha\text{-quasiperiodic lateral boundary condition: } u(2\pi,x_2)=e^{{\rm i}2\pi\alpha}u(0,x_2),\ x_2\in[a,b],\\
&c) \quad \text{Radiation condition over upper boundary } \Gamma_b \text{ and lower boundary } \Gamma_a \text{ incl. given traces}\\
&\qquad \text{of incident wave functions } u_b^{\rm inc}|_{\Gamma_b}\in H^{1/2}_\alpha(\Gamma_b) \text{ and } u_a^{\rm inc}|_{\Gamma_a}\in H^{1/2}_\alpha(\Gamma_a), \text{ respectively.}
\end{aligned} \tag{2.1}$$

Note that, for the case of TE polarization, in item a) the classical form $\Delta u + k^2 u = 0$ of the Helmholtz equation appears, which is equivalent to the equation in a) for constant wavenumbers k. In particular, this is the case for homogeneous materials in the substrate or the cover material, and we get the same radiation condition. For this radiation condition, we remark that the general representation of α-quasiperiodic Helmholtz solutions in the homogeneous cover material is

$$\begin{aligned}
u(x_1,x_2) - u_b^{\rm inc}(x_1,x_2) \;&=\; \sum_{l\in\mathbb{Z}} e^{{\rm i}(\alpha+l)x_1}\left\{c^+_{b,l}e^{{\rm i}\beta^b_l(x_2-b)} + c^-_{b,l}e^{-{\rm i}\beta^b_l(x_2-b)}\right\}, \quad x_2\ge b\,, \\
&\beta^b_l := \sqrt{[k^+]^2-[\alpha+l]^2}, \quad c^\pm_{b,l}\in\mathbb{C}.
\end{aligned} \tag{2.2}$$

The radiation condition on Γ_b requires $c^-_{b,l}=0$ for all coefficients of downgoing modes, i.e.

$$u(x_1,x_2)-u_b^{\mathrm{inc}}(x_1,x_2) \;=\; \sum_{l\in\mathbb{Z}} c^+_{b,l} e^{\mathbf{i}(\alpha+l)x_1} e^{\mathbf{i}\beta_l^b(x_2-b)}, \quad x_2\geq b. \tag{2.3}$$

Here, for simplicity, we have supposed $\beta_l^b\neq 0$. For the l with $\beta_l^b=0$, in Equ. (2.2) the term in brackets $\{c^+_{b,l}e^{\mathbf{i}\beta_l^b(x_2-b)}+c^-_{b,l}e^{-\mathbf{i}\beta_l^b(x_2-b)}\}$ must be modified. Depending on the application, it should be replaced by $\{c^+_{b,l}+c^-_{b,l}(x_2-b)\}$, by $\{c^+_{b,l}(x_2-b)+c^-_{b,l}\}$ or by $\{c^+_{b,l}\big(1+(x_2-b)\big)+c^-_{b,l}\big(1-(x_2-b)\big)\}$, respectively. This leads to a corresponding modification in (2.3). The radiation condition on Γ_a requires $c^+_{a,l}=0$ for all coefficients of upgoing modes, i.e.

$$u(x_1,x_2)-u_a^{\mathrm{inc}}(x_1,x_2)=\sum_{l\in\mathbb{Z}} c^-_{a,l}e^{\mathbf{i}(\alpha+l)x_1}e^{-\mathbf{i}\beta_l^a(x_2-a)},\ x_2\leq a\,,\ \beta_l^a:=\sqrt{[k^-]^2-[\alpha+l]^2}. \tag{2.4}$$

Again a corresponding modification is needed if $\beta_l^a=0$.

3 Eigenfunctions of ODE for reformulation of Helmholtz equation

For the scattering matrix algorithm, we have to generalize the radiation conditions (2.3) and (2.4) modeling inhomogeneous cover and substrate materials. In order to prepare this, we need the EVD of ordinary differential operators appearing in the reformulation of the Helmholtz equation as an ODE with operator valued coefficient function. The details will be needed also for the corresponding equations obtained by discretization.

Suppose the 2π periodic wavenumber function k is given as $k:=\omega\sqrt{\mu_0\varepsilon_0}\,\mathbf{n}$, where the refractive index $\mathbf{n}$, possibly depending on x_1 and x_2, is supposed to satisfy $\Re e\,\mathbf{n}>0$, $\Im m\,\mathbf{n}\geq 0$. With this k the TM wave equation is

$$\nabla\cdot k^{-2}\nabla u+u \;=\; 0, \quad \text{in } \mathbb{R}^2. \tag{3.1}$$

For piecewise constant k, this is nothing else than the Helmholtz equation with special transmission conditions over the curves of discontinuity for k. Now in the cover material and substrate (cf. Fig. 2), we assume $k(x_1,x_2)=k_+(x_1)$, $x_2>b$ and $k(x_1,x_2)=k_-(x_1)$, $x_2<a$. Equ. (3.1) is equivalent to the operator valued ODE

$$k^{-2}\partial_{x_2}^2 u=Lu:=-\partial_{x_2}k^{-2}\partial_{x_2}u-u, \quad k(x_1):=k_\pm(x_1). \tag{3.2}$$

We reduce this second-order ODE to a first-order ODE. Setting $v:=\partial_{x_2}u$ and $\vec{w}:=(u,v)^\top$, the ODE (3.2) is equivalent to $\partial_{x_2}\vec{u}=M\vec{u}$ with

$$M:=\begin{pmatrix} 0 & I\\ k^2L & 0\end{pmatrix}. \tag{3.3}$$

For this operator in space of univariate vector functions depending on x_1, the eigenvalues and eigenfunctions are defined by $M\vec{f_\lambda}=\lambda\vec{f_\lambda}$ for $\lambda\in\sigma_M$. Clearly, for $\vec{f_\lambda}=(f_\lambda,g_\lambda)^\top$, we get $g_\lambda=\lambda f_\lambda$ and $k^2Lf_\lambda=\lambda g_\lambda$. Consequently, for the eigenvalues $\pm\lambda$ of M, we obtain the eigenvector $(f_\lambda,\pm\lambda f_\lambda)^\top$ with f_λ satisfying

$$k^2\partial_{x_1}k^{-2}\partial_{x_1}f_\lambda+[k^2+\lambda^2]f_\lambda \;=\; 0. \tag{3.4}$$

As a first case, we discuss the EVD in (3.4) with a $k(x_1)$ **twice continuously differentiable**. We look for a solution f of (3.4) in the form $f = kh$.

$$\begin{aligned} k^2\partial_{x_1}k^{-2}\partial_{x_1}[kh] + [k^2+\lambda^2][kh] &= 0, \\ k\partial_{x_1}^2 h + \left\{[\partial_{x_1}^2 k] - 2k^{-1}[\partial_{x_1}k]^2 + [k^2+\lambda^2]k\right\} h &= 0, \\ \partial_{x_1}^2 h + \tilde{k}^2 h + \lambda^2 h &= 0, \\ \tilde{k}^2 &:= k^2 + k^{-1}[\partial_{x_1}^2 k] - 2k^{-2}[\partial_{x_1}k]^2. \end{aligned} \tag{3.5}$$

For this $f = kh$, we note that in the derivation of the variational form we have

$$\begin{aligned} \int_\Omega \{\nabla\cdot k^{-2}\nabla u\bar{v} + u\bar{v}\} &= \int_\Omega \left\{-k^{-2}\nabla u\overline{\nabla v} + u\bar{v}\right\} + \int_{\Gamma_a} k^{-2}\partial_{x_2}u\bar{v} + \int_{\Gamma_b} k^{-2}\partial_{x_2}u\bar{v}, \\ \int_0^{2\pi} k^{-2}\partial_{x_2}[f(x_1)e^{-\lambda(x_2-c)}]|_{x_2=c}\overline{[f(x_1)e^{-\lambda(x_2-c)}]|_{x_2=c}}\mathrm{d}x_1 &= \int_0^{2\pi}\lambda|h(x_1)|^2\mathrm{d}x_1, \end{aligned}$$

whereas, for the Helmholtz equation in the TE case,

$$\begin{aligned} \int_\Omega \{\Delta u\bar{v} + k^2 u\bar{v}\} &= \int_\Omega \left\{\nabla u\overline{\nabla v} + k^2 u\bar{v}\right\} + \int_{\Gamma_a} \partial_{x_2}u\bar{v} + \int_{\Gamma_b} \partial_{x_2}u\bar{v}, \\ \int_0^{2\pi} \partial_{x_2}[f(x_1)e^{-\lambda(x_2-c)}]\overline{[f(x_1)e^{-\lambda(x_2-c)}]|_{x_2=c}}\mathrm{d}x_1 &= \int_0^{2\pi}\lambda|f(x_1)|^2\mathrm{d}x_1. \end{aligned}$$

In other words, we get similar formulas for the PDE $\nabla\cdot k^{-2}u + u = 0$ and h as for $\Delta u + k^2 u = 0$ and f. In any case, we can use the results collected in [7, Lemma 4.5].

Lemma 3.1. *The spectrum is discrete, i.e., there holds* $\sigma_{[\partial_{x_1}^2+\tilde{k}^2 I]} = \{\lambda_n^2 : n\in\mathbb{Z}\}$. *We even get the asymptotics*

$$\begin{aligned} \lambda_n^2 &= (n+\alpha)^2 - \tilde{k}_{\mathrm{avg}}^2 + \mathcal{O}\left(\frac{1}{|n|^\kappa}\right),\ \tilde{k}_{\mathrm{avg}}^2 := \frac{1}{2\pi}\int_0^{2\pi}\tilde{k}^2(\tau)\mathrm{d}\tau,\ \kappa := \begin{cases} 1/2 & \text{if } \alpha = 0, 1/2 \\ 1 & \text{else} \end{cases}, \\ \lambda_n &= \sqrt{(n+\alpha)^2 - \tilde{k}_{\mathrm{avg}}^2} + \mathcal{O}\left(\frac{1}{|n|^{1+\kappa}}\right) = |n+\alpha| + \frac{\tilde{k}_{\mathrm{avg}}^2}{2}\frac{1}{|n|} + \mathcal{O}\left(\frac{1}{|n|^{1+\kappa}}\right),\ |n|\to\infty. \end{aligned} \tag{3.6}$$

Proof. To derive an asymptotics like this, we can argue as follows. The solution of the inhomogeneous ODE $\partial_{x_1}^2 h + \lambda^2 h = -\tilde{k}^2 h$ is

$$\begin{aligned} u(t) &= u(0)\cos(\lambda t) + \frac{u'(0)}{\lambda}\sin(\lambda t) - \frac{1}{\lambda}\int_0^t \sin\big(\lambda(t-\tau)\big)\tilde{k}^2(\tau)u(\tau)\mathrm{d}\tau, \\ u'(t) &= -u(0)\lambda\cos(\lambda t) + u'(0)\sin(\lambda t) - \int_0^t \cos\big(\lambda(t-\tau)\big)\tilde{k}^2(\tau)u(\tau)\mathrm{d}\tau. \end{aligned}$$

In other words,

$$\begin{aligned} \begin{pmatrix} u(t) \\ u'(t)/\lambda \end{pmatrix} &= \mathcal{R}_{\lambda t}\begin{pmatrix} u(0) \\ u'(0)/\lambda \end{pmatrix} - \frac{1}{\lambda}\int_0^t \begin{pmatrix} \sin\big(\lambda(t-\tau)\big) \\ \cos\big(\lambda(t-\tau)\big) \end{pmatrix}\tilde{k}^2(\tau)u(\tau)\mathrm{d}\tau, \\ \mathcal{R}_\phi &:= \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix}. \end{aligned} \tag{3.7}$$

Using this at $t = 2\pi$ and the α-quasiperiodicity $u^{(j)}(2\pi) = e^{\mathbf{i}2\pi\alpha}u(j)(0)$, $j = 0, 1$, we arrive at

$$\left[e^{\mathbf{i}2\pi\alpha}I - \mathcal{R}_{2\pi\lambda}\right]\begin{pmatrix} u(0) \\ u'(0)/\lambda \end{pmatrix} = -\frac{1}{\lambda}\int_0^{2\pi}\begin{pmatrix} \sin\big(\lambda(2\pi-\tau)\big) \\ \cos\big(\lambda(2\pi-\tau)\big) \end{pmatrix}\tilde{k}^2(\tau)u(\tau)\mathrm{d}\tau.$$

Solving this equation w.r.t. $(u(0), u'(0)/\lambda)^\top$ and substituting the result into (3.7), we obtain the equation $(u, u/\lambda)^\top = \frac{1}{\lambda}\mathcal{T}(u, u/\lambda)^\top$ with

$$\begin{aligned}\mathcal{T}\begin{pmatrix} u(t) \\ u'(t)/\lambda \end{pmatrix} &:= -\mathcal{R}_{\lambda t}\left[e^{\mathbf{i}2\pi\alpha} I - \mathcal{R}_{2\pi\lambda}\right]^{-1} \int_0^{2\pi} \begin{pmatrix} \sin\big(\lambda(2\pi-\tau)\big) \\ \cos\big(\lambda(2\pi-\tau)\big) \end{pmatrix} \tilde{k}^2(\tau) u(\tau) \mathrm{d}\tau \\ &\quad - \int_0^t \begin{pmatrix} \sin\big(\lambda(t-\tau)\big) \\ \cos\big(\lambda(t-\tau)\big) \end{pmatrix} \tilde{k}^2(\tau) u(\tau) \mathrm{d}\tau.\end{aligned}$$

We conclude

$$\begin{aligned} 1 \;\le\; \frac{1}{\lambda}\|\mathcal{T}\| &\le C\frac{1}{\lambda}\left|\det(e^{\mathbf{i}2\pi\alpha} I - \mathcal{R}_{2\pi\lambda})^{-1}\right|, \\ \left|\big(e^{\mathbf{i}2\pi\alpha} - \cos(2\pi\lambda)\big)^2 + \sin(2\pi\lambda)^2\right| &\le C\frac{1}{\lambda}, \\ \left|\cos\big(\pi(\alpha+\lambda)\big)\cos\big(\pi(\alpha-\lambda)\big)\right| &\le \frac{1}{4}C\frac{1}{\lambda}. \end{aligned} \tag{3.8}$$

So either $\cos(\pi(\lambda+\alpha)) = \mathcal{O}(|\lambda|^{-1/2})$ or $\cos(\pi(\lambda-\alpha)) = \mathcal{O}(|\lambda|^{-1/2})$. Since we have the identity $\cos(z) = \cos(\Re e\, z + \mathbf{i}\Im m\, z) = \cos(\Re e\, z)\cosh(\Im m\, z) - \mathbf{i}\sin(\Re e\, z)\sinh(\Im m\, z)$, we get the estimate $|\cos(\Re e\, z)| \le |\cos(z)|$. So either $\Re e\,\lambda = n - \alpha + \mathcal{O}(|\lambda|^{-1/2})$ or $\Re e\,\lambda = n + \alpha + \mathcal{O}(|\lambda|^{-1/2})$ for a suitable integer n. A small $\cos(\Re e\, z)$ means a $\sin(\Re e\, z)$ close to one s.t. we get the estimate $|\sinh(\Im m\, z)| < c|\cos(z)|$. So in any case $\Im m\,\lambda = \mathcal{O}(|\lambda|^{-1/2})$. Altogether we get the asymptotics $\lambda = n \pm \alpha + \mathcal{O}(|n|^{-1/2})$. Now assume $\alpha \neq 0, 1/2$. E.g. for $\lambda = n - \alpha + \mathcal{O}(|n|^{-1/2})$, we get the first cosine value in (3.8) as $\cos(\pi(\lambda+\alpha)) = \cos(\mathcal{O}(|n|^{-1/2}) = 1 - \mathcal{O}(|n|^{-1}) > 1/2$. With this, however, (3.8) implies $\cos(\pi(\lambda-\alpha)) = \mathcal{O}(|\lambda|^{-1})$. In other words, the above arguments lead us to the improved asymptotics $\lambda = n - \alpha + \mathcal{O}(|n|^{-1})$. Though this is less than (3.6), such an estimate for the discretized equation together with the stronger (3.6) is sufficient for our analysis. □

Now we look at the asymptotics of the eigenfunctions and discuss the basis property as well as the asymptotics of the eigenfunctions. Denoting $h_n := h_{\lambda_n}$. The asymptotics of the eigenfunctions has been mentioned in [7, Lemma 4.5] and we learn from this paper that, at least for $\alpha \neq 0, 1/2$ or for real-valued k, the functions $(1+n^2)^{s/2} h_n \in H^s_\alpha,\ n \in \mathbb{Z}$ form a Riesz basis for $-2 \le s \le 2$. Hence, we also have a Riesz basis $(1+n^2)^{s/2} f_n = (1+n^2)^{s/2} k h_n \in H^s_\alpha,\ n \in \mathbb{Z}$ for $-2 \le s \le 2$. We shall suppose throughout this paper that all eigenfunctions are of rank one. If rank-greater-than-one eigenfunction occur, then the subsequent algorithm must be adapted to that case, and, in the case of an infinite number of such eigenfunctions, the Riesz property of the basis might be violated. Note, however, that this is not a problem for real-valued $\tilde{k}$, since the eigenfunctions of selfadjoint operators all have rank one.

Lemma 3.2. *Assume the asymptotics $|\lambda_n| \to \infty$ for $|n| \to \infty$. Then we arrive at*

$$\begin{aligned} h_n(t) &\sim e^{\mathbf{i}\lambda_n t} - \frac{1}{\lambda_n}\frac{\sin(\lambda_n t)}{1 - e^{\mathbf{i}2\pi[\lambda_n+\alpha]}} \int_0^{2\pi} e^{\mathbf{i}2\lambda_n\tau} \tilde{k}^2(\tau)\mathrm{d}\tau \\ &\quad -\frac{1}{\lambda_n}\int_0^t \sin(\lambda_n[t-\tau])\tilde{k}^2(\tau) e^{\mathbf{i}\lambda_n\tau}\mathrm{d}\tau + \mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right) \\ &\sim e^{\mathbf{i}\lambda_n t} - \frac{1}{\lambda_n}\frac{e^{\mathbf{i}\lambda_n t}}{2\mathbf{i}} \int_0^t \tilde{k}^2(\tau)\mathrm{d}\tau + \mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right). \end{aligned}$$

Proof. To derive the asymptotics, we look for h in the representation $h(t) = e^{\mathbf{i}\lambda t}\eta(t)$, where $e^{\mathbf{i}\lambda t}$ is the eigenfunction for constant wavenumber functions $\tilde{k}$. To get an α-quasiperiodic h, the function η

must be $[\alpha-\lambda]$-quasiperiodic. We get

$$\begin{aligned} e^{\mathbf{i}\lambda t}\eta''(t)+2\mathbf{i}\lambda e^{\mathbf{i}\lambda t}\eta'(t)+[\mathbf{i}\lambda]^2 e^{\mathbf{i}\lambda t}\eta(t)+[\tilde{k}^2(t)+\lambda^2]e^{\mathbf{i}\lambda t}\eta(t) &= 0,\\ \eta''(t)+2\mathbf{i}\lambda\eta'(t)+\tilde{k}^2(t)\eta(t) &= 0. \end{aligned} \tag{3.9}$$

From (3.9), we get

$$\begin{aligned} [e^{\mathbf{i}2\lambda t}\eta'(t)]' &= e^{\mathbf{i}2\lambda t}[\eta''(t)+2\mathbf{i}\lambda\eta'(t)] = -e^{\mathbf{i}2\lambda t}\tilde{k}^2(t)\eta(t),\\ e^{\mathbf{i}2\lambda t}\eta'(t) &= \eta'(0)-\int_0^t e^{\mathbf{i}2\lambda\tau}\tilde{k}^2(\tau)\eta(\tau)\mathrm{d}\tau,\\ \eta'(t) &= \eta'(0)e^{-\mathbf{i}2\lambda t}-\int_0^t e^{-\mathbf{i}2\lambda[t-\tau]}\tilde{k}^2(\tau)\eta(\tau)\mathrm{d}\tau, \qquad (3.10)\\ \eta(t) &= \eta(0)-\frac{\mathbf{i}}{2\lambda}\left[1-e^{-\mathbf{i}2\lambda t}\right]\eta'(0)-\frac{\mathbf{i}}{2\lambda}\int_0^t\left[e^{-\mathbf{i}2\lambda[t-\tau]}-1\right]\tilde{k}^2(\tau)\eta(\tau)\mathrm{d}\tau. \qquad (3.11) \end{aligned}$$

W.l.o.g. we may set $\eta(0)=1$ or $\eta'(0)=1/\lambda$. Otherwise we would scale the function η. Set $\eta(0)=1$. From (3.10) and the $[\alpha-\lambda]$-quasiperiodicity we conclude

$$\begin{aligned} e^{-\mathbf{i}2\pi[\lambda-\alpha]}\eta'(0) &= \eta'(2\pi)=\eta'(0)e^{-\mathbf{i}4\pi\lambda}-\int_0^{2\pi}e^{-\mathbf{i}2\lambda[2\pi-\tau]}\tilde{k}^2(\tau)\eta(\tau)\mathrm{d}\tau,\\ \eta'(0) &= \frac{e^{\mathbf{i}4\pi\lambda}}{1-e^{\mathbf{i}2\pi[\lambda+\alpha]}}\int_0^{2\pi}e^{-\mathbf{i}2\lambda[2\pi-\tau]}\tilde{k}^2(\tau)\eta(\tau)\mathrm{d}\tau. \end{aligned}$$

Using the asymptotics of the eigenvalues $\lambda=\lambda_n$, $n\in\mathbb{Z}$ in (3.6)

$$\begin{aligned} e^{\mathbf{i}2\pi[\lambda_n+\alpha]}-1 &= \left[e^{\mathbf{i}4\pi\alpha}-1\right]-\pi\mathbf{i}e^{\mathbf{i}4\pi\alpha}\tilde{k}_{\mathrm{avg}}^2\frac{1}{n}+\mathcal{O}\left(\frac{1}{n^{1+\kappa}}\right),\\ \frac{1}{1-e^{\mathbf{i}2\pi[\lambda_n+\alpha]}} &= \begin{cases} \frac{1}{1-e^{\mathbf{i}4\pi\alpha}}-\pi\mathbf{i}\frac{e^{\mathbf{i}4\pi\alpha}}{[1-e^{\mathbf{i}4\pi\alpha}]^2}\tilde{k}_{\mathrm{avg}}^2\frac{1}{n}+\mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right) & \text{if } \alpha\neq 0,1/2\\ \frac{\lambda_n}{\pi\mathbf{i}\tilde{k}_{\mathrm{avg}}^2}+\mathcal{O}\left(n^{1/2}\right) & \text{if } \alpha=0,1/2 \end{cases} \end{aligned}$$

To get a bounded fraction on the left-hand side, we suppose that $\alpha\neq 0,1/2$. Recalling $h(t)=e^{\mathbf{i}\lambda t}\eta(t)$, Equ. (3.11) turns into

$$\begin{aligned} (I-T_n)h_n &= h_{n,0},\quad h_n = \sum_{l=0}^{\infty}h_{n,l},\quad h_{n,l} = T_n h_{n,l-1},\quad l\in\mathbb{Z}_+,\\ h_n &= h_{n,0}+T_n h_{n,0}+T_n^2 h_n. \end{aligned}$$

Here we have set

$$\begin{aligned} h_{n,0}(t) &:= e^{\mathbf{i}\lambda_n t},\quad 0\le t\le 2\pi,\\ T_n f(t) &:= e^{\mathbf{i}\lambda_n t}\frac{\mathbf{i}}{2\lambda_n}\left[1-e^{-\mathbf{i}2\lambda_n t}\right]\frac{e^{\mathbf{i}4\pi\lambda_n}}{1-e^{\mathbf{i}2\pi[\lambda_n+\alpha]}}\int_0^{2\pi}e^{-\mathbf{i}2\lambda_n[2\pi-\tau]}\tilde{k}^2(\tau)e^{-\mathbf{i}\lambda_n\tau}f(\tau)\mathrm{d}\tau\\ &\quad -e^{\mathbf{i}\lambda_n t}\frac{\mathbf{i}}{2\lambda_n}\int_0^t\left[e^{-\mathbf{i}2\lambda_n[t-\tau]}-1\right]\tilde{k}^2(\tau)e^{-\mathbf{i}\lambda_n\tau}f(\tau)\mathrm{d}\tau\\ &= -\frac{1}{\lambda_n}\frac{\sin(\lambda_n t)}{1-e^{\mathbf{i}2\pi[\lambda_n+\alpha]}}\int_0^{2\pi}e^{\mathbf{i}\lambda_n\tau}\tilde{k}^2(\tau)f(\tau)\mathrm{d}\tau-\frac{1}{\lambda_n}\int_0^t\sin(\lambda_n[t-\tau])\tilde{k}^2(\tau)f(\tau)\mathrm{d}\tau. \end{aligned}$$

Consequently,

$$\begin{aligned} h_n(t) &\sim e^{\mathbf{i}\lambda_n t} - \frac{1}{\lambda_n}\frac{\sin(\lambda_n t)}{1-e^{\mathbf{i}2\pi[\lambda_n+\alpha]}}\int_0^{2\pi} e^{\mathbf{i}2\lambda_n\tau}\tilde{k}^2(\tau)\mathrm{d}\tau \\ &\quad -\frac{1}{\lambda_n}\int_0^t \sin(\lambda_n[t-\tau])\tilde{k}^2(\tau)e^{\mathbf{i}\lambda_n\tau}\mathrm{d}\tau + \mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right) \\ &\sim e^{\mathbf{i}\lambda_n t} - \frac{1}{\lambda_n}\frac{\sin(\lambda_n t)}{1-e^{\mathbf{i}4\pi\alpha}}\int_0^{2\pi} e^{\mathbf{i}2\lambda_n\tau}\tilde{k}^2(\tau)\mathrm{d}\tau - \frac{1}{\lambda_n}\frac{e^{\mathbf{i}\lambda_n t}}{2\mathbf{i}}\int_0^t \tilde{k}^2(\tau)\mathrm{d}\tau \\ &\quad +\frac{1}{\lambda_n}\frac{e^{-\mathbf{i}\lambda_n t}}{2\mathbf{i}}\int_0^t e^{\mathbf{i}2\lambda_n\tau}\tilde{k}^2(\tau)\mathrm{d}\tau + \mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right). \end{aligned}$$

Since $\tilde{k}^2$ is continuously differentiable, applying partial integration to the first and third integral on the right-hand side, we arrive at

$$h_n(t) \sim e^{\mathbf{i}\lambda_n t} - \frac{1}{\lambda_n}\frac{e^{\mathbf{i}\lambda_n t}}{2\mathbf{i}}\int_0^t \tilde{k}^2(\tau)\mathrm{d}\tau + \mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right).$$

□

Now suppose the wave number $k(x_1)$ is only **piecewise twice continuously differentiable** and $k(x_1) \geq c_k > 0$. Then we get

Lemma 3.3. *The spectrum of operator k^2L is a discrete set of eigenvalues $\sigma_{k^2L} = \{\lambda_n^2 : n \in \mathbb{Z}_+\} \subset \mathbb{R}$ with $\lambda_n^2 \to \infty$, $|n| \to \infty$. The eigenfunction f_n corresponding to λ_n is of rank one. It is piecewise twice continuously differentiable and continuous, and $k^{-2}\partial_{x_1} f_n$ is continuous as well. Moreover, the scaled eigenfunctions $(1+|\lambda_n|^2)^{-s/2} f_n$, $n \in \mathbb{Z}_+$ of the differential operator k^2L form a Riesz basis in H^s_α for $-1 \leq s \leq 1$.*

Proof. Clearly, the eigenfunctions f_λ of the unbounded operator $k^2L\colon L^2(0,2\pi) \hookrightarrow L^2(0,2\pi)$ are in one-to-one correspondence with the eigenfunctions $k^{-1}f_\lambda$ of the operator $\tilde{L} := kLk$, which maps $k^{-1}H^1_\alpha$ into kH^1_α and the domain of definition of which is

$$\mathrm{dom}_{\tilde{L}} := \left\{h \in k^{-1}H^1_\alpha \colon k^{-2}\partial_{x_1} kh \in H^1_\alpha\right\}.$$

This is the operator of the variational form $(h,g) \mapsto a(h,g)$ with

$$a(h,g) := \int_0^{2\pi} \left\{k^{-2}(x_1)\,\partial_{x_1}[k(x_1)h(x_1)]\,\overline{\partial_{x_1}[k(x_1)g(x_1)]} - k^2(x_1)h(x_1)\overline{g(x_1)}\right\}\mathrm{d}x_1. \quad (3.12)$$

In other words, $\tilde{L}$ is selfadjoint and strongly elliptic, and its spectrum is a discrete set of real eigenvalues with the only cluster point ∞. We denote the eigenvalues by λ_n^2, $n \in \mathbb{Z}_+$ and the corresponding orthonormal eigenfunctions by h_n s.t. the $f_n := kh_n$ form a Riesz basis of eigenfunctions for operator k^2L in the space $L^2(0,2\pi)$. For a general function $h = \sum_{n\in\mathbb{Z}_+}\xi_n h_n$ with $\xi_n \in \mathbb{C}$, we obtain

$$\begin{aligned} \left\|k\sum_{n=1}^\infty \xi_n h_n\right\|^2_{H^1_\alpha} &\sim a(h,h) + c\,\|h\|^2_{L^2} = \sum_{n,m=1}^\infty \xi_n\overline{\xi_m}\left\langle \tilde{L}h_n, h_m\right\rangle + c\,\|h\|^2_{L^2}, \\ \left\|\sum_{n=1}^\infty \xi_n f_n\right\|^2_{H^1_\alpha} &\sim \sum_{n,m}\lambda_n^2\xi_n\overline{\xi_m}\,\langle h_n, h_m\rangle + c\,\|h\|^2_{L^2} \sim c\sum_n(1+|\lambda_n|^2)|\xi_n|^2. \end{aligned}$$

This is the Riesz basis property of the basis $(1+|\lambda_n|^{-2})^{1/2} f_n$ in H^1_α. The orthonormality of the h_n yields the Riesz basis property of the basis f_n in $L^2 = H^0_\alpha$. Interpolation provides us with the Riesz basis property of the basis $(1+|\lambda_n|^{-2})^{s/2} f_n$ in H^s_α for $0 \le s \le 1$. By duality arguments for the orthonormal basis, we get the Riesz property in H^s_α for $-1 \le s \le 1$.

Due to $k > c_k$, the operator kLk is selfadjoint and the ranks of the eigenfunctions are automatically equal to one. On each segment in $[0, 2\pi]$ where k is twice continuously differentiable, the equation $h_n'' + \tilde{k}^2 h_n + \lambda_n^2 h_n = 0$ holds (cf. (3.5)). Consequently, the solution is twice continuously differentiable over the closed interval. On the other hand, the equation $k^2 L f_n = \lambda_n f_n$ over the whole quasiperiodic interval implies that the global derivative of f_n is piecewise smooth. No Dirac delta should appear. Thus f_n is continuous and its derivative coincides with the piecewise derivative f_n'. Similarly, the derivative of the continuous quasiperiodic $[k^{-2} f_n']$ is piecewise smooth. Thus $[k^{-2} f_n']$ is continuous and its derivative coincides with the piecewise derivative $\partial_{x_1}[k^{-2} f_n']$. □

General case: In agreement with the last two cases, we suppose that there is a system of univariate eigenfunctions f_n, $n \in \mathbb{Z}_+$ of rank one with the corresponding eigenvalues λ_n such that $k^2 L f_n = \lambda_n^2 f_n$, such that $|\Im m\, \lambda_n| \le C |\Re e\, \lambda_n|$ for a fixed positive constant C, and such that the $(1+|\lambda_n|^2)^{-s/2} f_n$ form a Riesz basis in H^s_α for $-1 \le s \le 1$. Note that we change the index set from $\mathbb{Z}$ to $\mathbb{Z}_+$ such that $|\lambda_n| \le |\lambda_{n+1}|$ for all $n \in \mathbb{Z}_+$. Here, for the squareroots λ_n of the $\lambda_n^2 \ne 0$, we suppose that either $\Re e\, \lambda_n > 0$ or $\Re e\, \lambda_n = 0, \Im m\, \lambda_n < 0$. Note that $\pm\lambda_n$ are eigenvalues of M in (3.3). Depending on the choice of wavenumber function k as $k(x_1) = k^+(x_1) = k(x_1, b+0)$ or as $k(x_1) = k^-(x_1) = k(x_1, a-0)$, we write L_a or L_b for the differential operator L, $f_{a,n}$ or $f_{b,n}$ for the eigenfunction f_n, and $\lambda_{a,n}$ or $\lambda_{b,n}$ for the eigenvalue λ_n.

4 Radiation condition and unique solvability of scattering problem for special inhomogeneous super- and substrate

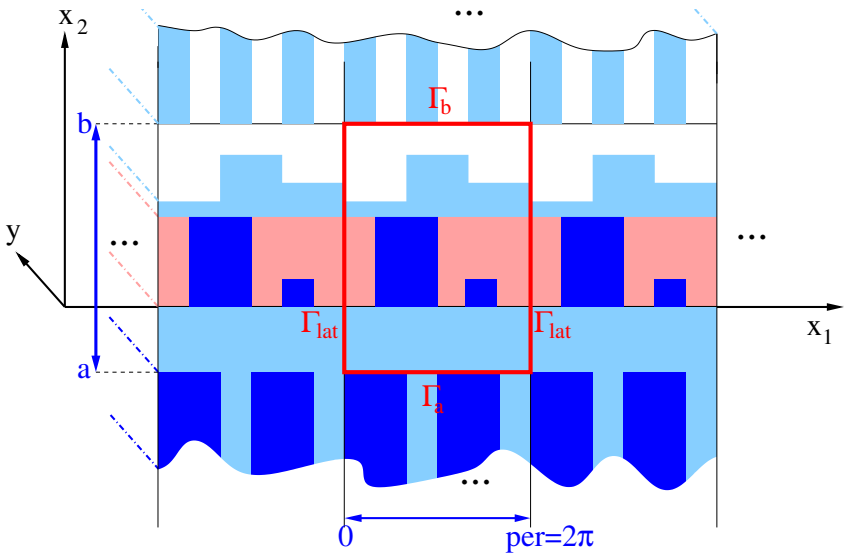

Figure 2: Geometry settings for inhomogeneous cover material and substrate.

Suppose $c = a$ or $c = b$. The eigenvalues λ of M in (3.3) are the square roots $\pm\sqrt{\lambda^2}$ for the eigenvalues λ^2 of (3.4). For definiteness, we choose the square root $\lambda = \sqrt{\lambda^2}$ s.t. either $\Re e\, \lambda > 0$

or that $\Re e\,\lambda=0$, $\Im m\,\lambda\le 0$, and consider both values $\pm\lambda$. For $\lambda=\lambda_{c,n}\neq 0$, we call the 2D wave modes $u^{\pm}_{c,n}(x_1,x_2):=e^{\mp\lambda_{c,n}(x_2-c)}f_{c,n}(x_1)$ upgoing for upper index $+$ and downgoing for index $-$. In the case $\lambda_{c,n}=0$, we define these waves by $u^{\pm}_{c,n}(x_1,x_2):=\big(1\pm(x_2-c)\big)f_{c,n}(x_1)$. The general representation of the Helmholtz solutions in the inhomogeneous cover material close to Γ_c is

$$u(x_1,x_2)\;=\;\sum_{n\in\mathbb{Z}_+}\left\{c^+_{c,n}u^+_{c,n}(x_1,x_2)+c^-_{c,n}u^-_{c,n}(x_1,x_2)\right\},\;c^{\pm}_{c,n}\in\mathbb{C}. \tag{4.1}$$

The expansion (2.2) for constant k is a special case of this, where the eigenvalue $\lambda_{c,n}=-\mathbf{i}\beta^b_n$, $n\in\mathbb{Z}$ corresponds to the eigenfunction $f_{c,n}(x_1)=e^{\mathbf{i}(\alpha+n)x_1}$. Of course the index set $\mathbb{Z}$ is to be changed into $\mathbb{Z}_+$. So, similarly to the homogeneous radiation conditions (2.3) and (2.4), we define, for the inhomogeneous medium,

Definition 4.1. *An α-quasiperiodic solution u of the 2D Helmholtz equation $\nabla\cdot k^{-2}\nabla u-u=0$ over the upper half-space Ω^+_b (over the lower half-space Ω^+_a) is said to satisfy the upper (lower) radiation condition if u admits the expansion $u(x_1,x_2)=u^{\mathrm{inc}}_b(x_1,x_2)+\sum_{n\in\mathbb{Z}_+}c^+_{b,n}u^+_{b,n}(x_1,x_2)$ for a sequence of coefficients $c^+_{b,n}\in\mathbb{C}$ (the expansion $u(x_1,x_2)=u^{\mathrm{inc}}_a(x_1,x_2)+\sum_{n\in\mathbb{Z}_+}c^-_{a,n}u^-_{a,n}(x_1,x_2)$ for a sequence of coefficients $c^-_{a,n}\in\mathbb{C}$). The sums converge in H^1_{loc}.*

Indeed, to see the convergence, we choose $b=0$ and, simplifying the notation, we set $\lambda_n=\lambda_{b,n}$ and $f_n:=f_{b,n}$. We take a general $u(x_1,x_2)=\sum_n c_n e^{-\lambda_n x_2}f_n(x_1)$ with the discrete $H^{1/2}_\alpha$ norm $\|(c_n)_n\|_\alpha:=\sum_n(1+|\lambda_n|^2)^{1/2}|c_n|^2<\infty$, and note that $\|(c_n)_n\|_\alpha\le C\|u\|_{H^{1/2}_\alpha}$ by the Riesz property for the scaled functions $(1+|\lambda_n|^2)^{1/4}f_n$ in $H^{1/2}_\alpha$. Then we get

$$\begin{aligned}
\|\partial_{x_2}u\|^2_{L^2}&=\int_0^1\int_0^{2\pi}\left|\sum_n\lambda_n c_n f_n(x_1,x_2)e^{-\lambda_n x_2}\right|^2\mathrm{d}x_1\mathrm{d}x_2\le C\int_0^1\sum_n\left|\lambda_n c_n e^{-\lambda_n x_2}\right|^2\mathrm{d}x_2\\
&\le\sum_n|\lambda_n c_n|^2\int_0^1 e^{-2\Re e\,\lambda_n x_2}\mathrm{d}x_2\le C\sum_n|\lambda_n|\,|c_n|^2\le C\|(c_n)_n\|_\alpha,\\
\|\partial_{x_1}u\|^2_{L^2}&=\int_0^1\int_0^{2\pi}\left|\partial_{x_1}\sum_n c_n f_n(x_1,x_2)e^{-\lambda_n x_2}\right|^2\mathrm{d}x_1\mathrm{d}x_2\le C\int_0^1\sum_n(1+|\lambda_n|^2)\left|c_n e^{-\lambda_n x_2}\right|^2\mathrm{d}x_2\\
&\le\sum_n(1+|\lambda_n|^2)\,|c_n|^2\int_0^1 e^{-2\Re e\,\lambda_n x_2}\mathrm{d}x_2\le C\sum_n(1+|\lambda_n|^2)^{1/2}\,|c_n|^2\le C\|(c_n)_n\|_\alpha,
\end{aligned}$$

where we have used $|\Im m\,\lambda_n|\le C\Re e\lambda_n$ to estimate $|\lambda_n|/|\Re e\,\lambda_n|$ by a constant. The corresponding estimate for the L^2 norm is similar. Hence the local H^1_α norm of u is bounded and the sum converges in this norm.

Using Def. 4.1, we can generalize the BVP for the scattering of incoming waves u^a and u^b by the grating with homogeneous cover material and substrate to that for a grating with inhomogeneous super- and substrate. From the proof of [7, Theorem 5.7], we obtain:

Theorem 4.2. *Suppose:*

i) For $c=a,b$, the systems of eigenfunctions $(1+|\lambda_{c,n}|^2)^{-s/2}f_{c,n}$ of (3.4) forms a Riesz basis in H^s_α for $-1\le s\le 1$. The eigenvalues $\lambda_{c,n}$ obey the estimate $|\Im m\,\lambda_{c,n}|\le C|\Re e\,\lambda_{c,n}|$ for a fixed positive constant C. There are no generalized eigenfunctions $f_{c,n}$ of rank greater than one.

ii) Any solution of the scattering problem with incident waves $u^{\mathrm{inc}}_a\equiv 0\equiv u^{\mathrm{inc}}_b$ is zero.

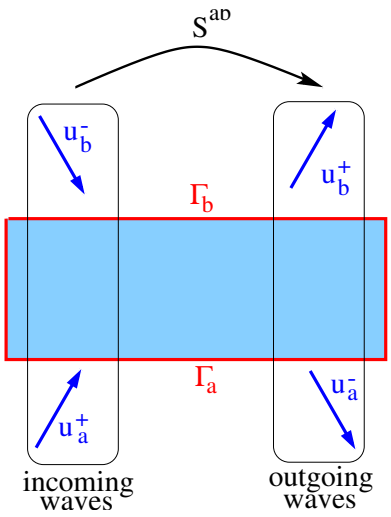

Figure 3: Scattering matrix.

Then, for any given $u_b^{\rm inc}|_{\Gamma_b} \in H_\alpha^{1/2}(\Gamma_b)$ *and* $u_a^{\rm inc}|_{\Gamma_a} \in H_\alpha^{1/2}(\Gamma_a)$*, there is a unique solution* $u \in H_\alpha^1(\Omega)$ *of the scattering problem (2.1) with the radiation conditions of Def. 4.1. In particular, there is a bounded solution operator (scattering operator or scattering matrix)* S^{ab} *mapping:*

$$\begin{pmatrix} H_\alpha^{1/2}(\Gamma_a) \\ H_\alpha^{1/2}(\Gamma_b) \end{pmatrix} \ni \begin{pmatrix} u_a^{\rm inc}|_{\Gamma_a} \\ u_b^{\rm inc}|_{\Gamma_b} \end{pmatrix} =: \begin{pmatrix} u_a^+ \\ u_b^- \end{pmatrix} \mapsto \begin{pmatrix} u_b^+ \\ u_a^- \end{pmatrix} := \begin{pmatrix} [u - u_b^{\rm inc}]|_{\Gamma_b} \\ [u - u_a^{\rm inc}]|_{\Gamma_a} \end{pmatrix} \in \begin{pmatrix} H_\alpha^{1/2}(\Gamma_b) \\ H_\alpha^{1/2}(\Gamma_a) \end{pmatrix} .$$

For assumption i), we refer to the two cases discussed in Sect. 3. In particular, there is no eigenfunction of rank greater than one for real-valued $k^\pm$ and the system of eigenfunctions is even orthonormal in the sense of $\langle k^{-2} f_n, f_m \rangle = \delta_{n,m}$. Suppose k is not a real-valued function. To our knowledge, there is no example of a rank-greater-one eigenfunction known yet. If such a function exists, then the system of upgoing and downgoing waves is to be modified and an adaption of the scattering matrix algorithm is needed. Such an adaption might be difficult if it is not clear which eigenfunction has the rank greater one. To prove a general Riesz basis property seems to be extremely difficult if infinitely many rank-greater-one eigenvalues exists.

If assumption i) is satisfied, then the variational form can be shown to be strongly elliptic. Surely, there exist trivial solutions of the scattering problem with no incoming wave, so called eigenmodes or trapped modes. If wavenumber functions with non-real values are involved (absorbing materials), then the uniqueness of ii) can be shown. If k is real-valued, then the existence of eigenmodes seems to be an exceptional case.

5 SMA with no discretization w.r.t. variable x_1

Now we introduce the SMA on the continuous level. The RCWA will be the discretization of this SMA and will be considered in Sects. 6–7. The key instrument of the SMA is the S-matrix of Thm. 4.2, which has a natural 2×2 block structure. To see this clearly we need projections in the space of boundary values, i.e. in the space of Dirichlet and Neumann data.

$$
\begin{aligned}
u_b^{\pm}(x_1,b) &= \sum_{n\in\mathbb{Z}_+} c_{b,n}^{\pm} f_{b,n}(x_1) = \sum_{n\in\mathbb{Z}_+:\lambda_{b,n}\neq 0} c_{b,m}^{\pm} f_{b,n}(x_1) e^{\pm\lambda_{c,n}[x_2-b]}\Big|_{x_2=b} \\
&\qquad + \sum_{n\in\mathbb{Z}_+:\lambda_{b,n}=0} c_{b,m}^{\pm} f_{b,n}(x_1)\Big(1\pm(x_2-b)\Big)\Big|_{x_2=b}, \\
\partial_{x_2} u_b^{\pm}(x_1,b) &= D_t N_b^{\pm}\Big(u_b^{\pm}|_{\mathbb{R}_b^3}\Big)(x_1) := \sum_{n\in\mathbb{Z}_+:\lambda_{b,n}} \partial_{x_2} u_{b,n}^{+}(x_1,b) \\
&= \pm \sum_{n\in\mathbb{Z}_+:\lambda_{b,n}\neq 0} \lambda_{c,n} c_{b,m}^{\pm} f_{b,n}(x_1) \pm \sum_{n\in\mathbb{Z}_+:\lambda_{b,n}=0} c_{b,m}^{\pm} f_{b,n}(x_1). \qquad (5.1)
\end{aligned}
$$

Due to the Riesz basis property, each trace of $u_b^{\pm}\in H_\alpha^{1/2}(\Gamma_b)$ has a unique continuation to the upper and lower half space s.t. the trace is a function $\partial_{x_2}u^{\pm}\in H_\alpha^{-1/2}(\Gamma_b)$. In this sense, the Dirichlet traces $u^{\pm}$ of the upgoing and downgoing waves can be embedded into the H_α^1 space of wave solutions above and below Γ_b, respectively. It can be embedded into the space of boundary data consisting of couples of Dirichlet and Neumann data. We identify

$$
\begin{aligned}
u_b^{\pm} &\leftrightarrow (u_b^{\pm}, k_b^{-2}\partial_{x_2}u_b^{\pm}),\ k_b := k(\cdot,b+0), \\
H_\alpha^{1/2}(\Gamma_b) &\leftrightarrow \left[H_\alpha^{1/2}\times H_\alpha^{-1/2}\right]_{\pm}(\Gamma_b) \subseteq H_\alpha^{1/2}(\Gamma_b)\times H_\alpha^{-1/2}(\Gamma_b). \qquad (5.2)
\end{aligned}
$$

Remark 5.1. *Note that the factor k_b^{-2} is new for the TM case. It does not appear for TE polarization. It is introduced since over the interface Γ_b the function u and $k^{-2}\partial_{x_2}u$ are continuous, i.e.,*

$$
\begin{aligned}
u(x_1,b+0) &= u(x_1,b-0), \\
k^{-2}(x_1,b+0)\partial_{x_2}u(x_1,b+0) &= k^{-2}(x_1,b-0)\partial_{x_2}u(x_1,b-0),\quad 0\le x_1\le 2\pi.
\end{aligned}
$$

These equalities hold in the trace spaces. In other words, we correctly have (5.2) with

$$
H_\alpha^{1/2}(\Gamma_b) \leftrightarrow H_\alpha^{1/2}(\Gamma_b)\times k_b^{-2}H_\alpha^{-1/2}(\Gamma_b),
$$

which makes it difficult to set $k^{-2}(x_1,b+0)\partial_{x_2}u(x_1,b+0)$ equal to $k^{-2}(x_1,b-0)\partial_{x_2}u(x_1,b-0)$. However, if the multiplication operator $k_b^{-2}I$ is bounded in the Sobolev space $H_\alpha^{-1/2}(\Gamma_b)$, then we have $H_\alpha^{-1/2}(\Gamma_b)=k_b^{-2}H_\alpha^{-1/2}(\Gamma_b)$ and the same trace space from above and below. In particular this is true for a Hölder continuous function k_b^{-2}. We shall use (5.2) mainly if $k_b^{-2}I$ is bounded.

Lemma 5.2. *If Assumption i) of Thm. 4.2 holds and if $k_b^{-2}I$ is a bounded operator in $H_\alpha^{-1/2}$, then the space of Dirichlet and Neumann data is the direct sum*

$$
H_\alpha^{1/2}(\Gamma_b)\times H_\alpha^{-1/2}(\Gamma_b) = \left[H_\alpha^{1/2}\times H_\alpha^{-1/2}\right]_{+}(\Gamma_b) \ \oplus\ \left[H_\alpha^{1/2}\times H_\alpha^{-1/2}\right]_{-}(\Gamma_b),
$$

and the projections $P_b^{\pm}$ of the space $H_\alpha^{1/2}(\Gamma_b)\times H_\alpha^{-1/2}(\Gamma_b)$ onto $\left[H_\alpha^{1/2}\times H_\alpha^{-1/2}\right]_{\pm}(\Gamma_b)$ parallel to $\left[H_\alpha^{1/2}\times H_\alpha^{-1/2}\right]_{\mp}(\Gamma_b)$ are bounded. In particular, we get $\operatorname{im} P_b^{\pm} = \left[H_\alpha^{1/2}\times H_\alpha^{-1/2}\right]_{\pm}(\Gamma_b)$.

Proof. From (5.1) we see $D_tN_b^- = -D_tN_b^+$. Then the representation

$$
\left(u_D, k_b^{-2}u_N\right) = \left(u^+, k_b^{-2}D_tN_b^+u^+\right) + \left(u^-, -k_b^{-2}D_tN_b^+u^-\right)
$$

leads us to

$$P_b^{\pm}\big(u_D, k_b^{-2}u_N\big) \;=\; \left(\frac{1}{2}u_D \pm \frac{1}{2}[D_tN_b^+]^{-1}u_N\,,\, \pm\frac{1}{2}k_b^{-2}D_tN_b^+u_D + \frac{1}{2}k_b^{-2}u_N\right). \quad (5.3)$$

Using (5.1) and the Riesz property of the $f_{b,n}$, $n\in\mathbb{Z}_+$, we get the boundedness of the Dirichlet-to-Neumann mappings and the last formula proves the continuity of the projections. □

Note that in the TE case (cf. [7, Lemma 6.1]), the assumption on the boundedness of $k_b^{-2}I$ in $H_\alpha^{-1/2}$ is redundant.

Analogously to the projections $P_b^{\pm}$ in $H^{1/2}\times H^{-1/2}$ over Γ_b based on the eigenfunctions $f_{b,n}$ for k^2L with $k(x_1)=k(x_1,b+0)$ we have the projections $P_a^{\pm}$ in $H^{1/2}\times H^{-1/2}$ over Γ_a based on $k(x_1)=k(x_1,a+0)$. From (5.3) and $P_a^{\pm}(u,v)=(u_a^{\pm},\pm k_a^{-2}D_tN_a^+u_a^{\pm})$, we get the formula

$$\begin{aligned} P_b^{\pm}\big(u_a^+ + u_a^-, k_a^{-2}D_tN_a^+u_a^+ - k_a^{-2}D_tN_a^+u_a^-\big) \;&=\; (5.4)\\ \left(\frac{1}{2}[u_a^+ + u_a^-] \pm \frac{1}{2}[D_tN_b^+]^{-1}\frac{k_b^2}{k_a^2}[D_tN_a^+u_a^+ - D_tN_a^+u_a^-]\,,\right.&\\ \left. \pm\frac{1}{2}k_b^{-2}D_tN_b^+[u_a^+ + u_a^-] + \frac{1}{2}k_a^{-2}[D_tN_a^+u_a^+ - D_tN_a^+u_a^-]\right), \quad &(5.5)\end{aligned}$$

which, by the identification (5.2), can be written as $P_b^{\pm}\colon H_\alpha^{1/2}\to H_\alpha^{1/2}$ and

$$P_b^{\pm}[u_a^+] = \frac{1}{2}\Big[u_a^+ \pm [D_tN_b^+]^{-1}\frac{k_b^2}{k_a^2}D_tN_a^+u_a^+\Big],\; P_b^{\pm}[u_a^-] = \frac{1}{2}\Big[u_a^- \mp [D_tN_b^+]^{-1}\frac{k_b^2}{k_a^2}D_tN_a^+u_a^-\Big]. \quad (5.6)$$

In this sense, we arrive at $\operatorname{im}P_b^{\pm}=H_\alpha^{1/2}$ and $P_b^{\pm}\colon H_\alpha^{1/2}\to H_\alpha^{1/2}$ provided that $k_a^{-2}I$ and $k_b^{-2}I$ are bounded operators in $H_\alpha^{-1/2}$. Whereas $P^{\pm}$ in (5.4) is a projection, the identified operators for $P^{\pm}$ on the right-hand sides of (5.6) are not. The validity of (5.6) depends on the knowledge that the functions $u_a^{\pm}$ are the Dirichlet traces of outgoing and downgoing waves, respectively. Note that, for the TE case, (5.6) holds with the factor k_b^2/k_a^2 deleted.

So the S-matrix (cf. Fig. 3) acting in the boundary-value space $H_\alpha^{1/2}\times H_\alpha^{-1/2}$ consists of the four blocks $S_{++}^{ab}:=P_b^+S^{ab}|_{\operatorname{im}P_a^+}$, $S_{+-}^{ab}:=P_b^+S^{ab}|_{\operatorname{im}P_b^-}$, $S_{-+}^{ab}:=P_a^-S^{ab}|_{\operatorname{im}P_a^+}$, and $S_{--}^{ab}:=P_a^-S^{ab}|_{\operatorname{im}P_b^-}$, which we identify by their corresponding operators in $H_\alpha^{1/2}$ (cf. (5.2)). Its action corresponds to a system of two linear equations.

$$S^{ab} \;=\; \begin{pmatrix} S_{++}^{ab} & S_{+-}^{ab} \\ S_{-+}^{ab} & S_{--}^{ab}\end{pmatrix}, \qquad \begin{aligned} u_b^+ &= S_{++}^{ab}u_a^+ + S_{+-}^{ab}u_b^-,\\ u_a^- &= S_{-+}^{ab}u_a^+ + S_{--}^{ab}u_b^-.\end{aligned}$$

Note that the four blocks of S^{ab} identified as operators acting in the spaces $\operatorname{im}P_a^{\pm}=H_\alpha^{1/2}(\Gamma_a)$ and $\operatorname{im}P_b^{\pm}=H_\alpha^{1/2}(\Gamma_b)$ are continuous by the mapping property of the variational operator of Thm. 4.2.

We start the derivation of the algorithm with the case of a grating consisting of two adjacent slices (cf. (4)). Suppose the S-matrices S^{ab} and S^{bc} of the slices between Γ_a and Γ_b and between Γ_b and Γ_c, respectively, are known (cf. Sect. 6). How does the matrix S^{ac} between Γ_a and Γ_c looks like? In other words, we know

$$\begin{aligned} u_c^+ &= S_{++}^{bc}u_b^+ + S_{+-}^{bc}u_c^- &(5.7)\\ u_b^- &= S_{-+}^{bc}u_b^+ + S_{--}^{bc}u_c^- &(5.8)\\ u_b^+ &= S_{++}^{ab}u_a^+ + S_{+-}^{ab}u_b^- &(5.9)\\ u_a^- &= S_{-+}^{ab}u_a^+ + S_{--}^{ab}u_b^-, &(5.10)\end{aligned}$$

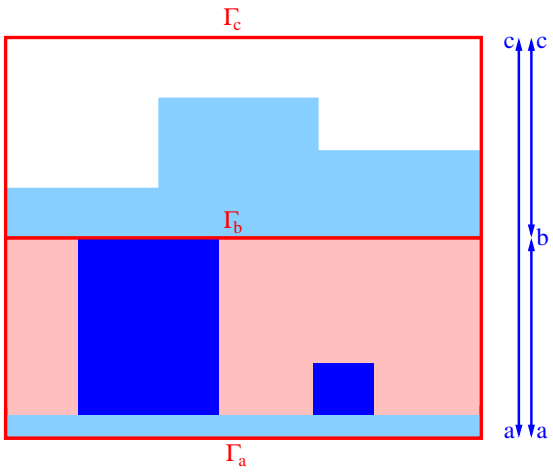

Figure 4: Step from two slices to their union.

and we look for

$$u_c^+ \;=\; S_{++}^{ac}u_a^+ + S_{+-}^{ac}u_c^-, \qquad u_a^- \;=\; S_{-+}^{ac}u_a^+ + S_{--}^{ac}u_c^-. \tag{5.11}$$

We set the traces $u_\pm^b = u_\pm^b(\cdot, b)$ of the functions $u_\pm^b(\cdot,\cdot)$ in the slice between Γ_b and Γ_c to the corresponding upper output and input functions $u_\pm^b$ of the S-matrix for the slice between Γ_b and Γ_c. Both traces are in the trace space $H_\alpha^{1/2}\times k_{b+0}^{-2}H_\alpha^{-1/2}$. Then we eliminate these unknown functions from the linear system (5.7)–(5.10). Defining $D := (I - S_{-+}^{bc}S_{+-}^{ab})$, we arrive at the linear system (5.11) with the operator coefficients

$$S^{ac} = \begin{pmatrix} S_{++}^{ac} & S_{+-}^{ac} \\ S_{-+}^{ac} & S_{--}^{ac} \end{pmatrix} = \begin{pmatrix} S_{++}^{bc}\left[I + S_{+-}^{ab}D^{-1}S_{-+}^{bc}\right]S_{++}^{ab} & S_{+-}^{bc} + S_{++}^{bc}S_{+-}^{ab}D^{-1}S_{--}^{bc} \\ S_{-+}^{ab} + S_{--}^{ab}D^{-1}S_{-+}^{bc}S_{++}^{ab} & S_{--}^{ab}D^{-1}S_{--}^{bc} \end{pmatrix}. \tag{5.12}$$

Lemma 5.3. *Suppose the BVP (2.1) for the three gratings between Γ_a and Γ_b, between Γ_a and Γ_c, and between Γ_b and Γ_c are uniquely solvable s.t. the S-matrices S^{ab}, S^{ac}, and S^{bc} exist. Furthermore, suppose $k_b(x_1) := k(x_1, b)$ and $k_c(x_1) := k(x_1, c)$ are piecewise twice continuously differentiable w.r.t. x_1. Then the operator $D := (I - S_{-+}^{bc}S_{+-}^{ab})$ is invertible.*

Proof. Due to the definition D and due to the compactness of S_{-+}^{bc} (cf. the subsequent (6.7) and use the compactness of T_+^{ab} following from (6.8)), the operator D is a Fredholm operator of index zero. It remains to prove that codimension of $\operatorname{im} D \subseteq \operatorname{im} P^-$ is zero, i.e. that the image space of D is dense.

For incoming waves $u_a^+ = 0$ and u_c^-, there exists a solution u in the grating between Γ_a and Γ_c. Taking the restrictions to Γ_a, Γ_b, and Γ_c and their projections to the up- and downgoing waves, we get the waves u_a^-, $u_b^\pm$, and u_c^+. The Eqns. (5.8) and (5.9) lead to the system

$$\begin{aligned} -S_{-+}^{bc}u_b^+ + u_b^- &= S_{--}^{bc}u_c^-, \\ u_b^+ - S_{+-}^{ab}u_b^- &= S_{++}^{ab}u_a^+ = 0. \end{aligned}$$

Multiplying the last equation by S_{-+}^{bc} and adding the result to the first, we arrive at $Du_b^- = S_{--}^{bc}u_c^-$. From the subsequent Eqns. (6.7) and (6.8), we observe that the image space of D is dense. □

Now consider the general case and split the rectangular domain of Fig. 2 into n smaller slices (cf. Fig. 5). We denote the S-matrices of the slices by S^j and the Dirichlet boundary values on the slice

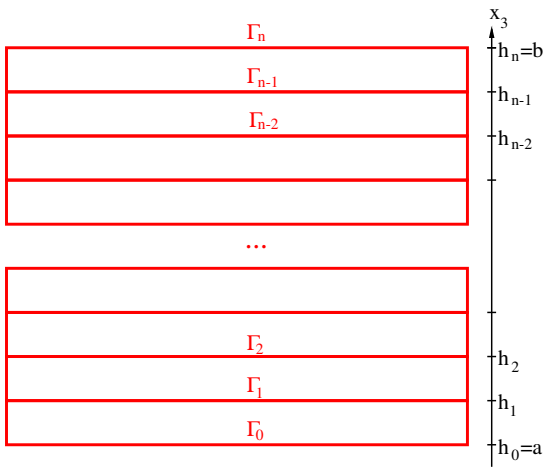

Figure 5: Step from many small slices to their union.

boundaries by $u_j^{\pm}$. Furthermore, we introduce the accumulated S-matrix $\mathbb{S}^j$ over the union over all slices below the $(j+1)$th slice. I.e., we have

$$u_j^{\pm} := u_{h_j}^{\pm}, \quad S^j := S^{h_{j-1}h_j}, \quad \mathbb{S}^j := S^{h_0h_j}, \ j=1,\cdots,n\ .$$

Suppose, for each small slice between $\Gamma_{h_{j-1}}$ and Γ_{h_j}, we can compute the S-matrix S^j, which requires a solver for the BVP (2.1) (cf. Sect. 6). With this we get the

Scattering matrix algorithm.

i) Compute recursively the accumulated S-matrix $\mathbb{S}^n$:

  i)- i) Initialization:
  Set $j=1$ and compute $\mathbb{S}^j=S^1$ (cf. Sect. 6).

  i)-ii) Iteration for j running from 2 to n:
  Compute S^j (cf. Sect. 6).
  Apply the two-step formula (5.12) with $S^{ab}=\mathbb{S}^{j-1}$, $S^{bc}=S^j$, and $S^{ac}=\mathbb{S}^j$ to compute $\mathbb{S}^j$ from $\mathbb{S}^{j-1}$ and S^j. (5.13)

ii) Given the incoming wave modes u_0^+ and u_n^-, compute the reflected and transmitted waves $u_n^+=\mathbb{S}_{++}^n u_0^+ + \mathbb{S}_{+-}^n u_n^-$ and $u_0^-=\mathbb{S}_{-+}^n u_0^+ + \mathbb{S}_{--}^n u_n^-$ and their Rayleigh coefficients $c_{b,n}^+$ and $c_{a,n}^-$, respectively.
Compute the scaled squared moduli $|c_{b,n}^+|^2$ and $|c_{a,n}^-|^2$ to obtain the efficiencies (intensities) of the reflected and transmitted wave modes.
Compute the arguments of the complex numbers $c_{b,n}^+/|c_{b,n}^+|$ and $c_{a,n}^-/|c_{a,n}^-|$ to get the phase shifts of the modes.

Remark 5.4. *Note that, in applications, the radiation condition of Def. (4.1) over Γ_{h_0} and Γ_{h_n} might be the classical one of (2.3) and (2.4). However, splitting the whole domain of the grating into smaller slices, the wave-number function on some of the vertical slice boundaries Γ_{h_j} will not be constant and we rely on the Def. (4.1). This condition is valid at least on an infinitesimal small neighbourhood of the slice boundary. Though the developers of the RCWA never thought about a radiation condition for all $x_2>h_j$ or all $x_2<h_j$, they use this condition in the S-matrices of the RCWA.*

Remark 5.5. *The above defined scattering matrix algorithm updates the four blocks $\mathbb{S}_{++}^j$, $\mathbb{S}_{+-}^j$, $\mathbb{S}_{-+}^j$, and $\mathbb{S}_{--}^j$ in each step. A reduced algorithm is possible if $u_n^-\equiv 0$. Then it is sufficient to update two blocks of the S-matrix and two vectors.*

Remark 5.6. *If we are interested in the solution over the slices, then we can go backwards. Using the two-step equations (5.9) and (5.8) for the two slices $\mathbb{S}^{n-1}$ and S^n, we compute u^+_{n-1} and u^-_{n-1}. Using (5.9) and (5.8) for the two slices $\mathbb{S}^{n-2}$ and S^{n-1}, we compute $u^\pm_{n-2}$. Using (5.9) and (5.8) for the two slices $\mathbb{S}^{n-3}$ and S^{n-2}, we compute $u^\pm_{n-3}$. Going recursively up to 1, we get $u^\pm_1$ from (5.9) and (5.8) for the two slices $\mathbb{S}^1$ and S^2. Finally, over each slice between $\Gamma_{h_{j-1}}$ and Γ_{h_j}, we apply the solver for (2.1), which has been used for the computation of the S-matrix S^j (cf. Sect. 6). Knowing the boundary data $u^\pm_{j-1}$, the solver provides us with the values of the wave solution between $\Gamma_{h_{j-1}}$ and Γ_{h_j}.*

6 Solution of the scattering problem over a slice and computation of the S-matrix

Clearly, the scattering problem over the slice is equivalent to a variational formulation, which can be solved numerically by FEM. Then the combination of the iteration of the scattering matrix algorithm in Sect. 5 with FEM is nothing else than a DDM for the FEM. In engineering applications, however, the following different approach is used (cf. (6.6) and (6.7)), which reduces the computation to the solution of the equivalent operator valued ODE.

To prepare the formula of the S-matrix, we need a few definitions. Recall the identification in (5.2) and the splitting of the boundary data in Lemma 5.2. Analogously to the projections $P^\pm_b$ in $H^{1/2}\times H^{-1/2}$ over Γ_b based on the eigenfunctions $f_{b,n}$ for the differential operator k^2L with $k(x_1)=k(x_1,b+0)$, we define the projections $P^\pm_{b-0}$ in $H^{1/2}\times H^{-1/2}$ over Γ_b based on the eigenfunctions $f_{b-0,n}$ for k^2L with $k(x_1)=k(x_1,b-0)$. We introduce the transition operators T^+_{ab} and T^-_{ba}

$$\begin{aligned} T^+_{ab} &: \operatorname{im} P^+_a \to \operatorname{im} P^+_{b-0} \subset H^{1/2}(\Gamma_b)\times H^{-1/2}(\Gamma_b),\\ T^-_{ba} &: \operatorname{im} P^-_{b-0} \to \operatorname{im} P^+_a \subset H^{1/2}(\Gamma_a)\times H^{-1/2}(\Gamma_a). \end{aligned}$$

The operator T^+_{ab} maps $u^+_a \leftrightarrow (u^+_a, k_a^{-2}v^+_a := k_a^{-2}D_tN_au^+_a)$ to $\big(u(\cdot,b_0), k^{-2}_{b-0}\partial_{x_2}u(\cdot,b-0)\big)$, where u is the solution of the initial value problem

$$\begin{array}{lrcll} a) & \partial^2_{x_2}u(x_1,x_2) &=& k^2[Lu](x_1,x_2), & 0\le x_1\le 2\pi,\ a\le x_2\le b,\\ b) & u(x_1,a) &=& u^+_a(x_1), & 0\le x_1\le 2\pi,\\ c) & \partial_{x_2}u(x_1,a) &=& v^+_a(x_1), & 0\le x_1\le 2\pi, \end{array} \tag{6.1}$$

of the operator valued ODE $\partial^2_{x_2}u = k^2Lu$ equivalent to the Helmholtz equation (cf. (3.2)). Similarly, T^-_{ba} maps $u^-_{b-0} \leftrightarrow (u^-_{b-0}, k^{-2}_{b-0}v^-_{b-0} := -k^{-2}_{b-0}D_tN_{b-0}u^-_{b-0})$ to $\big(u(\cdot,a), k_a^{-2}\partial_{x_2}u(\cdot,a)\big)$, where u is the solution of the problem

$$\begin{array}{lrcll} a) & \partial^2_{x_2}u(x_1,x_2) &=& k^2[Lu](x_1,x_2), & 0\le x_1\le 2\pi,\ a\le x_2\le b,\\ b) & u(x_1,b) &=& u^-_{b-0}(x_1), & 0\le x_1\le 2\pi,\\ c) & \partial_{x_2}u(x_1,b) &=& v^+_{b-0}(x_1), & 0\le x_1\le 2\pi. \end{array} \tag{6.2}$$

Note that, for the case of x_2 invariant wavenumber $k(x_1,x_2)=k(x_1)$ in the slice $[0,2\pi]\times[a,b)$, we get $P^\pm_{b-0}=P^\pm_a$, and the transition operators T^+_{ab} and T^-_{ba} are given by (cf. (4.1))

$$\begin{aligned} T^+_{ab}[u^+_{a,n}(\cdot,a)] &= u^+_{a,n}(\cdot,b-0)\in\operatorname{im}P^+_a,\\ T^-_{ba}[u^-_{a,n}(\cdot,b)] &= u^-_{a,n}(\cdot,a)\in\operatorname{im}P^-_a. \end{aligned}$$

W.r.t. the basis $f_{a,n}$, $n\in\mathbb{Z}_+$, both transition operators T^+_{ab} and T^-_{ba} have the same diagonal matrix $(e^{-\lambda_{a,n}[b-a]}\delta_{m,n})_{m,n\in\mathbb{Z}_+}$ and are bounded.

Next we derive the formula for the S-matrix. The boundary values $v_a^+ \in \operatorname{im} P_a^+$ over Γ_a and $v_b^- = 0$ on the curve Γ_b lead to a wave solution with boundary values $P_b^+ T_{ab}^+ v_a^+ \in \operatorname{im} P_b^+$ and $P_b^- T_{ab}^+ v_a^+ \in \operatorname{im} P_b^-$.

$$S^{ab} : \begin{pmatrix} v_a^+ \\ P_b^- T_{ab}^+ v_a^+ \end{pmatrix} \mapsto \begin{pmatrix} P_b^+ T_{ab}^+ v_a^+ \\ 0 \end{pmatrix}. \tag{6.3}$$

Take $v_{b-0}^- \in \operatorname{im} P_{b-0}^-$. Then the boundary values $P_b^+ v_{b-0}^- \in \operatorname{im} P_b^+$ and $P_b^- v_{b-0}^- \in \operatorname{im} P_b^-$ on the curve Γ_b as well as $P_a^+ T_{ba}^- v_{b-0}^- \in \operatorname{im} P_a^-$ and $P_a^- T_{ba}^- v_{b-0}^- \in \operatorname{im} P_a^-$ on Γ_a lead to

$$S^{ab} : \begin{pmatrix} P_a^+ T_{ba}^- v_{b-0}^- \\ P_b^- v_{b-0}^- \end{pmatrix} \mapsto \begin{pmatrix} P_b^+ v_{b-0}^- \\ P_a^- T_{ba}^- v_{b-0}^- \end{pmatrix}. \tag{6.4}$$

For the functions $u_a^+ = v_a^+ + P_a^+ T_{ba}^- v_{b-0}^-$ and $u_b^- = P_b^- T_{ab}^+ v_a^+ + P_b^- v_{b-0}^-$, Equations (6.3) and (6.4) yield

$$\begin{aligned} \begin{pmatrix} u_a^+ \\ u_b^- \end{pmatrix} &= \begin{pmatrix} I|_{\operatorname{im} P_a^+} & P_a^+ T_{ba}^- \\ P_b^- T_{ab}^+ & P_b^-|_{\operatorname{im} P_{b-0}^-} \end{pmatrix} \begin{pmatrix} v_a^+ \\ v_{b-0}^- \end{pmatrix}, \\ S^{ab} \begin{pmatrix} u_a^+ \\ u_b^- \end{pmatrix} &= \begin{pmatrix} P_b^+ T_{ab}^+ & P_b^+|_{\operatorname{im} P_{b-0}^-} \\ 0 & P_a^- T_{ba}^- \end{pmatrix} \begin{pmatrix} v_a^+ \\ v_{b-0}^- \end{pmatrix}. \end{aligned} \tag{6.5}$$

Assuming that the determinant operator $D_{ab}^- := \{P_b^-|_{\operatorname{im} P_{b-0}^-} - P_b^- T_{ab}^+ P_a^+ T_{ba}^-\} \colon \operatorname{im} P_{b-0}^- \to \operatorname{im} P_b^-$ of the first matrix in (6.5) is invertible, we arrive at

$$\begin{aligned} S^{ab} &= \begin{pmatrix} P_b^+ T_{ab}^+ & P_b^+|_{\operatorname{im} P_{b-0}^-} \\ 0 & P_a^- T_{ba}^- \end{pmatrix} \begin{pmatrix} I|_{\operatorname{im} P_a^+} + P_a^+ T_{ba}^- [D_{ab}^-]^{-1} P_b^- T_{ab}^+ & -P_a^+ T_{ba}^- [D_{ab}^-]^{-1} \\ -[D_{ab}^-]^{-1} P_b^- T_{ab}^+ & [D_{ab}^-]^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \left\{ P_b^+ - P_b^+ \left[P_{b-0}^- - T_{ab}^+ P_a^+ T_{ba}^- \right] [D_{ab}^-]^{-1} P_b^- \right\} T_{ab}^+ & \left[P_b^+ - P_b^+ T_{ab}^+ P_a^+ T_{ba}^- \right] [D_{ab}^-]^{-1} \\ -P_a^- T_{ba}^- [D_{ab}^-]^{-1} P_b^- T_{ab}^+ & P_a^- T_{ba}^- [D_{ab}^-]^{-1} \end{pmatrix}. \end{aligned} \tag{6.6}$$

Note that, for the case of x_2 invariant wavenumber $k(x_1, x_2) = k(x_1)$ in the slice $[0, 2\pi] \times [a, b)$, the formula (6.6) simplifies to $S^{ab} \colon \operatorname{im} P_a^+ \times \operatorname{im} P_b^- \to \operatorname{im} P_b^+ \times \operatorname{im} P_a^-$ with

$$S^{ab} = \begin{pmatrix} P_b^+ T_{ab}^+ - P_b^+ [D_{ab}^-]^{-1} P_b^- T_{ab}^+ & P_b^+ [D_{ab}^-]^{-1} \\ -T_{ba}^- [D_{ab}^-]^{-1} P_b^- T_{ab}^+ & T_{ba}^- [D_{ab}^-]^{-1} \end{pmatrix}, \tag{6.7}$$

where $D_{ab}^- = P_b^- : \operatorname{im} P_a^- \to \operatorname{im} P_b^-$.

Lemma 6.1. *Consider a grating in the domain $[0, 2\pi] \times [a, b]$ with a wavenumber function k s.t. $k(x_1, x_2) = k_a(x_1)$ for $a \le x_2 < b$ and in the substrate $x_2 \le a$ and s.t. $k(x_1, x_2) = k_b(x_1)$ in the cover material $b \le x_2$. Suppose k_a and k_b are piecewise twice continuously differentiable w.r.t. x_1. Furthermore suppose the BVP (2.1) over this grating is uniquely solvable (cf. Thm. 4.2), i.e., there exists the bounded S-matrix S^{ab}. Then the operator $D_{ab}^- = P_b^- \colon \operatorname{im} P_a^- \to \operatorname{im} P_b^-$ is invertible.*

Proof. If $v_a^+ = 0$ and $P_b^- v_{b-0}^- = 0$ and if u_a^+ and u_b^- are defined by (6.5), then, due to $P_a^+ T_{ba}^- = 0$, we get $u_a^+ = 0$ and $u_b^- = 0$ s.t. $S^{ab}(u_a^+, u_b^-)^\top = (0, 0)^\top$. In particular, $P_a^- T_{ba}^- v_{b-0}^- = T_{ba}^- v_{b-0}^- = 0$ s.t. $v_{b-0}^- = 0$. In other words, the null space of the operator $P_b^- \colon \operatorname{im} P_a^- \to \operatorname{im} P_b^- P_a^- \subseteq \operatorname{im} P_b^-$ is trivial and the right inverse maps $[P_b^-|_{\operatorname{im} P_a^-}]^{-1} \colon \operatorname{im} P_b^- P_a^- \to \operatorname{im} P_a^-$.

It remains to show that the image of D_{ab}^- coincides with the space $\operatorname{im} P_b^-$. Consider the scattering problem (2.1) with the incoming functions $u_a^+ = 0$ from below and an arbitrary $u_b^- \in \operatorname{im} P_b^-$ from above. Then there exists a unique wave solution $u \in H_\alpha^1([0, 2\pi] \times [a, b])$ and a unique solution pair

$u_b^+ \in \operatorname{im} P_b^+ \subset H^{1/2}(\Gamma_b)$ and $u_a^- \in \operatorname{im} P_a^- \subset H^{1/2}(\Gamma_a)$. Clearly, we have $u_{b-0}^{\pm} = P_{b-0}^{\pm}(u_b^+ + u_b^-)$ s.t. we get $[T_{ab}^+]^{-1}P_{b-0}^+(u_b^+ + u_b^-) = u_a^+ = 0$ and $T_{ba}^- P_{b-0}^-(u_b^+ + u_b^-) = u_a^-$, where $P_{b-0}^{\pm} = P_a^{\pm}$. In other words, we have $P_{b-0}^+(u_b^+ + u_b^-) = 0$ and $P_{b-0}^-(u_b^+ + u_b^-) = [T_{ba}^-]^{-1}u_a^-$. This implies the inclusion $(u_b^+ + u_b^-) \in \operatorname{im} P_{b-0}^-$ and $u_b^+ + u_b^- = u_{b-0}^- = [T_{ab}^-]^{-1}u_a^-$. For the boundary data of the wave solution $u = u_b^+ + u_b^- = u_{b-0}^- \in \operatorname{im} P_{b-0}^- \subset H_\alpha^{1/2}$, this gives $P_b^- u_{b-0}^- = u_b^- \in \operatorname{im} D_{ab}^-$, and the arbitrary $u_b^- \in \operatorname{im} P_b^-$ is in the image of D_{ab}^-. □

We conjecture that, generally, the operator D_{ab}^- is a Fredholm operator of index zero (invertible operator plus compact operator). Unfortunately, in contrast to the case of TE polarization (cf. [7, proof of Lemma 6.2]), we cannot prove this fact for TM. For small widths $b-a$ and for an invertible BVP (2.1), the above proof might be helpful to derive the invertible of a general D_{ab}^-.

The matrix $\mathbf{S}^{ab}$ of S^{ab} w.r.t. the four bases, namely with $f_{a,n},\ n \in \mathbb{Z}_+$ in $\operatorname{im} P_a^+$, with the same basis in $\operatorname{im} P_a^-$, with the basis $f_{b,n},\ n \in \mathbb{Z}_+$ in $\operatorname{im} P_b^+$, and with the same basis in $\operatorname{im} P_b^-$ is

$$\mathbf{S}^{ab} = \begin{pmatrix} \left(\Theta_{++}^{ab} - \Theta_{+-}^{ab}[\Theta_{--}^{ab}]^{-1}\Theta_{-+}^{ab}\right) T^{ab} & \Theta_{+-}^{ab}[\Theta_{--}^{ab}]^{-1} \\ -T^{ab}\Theta_{-+}^{ab}T^{ab} & T^{ab}[\Theta_{--}^{ab}]^{-1} \end{pmatrix},\ T^{ab} = \left(e^{-(b-a)\lambda_{a,n}}\delta_{n,m}\right)_{n,m\in\mathbb{Z}_+}. \quad (6.8)$$

Here Θ^{ab} with its four blocks $\Theta_{+\pm}^{ab}$ and $\Theta_{-\pm}^{ab}$ is the matrix of the basis transform in $H_\alpha^{1/2} \times H_\alpha^{-1/2}$ from the bases $(f_{a,n}, \pm k_a^{-2}\lambda_{a,n}f_{a,n}),\ n \in \mathbb{Z}_+$ to the bases $(f_{b,n}, \pm k_b^{-2}\lambda_{a,n}f_{b,n}),\ n \in \mathbb{Z}_+$.

Finally, we shall present a formula for the S-matrix alternative to (6.6), which is frequently used and yields, in most cases, almost the same results. Only for large truncation indices N (cf. the subsequent (7.1)) and for deep gratings (i.e. for big widths $b-a$), there appear exponentials with large real arguments leading to overflow problems in the numerical computation. This alternative formula has no essential advantages in comparison with (6.6) but contains unbounded operators, which make the analysis difficult. Problems of (6.6) with the invertibility of D_{ab}^- correspond to problems of the subsequent (6.10) with the invertibility of E_{ab}^-. So we mention only the formula. We even restrict ourselves to the case of gratings with a wavenumber function independent of x_2 for $a \le x_2 < b$ (compare the special case (6.7) of (6.6)).

Define the operator $\mathbb{T}_{ab}$ mapping $H_\alpha^{1/2}(\Gamma_a) \times H_\alpha^{-1/2}(\Gamma_a)$ to $H_\alpha^{1/2}(\Gamma_b) \times H_\alpha^{-1/2}(\Gamma_b)$ by (6.1) but with general initial values (u_a, v_a) instead of (u_a^+, v_a^+). Then

$$\begin{pmatrix} u_b^+ \\ u_b^- \end{pmatrix} = \begin{pmatrix} [P_b^+\mathbb{T}_{ab}|_{\operatorname{im} P_a^+}] & [P_b^+\mathbb{T}_{ab}|_{\operatorname{im} P_a^-}] \\ [P_b^-\mathbb{T}_{ab}|_{\operatorname{im} P_a^+}] & [P_b^-\mathbb{T}_{ab}|_{\operatorname{im} P_a^-}] \end{pmatrix} \begin{pmatrix} u_a^+ \\ u_a^- \end{pmatrix}. \quad (6.9)$$

Supposing the existence of the inverse of $E_{ab}^- := [P_b^-\mathbb{T}_{ab}|_{\operatorname{im} P_a^-}]$ and writing this vector equation as a system of two equations and solving the latter w.r.t. the unknowns u_b^+ and u_a^-, we get the vector equation $(u_b^+, u_a^-)^\top = S^{ab}(u_a^+, u_b^-)^\top$ with

$$S^{ab} = \begin{pmatrix} [P_b^+\mathbb{T}_{ab}|_{\operatorname{im} P_a^+}] - [P_b^+\mathbb{T}_{ab}|_{\operatorname{im} P_a^-}][E_{ab}^-]^{-1}[P_b^-\mathbb{T}_{ab}|_{\operatorname{im} P_a^+}] & [P_b^+\mathbb{T}_{ab}|_{\operatorname{im} P_a^-}][E_{ab}^-]^{-1} \\ -[E_{ab}^-]^{-1}[P_b^-\mathbb{T}_{ab}|_{\operatorname{im} P_a^+}] & [E_{ab}^-]^{-1} \end{pmatrix}. \quad (6.10)$$

If this formula is used for the SMA of (5.13), then it should be used at most in the initialization step i)-i) to compute $\mathbb{S}^1$. For the updates of the $\mathbb{S}^j$ in i)-ii), a different two-step formula should be used, which computes $\mathbb{S}^j$ from $\mathbb{S}^{j-1}$ and from the T-matrix $\mathbb{T}_{h_{j-1},h_j}$ directly. Indeed, this new formula can be derived similarly to (5.12), replacing (5.7)–(5.8) by (6.9).

7 Discretization used by RCWA and FMM

Whereas in the FEM discretization of the SMA the domain of each slice is split into triangular subdomains and the functions are approximated by low order polynomial functions over each triangle, the classical SMA, i.e. the RCWA or the FMM, are based on approximation by truncated Fourier series w.r.t. variable x_1. Of course, the Fourier coefficients depend on x_2. In other words, the α-quasiperiodic function is expanded as the sum (cf. (2.2))

$$u(x_1,x_2) \;=\; \sum_{l\in\mathbb{Z}} \hat{u}_l(x_2) e^{\mathbf{i}(\alpha+l)x_1}.$$

Then a truncation index $N>0$ is fixed and an approximate function

$$u_N(x_1,x_2) \;=\; \sum_{l=-N}^{N} \hat{u}_{N,l}(x_2) e^{\mathbf{i}(\alpha+l)x_1} \;\approx\; \mathcal{P}_N u(x_1,x_2) \;:=\; \sum_{l=-N}^{N} \hat{u}_l(x_2) e^{\mathbf{i}(\alpha+l)x_1} \tag{7.1}$$

is sought. Setting $v:=k^{-2}\partial_{x_2}u$ and $\vec{u}:=(u,v)^\top$, the PDE $\nabla\cdot k^{-2}\nabla u + u = 0$ is equivalent (compare Equ. (3.1) for the case of x_2-independent wavefunction) to the ODE $\partial_{x_2}\vec{u} = M_{x_2}\vec{u}$ with operator valued coefficients (compare (3.3) and (3.2)),

$$M_{x_2} := \begin{pmatrix} 0 & k^2 I \\ L & 0 \end{pmatrix}, \quad Lu := -\partial_{x_1} k^{-2} \partial_{x_1} u - u,$$

which is approximated by the projected equation $\partial_{x_2}\vec{u}_N = M_{x_2,N}\vec{u}_N$ including the operator valued matrix coefficient $M_{x_2,N}$ defined as

$$\begin{aligned} M_{x_2,N} \;&:=\; \begin{pmatrix} 0 & \left[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}\right]^{-1} \\ \left[\mathcal{P}_N L|_{\operatorname{im}\mathcal{P}_N}\right] & 0 \end{pmatrix}, \\ \left[\mathcal{P}_N L|_{\operatorname{im}\mathcal{P}_N}\right] u_N \;&=\; -\partial_{x_1}\left[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}\right]\partial_{x_1} u_N - u_N. \end{aligned} \tag{7.2}$$

Note that the matrix of $[\mathcal{P}_N k^{-2}(\cdot,x_2) I|_{\operatorname{im}\mathcal{P}_N}]$ w.r.t. the basis functions $x_1\mapsto e^{\mathbf{i}(\alpha+l)x_1}$ with the indices $-N\le l\le N$ is a Toeplitz matrix and that of $[\mathcal{P}_N\partial_{x_1}|_{\operatorname{im}\mathcal{P}_N}] = \partial_{x_1}|_{\operatorname{im}\mathcal{P}_N}$ is the diagonal matrix $\big(\delta_{l,k}\mathbf{i}(\alpha+l)\big)_{l,k=-N}^{N}$.

Remark 7.1. *For a piecewise smooth multiplicator function g, the use of $[\mathcal{P}_N g^{-1} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}$ instead of $[\mathcal{P}_N g I|_{\operatorname{im}\mathcal{P}_N}]$ improves the approximation (cf. [10]) if gu is smoother than u. In (7.2) the inverse Galerkin approximation appears naturally from the reduction of the second-order differential equation $\partial_{x_2}[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_2}u_N = -\partial_{x_1}[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}u_N - u_N$ to a system of two first-order equations.*

Starting from (7.2), we define the algorithms and formulas from Sects. 3-6 on a discrete level. The projections $\mathcal{P}_N$ onto the truncated Fourier series are bounded in $H_\alpha^{\pm 1/2}$ with a norm uniformly bounded w.r.t. N. The corresponding operator on the spaces of Dirichlet and Neumann data over Γ_c with $c=a,b$ will be denoted by $\mathcal{P}_{N,N}^c := \mathcal{P}_N^c\otimes\mathcal{P}_N^c \in \mathcal{L}([H_\alpha^{1/2}(\Gamma_c)\times H_\alpha^{-1/2}(\Gamma_c)])$. So the discrete version of the space of Dirichlet and Neumann data is $\operatorname{im}\mathcal{P}_{N,N}^c$. Due to the assumptions $\Re e\, k>0$ and $\Im m\, k\ge 0$, there is a $\zeta\in\mathbb{C}$ s.t. $\Re e\,[\zeta k_c^{-2}]\ge\varepsilon>0$ s.t. the operator of multiplication by $[\zeta k_c^{-2}]$ is positive definite. Consequently, for bounded multiplication operators $k_c^{-2}I\in\mathcal{L}(H_\alpha^{-1/2})$, the Galerkin method applies, the operators $\mathcal{P}_N k_c^{-2} I|_{\operatorname{im}\mathcal{P}_N}$ are invertible, and the $[\mathcal{P}_N k_c^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}$ are bounded

uniformly w.r.t. N. We obtain discrete eigenvalues $\lambda_{c,n,N}$ and eigenfunctions $f_{c,n,N}$ replacing M_c by $M_{c,N}$ s.t. (cf.(3.4))

$$[\mathcal{P}_N^c k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N^c}]^{-1}\partial\,[\mathcal{P}_N^c k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N^c}]\partial\, f_{c,n,N} + \left[[\mathcal{P}_N^c k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N^c}]^{-1} + \lambda_{c,n,N}^2 I\right] f_{c,n,N} = 0, \quad (7.3)$$

where $\sigma_{M_{c,N}} = \{\pm\lambda_{c,n,N} : n = 1, \cdots, 2N+1\}$ and $f_{c,n,N} \in \mathrm{im}\,\mathcal{P}_N^c \subset H_\alpha^1(\Gamma_c)$. Note that in the case of TE polarization, we get $\Delta u + k^2 u = 0$ in part a) of (2.1) and, instead of (7.3), we have the simpler discretized eigenvalue equation

$$\partial^2 f_{c,n,N} + \left[[\mathcal{P}_N^c k_c^2 I|_{\mathrm{im}\,\mathcal{P}_N^c}] f_{c,n,N} + \lambda_{c,n,N}^2 I\right] f_{c,n,N} = 0. \quad (7.4)$$

Similarly to (5.1), we get

$$D_t N_{c,N}^{\pm}\left\{\sum_{n=1}^{2N+1} \xi_n f_{c,n,N}\right\} = \pm \sum_{\substack{n=1 \\ \lambda_{c,n,N}\neq 0}}^{2N+1} \xi_n \lambda_{c,n,N} f_{c,n,N} \pm \sum_{\substack{n=1 \\ \lambda_{c,n,N}= 0}}^{2N+1} \xi_n f_{c,n,N}, \quad \xi_n \in \mathbb{C} \quad (7.5)$$

and, similarly to Equ. (5.2), we can identify the Dirichlet data $u_{c,N}^{\pm} = \sum_{n=1}^{2N+1} \xi_n f_{c,n,N}$ for upgoing and downgoing waves, respectively, with couples of discrete Dirichlet and Neumann data s.t.

$$\begin{aligned} u_{c,N}^{\pm} &\leftrightarrow (u_{c,N}^{\pm}, [\mathcal{P}_N^c k_c^{-2}|_{\mathrm{im}\,\mathcal{P}_N^c}]\partial_{x_2} u_{c,N}^{\pm}),\ k_c := k(\cdot, c+0), \\ H_\alpha^{1/2}(\Gamma_c) \supset \mathrm{im}\,\mathcal{P}_N^c &\leftrightarrow [\mathrm{im}\,\mathcal{P}_N^c]_{\pm}(\Gamma_c) \subseteq \mathrm{im}\,\mathcal{P}_{N,N}^c \subset H_\alpha^{1/2}(\Gamma_c) \times H_\alpha^{-1/2}(\Gamma_c). \end{aligned} \quad (7.6)$$

Like in Lemma 5.2 the full space $\mathrm{im}\,\mathcal{P}_{N,N}^c$ is the direct sum of these two subspaces, and we denote the projection onto the upgoing waves parallel to the downgoing waves by $P_{c,N}^+$ and $I - P_{c,N}^+$ by $P_{c,N}^-$. So the space of Dirichlet data $\mathrm{im}\,\mathcal{P}_N^c$ of upgoing and downgoing waves is identified with $\mathrm{im}\,P_{c,N}^+ \subset \mathcal{P}_{N,N}^c$ and $\mathrm{im}\,P_{c,N}^- \subset \mathcal{P}_{N,N}^c$, respectively. Replacing the Helmholtz equation by the discretized operator valued ODE and using the just mentioned discretized splitting into upgoing and downgoing waves, we can consider the discretized BVP (2.1) over the full grating and over each slice of Sect. 5. We get a discretized solution operator (scattering matrix) S_N^{ab} mapping $\mathrm{im}\,P_{a,N}^+ \otimes \mathrm{im}\,P_{b,N}^-$ into $\mathrm{im}\,P_{b,N}^+ \otimes \mathrm{im}\,P_{a,N}^-$

$$\begin{aligned} S_N^{ab} &= \begin{pmatrix} S_{++,N}^{ab} & S_{+-,N}^{ab} \\ S_{-+,N}^{ab} & S_{--,N}^{ab} \end{pmatrix}, \\ S_{++,N}^{ab} &= \left\{P_{b,N}^+ - P_{b,N}^+\left[P_{b-0,N}^- - T_{ab,N}^+ P_{a,N}^+ T_{ba,N}^-\right][D_{ab,N}^-]^{-1} P_{b,N}^-\right\} T_{ab,N}^+, \\ S_{+-,N}^{ab} &= \left[P_{b,N}^+ - P_{b,N}^+ T_{ab,N}^+ P_{a,N}^+ T_{ba,N}^-\right][D_{ab,N}^-]^{-1}, \\ S_{-+,N}^{ab} &= -P_{a,N}^- T_{ba,N}^- [D_{ab,N}^-]^{-1} P_{b,N}^- T_{ab,N}^+, \\ S_{--,N}^{ab} &= P_{a,N}^- T_{ba,N}^- [D_{ab,N}^-]^{-1}. \end{aligned} \quad (7.7)$$

Here the transition operators $T_{ab,N}^+$ and $T_{ba,N}^-$ are the solution operators of the discretized ODE

$$\partial_{x_2}[\mathcal{P}_N^{x_2} k^{-2} I|_{\mathrm{im}\,\mathcal{P}_N^{x_2}}]\partial_{x_2} u_N = L_{x_2,N} u_N,\ L_{x_2,N} u_N := -\partial_{x_1}[\mathcal{P}_N^{x_2} k^{-2} I|_{\mathrm{im}\,\mathcal{P}_N^{x_2}}]\partial_{x_1} u_N - u_N \quad (7.8)$$

corresponding to the initial value problems (6.1) and (6.2), respectively. The discretized operator $D_{ab,N}^- : \mathrm{im}\,P_{b-0,N}^- \to \mathrm{im}\,P_{b,N}^-$ for the operator D_{ab}^- used in the S-matrix formula (6.6) is defined as $D_{ab,N}^- := \{P_{b,N}^-|_{\mathrm{im}\,P_{b-0,N}^-} - P_{b,N}^- T_{ab,N}^+ P_{a,N}^+ T_{ba,N}^-\}$. Note that the projections $P_{c,N}^{\pm}$ can be computed by (7.5) and by (cf. (5.6))

$$\begin{aligned} P_{b,N}^{\pm}[u_{a,N}^+] &= \frac{1}{2}\left[u_{a,N}^+ \pm [D_t N_{b,N}^+]^{-1}\left[\mathcal{P}_N k_b^{-2} I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\left[\mathcal{P}_N k_a^{-2} I\big|_{\mathrm{im}\,\mathcal{P}_N}\right] D_t N_{a,N}^+ u_{a,N}^+\right], \\ P_{b,N}^{\pm}[u_{a,N}^-] &= \frac{1}{2}\left[u_{a,N}^- \mp [D_t N_{b,N}^+]^{-1}\left[\mathcal{P}_N k_b^{-2} I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\left[\mathcal{P}_N k_a^{-2} I\big|_{\mathrm{im}\,\mathcal{P}_N}\right] D_t N_{a,N}^+ u_{a,N}^-\right]. \end{aligned} \quad (7.9)$$

We shall discuss the computation of the discretization $T^+_{ab,N}$ and $T^-_{ba,N}$ next. If this is done, then (7.7) enables us to compute S^{ab}_N. Approximating $S^j = S^{h_{j-1}h_j}$ by discretized $S^j_N = S_N^{h_{j-1}h_j}$ and the incoming waves u^-_n and u^+_0 by their truncations $\mathcal{P}_N u^-_n$ and $\mathcal{P}_N u^+_0$, we can perform the SMA (5.13) on the discrete level.

Finally, we have to fix how to compute the transition matrices $T^+_{ab,N}$ and $T^-_{ba,N}$. We can use any integration algorithm for initial problems of ordinary differential operators (cf. the discussion of this point e.g. in [13,15], where the resulting scattering matrix algorithm is called FMM). Unfortunately, fast explicit methods are often not stable, and the integration error blows up for large widths $h_j - h_{j-1}$ of the slice. To overcome this problem, the widths $h_j - h_{j-1}$ are reduced by increasing the number of slices, and with thin slices a stable algorithm is achieved. Of course, the computing time increases with an increasing number of slices.

An alternative method is the classical approach of the RCWA (cf. [11]). **Firstly**, we suppose that the wavenumber in the slice is independent of x_2, i.e. $k(x_1, x_2) = k(x_1, a)$. Then $P^+_{b-0,N} = P^+_{a,N}$ and the eigenfunction decomposition can be applied. E.g. for T^+_{ab}, we use the identification (7.6) of $u^+_{a,N}$ with $(u^+_{a,N}, [\mathcal{P}_N k_a^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}] D_t N^+_{a,N} u^+_{a,N})$ and arrive at

$$T^+_{ab,N}\left\{\sum_{n=1}^{2N+1}\xi_n f_{a,n,N}\right\} = \sum_{\substack{n=1\\ \lambda_{a,n,N}\neq 0}}^{2N+1}\left[e^{\mathbf{i}\lambda_{a,n,N}(b-a)}\xi_n\right] f_{a,n,N} + \sum_{\substack{n=1\\ \lambda_{a,n,N}=0}}^{2N+1}\left[\big(1+(b-a)\big)\xi_n\right] f_{a,n,N},$$

for any $\xi_n \in \mathbb{C}$. Note that this formula reveals the importance of the decomposition of the waves into upgoing and downgoing ones. For an improper decomposition, there would appear coefficients $[e^{-\lambda_{a,n,N}(b-a)}\xi_n]$ blowing up with larger width $b-a$ and with $n\to\infty$.

Secondly, if the wavenumber k of the slice is dependent on x_2, then we split the slice into the union of very thin subslices and approximate k over each subslice by a wavenumber function independent on x_2. For example consider an echelle grating like in Fig. 6, where two layers cover the lower boundary Γ_a and a triangle is set upon the upper layer, surrounded by the turquoise lines and consisting of the same material as the upper layer. Then the corresponding wavenumber function can be approximated by the staircase geometry indicated by the additional blue layers. Replacing the slice by the union of subslices, we have an approximate geometry, for which the case of x_2-independent wavenumbers applies. Of course the price for this solution is an extra numerical error due to the approximation of the wavenumber and an increased computing time due to the increased number of slices.

Altogether, the parameters of discretization are the following.

- The first parameter is the truncation parameter N in (7.1).
- To get the Galerkin operators in (7.2), the Fourier coefficients of the reciprocal squared wavenumber function k^{-2} must be computed. In general, this requires a quadrature of the integrals for the Fourier coefficients. So the second discretization parameters are those of quadrature.
- The third parameter is the stepsize of the slicing $h := \min_{j=1,\cdots,n}[h_j - h_{j-1}]$ (cf. Fig. 5).
- If the FMM is applied, then the matrices $T^+_{ab,N}$ and $T^-_{ba,N}$ are computed by a numerical integration of initial value problems for an ODE, i.e., by a FDM. So the last discretization parameter is the stepsize of such a FDM.

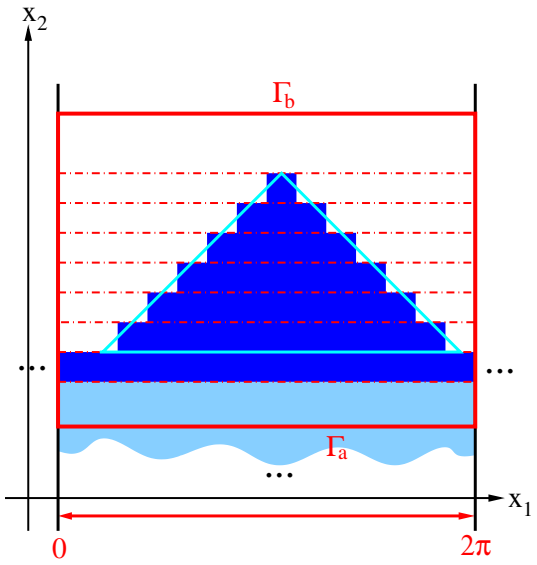

Figure 6: Staircase example for subslices to approximate an x_2-depending wavenumber function.

8 Analysis of convergence

Before we start, we have to comment on the analysis in [4]. In this paper it is used that the RCWA is equivalent to a discretized variational equation for the standard variational equation with the wavenumber function replaced by an approximation, which is piecewise constant w.r.t. x_2, and with a trial space

$$\operatorname{span}\left\{x_1 \mapsto e^{\mathbf{i}(l+\alpha)x_1} : \, l = -N, \cdots, N\right\} \otimes H^1(a,b),$$

i.e., the space of truncated Fourier series with x_2-dependent coefficients. At the first glance, the paper seems to be disappointing since this equivalence assumes (tacit assumption in the proof of [4, Thm. 7])

- For the S-matrix computation, the integration of the ODE is exact:
  This is acceptable if the RCWA with (6.6) is used. This might be not acceptable for the FMM.
- On the common boundary between consecutive slices, the boundary data for the upgoing and downgoing are identified, i.e. all the operations in the iteration steps are exact:
  Hence, the error propagation in the iteration is neglected in this first step of analysis. Such a propagation analysis would be of interest for the stepsize of the slicing tending to zero, which is not treated in the current paper either.
- All matrices, for which the inverse is required in the algorithm, are supposed to be invertible:
  In particular, this requires the invertibility of the D^-_{ab} for the computation of the scattering matrices via (6.6) and that of D for the iteration step in (5.12).

However, these assumptions mean that some errors are neglected, but others are treated by a hard and deep analysis. So an estimate for the approximation error independent of the algorithmic implementation is provided and, therewith, a lower estimate for the convergence with $\max_j |h_j - h_{j-1}| \to 0$ and with $N \to \infty$. Moreover, the general RCWA is based on an EVD. For this, the asymptotic analysis of the convergence of eigenvalues and eigenfunctions is extremely difficult. Observing empirically computation errors of the EVD less than a small threshold, it is natural to neglect the algorithmic errors due to EVD. Finally, note that in [4, Equ. (56)] there appears a monotonicity condition on the electric permittivity, called non-trapping conditions. Though this is only a sufficient condition, it is a clever assumption to exclude trapped eigenmodes, i.e. to guarantee condition ii) of Thm. 4.2 for all possible slices. Without this, the validity of ii) remains open unless an absorbing material is involved.

In the current paper, the grating is supposed to be the union of a fixed finite number of slices with wavenumber function independent of x_2 inside each slice. For the RCWA no finer slicing is needed, and it remains to analyze the convergence for $N\to\infty$. We start with $N=\infty$, i.e., we consider the SMA iteration on a continuous level with no truncation of the Fourier series. From the derivation of (6.7) and of the iteration (cf. Sect. 5), from Thm. 4.2, and from the Lemmata 5.3 and 6.1, we conclude (compare [7, Thm. 6.3] for the TE polarization)

Theorem 8.1. *Suppose the slicing is fixed s.t., for $j=1,\cdots,n$, the wavenumber functions $k(x_1,x_2)$ are independent of x_2 in $h_{j-1}\le x_2<h_j$. Further, for $j=1,\cdots,n$ suppose the multiplication operators $k_{h_j}^{-2}I\in\mathcal{L}(H_\alpha^{-1/2})$ and the S-matrices $S^{h_{j-1}h_j}$ and $S^{h_0h_j}$ are bounded operators (cf. Thm. 4.2). Choose any pair (u_a^+,u_b^-) with $u_a^+\in\operatorname{im}P_a^+=H_\alpha^{1/2}(\Gamma_a)$ and $u_b^-\in\operatorname{im}P_b^-=H_\alpha^{1/2}(\Gamma_b)$. Then the iterative SMA method (5.13), where the S-matrices are computed by (6.7) and the transition operators by exact integration of the initial value problems (6.1) and (6.2) (e.g. by (6.8)), yields the outgoing waves $u_b^+\in\operatorname{im}P_b^+=H_\alpha^{1/2}(\Gamma_b)$ and $u_a^-\in\operatorname{im}P_a^-=H_\alpha^{1/2}(\Gamma_a)$, i.e. the true solutions for the scattering problem (2.1).*

Again we note that no boundedness of any multiplication operator $k_c^{-2}I$ in the space $H_\alpha^{-1/2}$ is needed for the case of TE polarization.

Next we fix the slicing and look at the RCWA discretization with finite truncation index N tending to infinity. The first question is, how to deal with the EVD. From to general theory of approximate eigenvalue computation of operators by computing the eigenvalues of approximate operators (cf. e.g. [17]), it seems natural to require the following property of the EVD algorithm applied in the SMA:

Assumption on approximation of EVD:

$$\begin{aligned}&\text{For operator } k^2L\colon L^2\hookrightarrow L^2 \text{ and for the approximate operators}\\&A_N:=[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\mathcal{P}_NL|_{\operatorname{im}\mathcal{P}_N}\in\mathcal{L}\big(\operatorname{im}\mathcal{P}_N\big) \text{ suppose}\\&A_N\mathcal{P}_N\to k^2L\in\mathcal{L}(H_\alpha^1[0,2\pi],H_\alpha^{-1}[0,2\pi]). \text{ For the eigenfunctions}\\&\text{with } k^2Lf_n=\lambda_nf_n,\ n\in\mathbb{Z}_+ \text{ and } A_Nf_{n,N}=\lambda_{n,N}f_{n,N},\ n=1,\cdots,2N+1\\&\text{suppose that, for any } \varepsilon>0 \text{ and } n_1\in\mathbb{Z}_+,\ \text{there is an } N_0=N_0(\varepsilon,n_1),\ \text{s.t.}\\&|\lambda_n-\lambda_{N,n}|\le\varepsilon \text{ and } \|f_n-f_{n,N}\|_{H_\alpha^{1/2}}\le\varepsilon,\ 1\le n\le\min\{n_1,2N+1\}\\&\text{for any } N \text{ with } N_0\le N.\end{aligned}\tag{8.1}$$

Tacitly, we assume that all these eigenfunctions are of rank one.

To prepare the proof on the convergence of the discretized RCWA, we recall a well-known result on the approximation of eigenvalues (cf. e.g. [17, Sect. 4.2]) and present two lemmata on the stable convergence of the Dirichlet-to-Neumann maps. Consider operators $E,F\in\mathcal{L}(H,H^*)$ in the Hilbert spaces H and its dual H^*. We suppose $H\subseteq L^2\subseteq H^*$ s.t. the duality between H and H^* is an extension of the L^2 scalarproduct. Suppose there are orthogonal projections $\mathbb{P}_N\in\mathcal{L}(L^2)$ strongly converging to the identity for $N\to\infty$ s.t. the $\mathbb{P}_N$ converge strongly in H and H^* as well. Consider approximate operators $E_N,F_N\in\mathcal{L}(\operatorname{im}\mathbb{P}_N,\operatorname{im}\mathbb{P}_N^*)$ s.t. the strong convergences $E_N\mathbb{P}_N\to E$ and $F_N\mathbb{P}_N\to F$ hold. Eigenfunctions f_N,f and eigenvalues λ_N,λ are defined by the equations $[E_N+F_N-\lambda_NI_N]f_N=0$ and $[E+F-\lambda I]f=0$, respectively. Finally, recall that the converge $E_N\mathbb{P}_N\to E$ is called stable if there is an $N_0\in\mathbb{Z}_+$ s.t. $\sup_{N\ge N_0}\|E_N^{-1}\mathbb{P}_N\|_{\mathcal{L}(H^*,H)}<\infty$. If E is invertible, then $E_N\mathbb{P}_N\to E$ is stable if and only if $E_N^{-1}\mathbb{P}_N\to E^{-1}$. The convergence $F_N\mathbb{P}_N\to F$ is called compact if, for any bounded sequence $x_N\in H$, the closure of the set $\{F_N\mathbb{P}_Nx_N\colon N\in\mathbb{Z}_+\}$ is compact. Equivalently, the convergence is compact if, for any bounded sequence $x_N\in H,\ N\in\mathbb{Z}_+$,

for any subsequence $x_N,\ N\in\mathbb{N}'\subset\mathbb{Z}_+$, and for any small positive number ε, there is a subsequence $x_N,\ N\in\mathbb{N}''\subset\mathbb{N}'$ and an $x\in H$ s.t. $\|F_N\mathbb{P}_N x_N - x\|\le\varepsilon$ holds for $N\in\mathbb{N}''$.

Theorem 8.2. *Suppose:*

i) There is a domain $\Lambda\subset\mathbb{C}$ s.t., for each $\lambda\in\Lambda$, the operator $[E+F-\lambda I]$ is a Fredholm operator of index zero.

ii) There is a $\lambda^{\#}\in\Lambda$ s.t. the operator $[E+F-\lambda^{\#}I]\in\mathcal{L}(H,H^*)$ is invertible.

iii) The convergence $[E_N-\lambda I_N]\to[E-\lambda I]$ is stable for any $\lambda\in\Lambda$. The convergence $F_N\to F$ is compact.

Then there holds:

a) For a sequence λ_N of eigenvalues for $[E_N+F_N-\lambda_N I_N]f_N=0$ with a limit $\lambda_N\to\lambda$, the limit λ is an eigenvalue for $[E+F-\lambda I]f=0$.

b) For any eigenvalue λ for $[E+F-\lambda I]f=0$, there exists eigenvalues λ_{N_k} and eigenfunctions f_{N_k} with $[E_{N_k}+F_{N_k}-\lambda_{N_k}I_{N_k}]f_{N_k}=0$ s.t., for $k\to\infty$, we get $\lambda_{N_k}\to\lambda$ and $f_{N_k}\to f$ in H, where f is an eigenfunction with $[E+F-\lambda I]f=0$.

If the Conditions i),ii), and iii) hold for the domain Λ replaced by $\Lambda=\{0\}$, then the convergence $[E_n+F_n]\mathbb{P}_N\to[E+F]$ is stable.

Now fix an x_2-coordinate c and denote by k_c the restriction $k_c(x_1):=k(x_1,c)$ of the wavenumber function $k(x_1,x_2)$. Recall the definition (7.3) for the eigenvalues $\lambda_{c,n,N}$ and eigenfunctions $f_{c,n,N}$. To simplify the formulas, we assume $\lambda_{c,n,N}\neq0$ s.t. we get the Dirichlet-to-Neumann maps and their inverses by (cf. (7.5))

$$\begin{aligned}
[D_tN_c^+]^{\pm1}\Big(\sum_{n\in\mathbb{Z}_+}\xi_n f_{c,n}\Big) &= \sum_{n\in\mathbb{Z}_+}\sqrt{\lambda_{c,n}}^{\pm1}\xi_n f_{c,n},\\
[D_tN_{c,N}^+]^{\pm1}\Big(\sum_{n=1}^{2N+1}\xi_n f_{c,n,N}\Big) &= \sum_{n=1}^{2N+1}\sqrt{\lambda_{c,n,N}}^{\pm1}\xi_n f_{c,n,N},\quad \xi_n\in\mathbb{C}.
\end{aligned}\tag{8.2}$$

Note that $D_tN_{c,N}^- = -D_tN_{c,N}^+$ and $D_tN_c^- = -D_tN_c^+$. So we only have to consider the Dirichlet-to-Neumann operators with plus sign. We shall say that the basis $f_{c,n,N},\ n=1,\cdots,2N+1$ satisfies the Riesz property in H^s_α uniformly if there is a constant $C\ge1$ independent of N s.t., for all $\xi_n\in\mathbb{C}$,

$$\frac{1}{C}\sqrt{\sum_{n=1}^{2N+1}(1+|\lambda_{c,n,N}|^2)^s|\xi_n|^2}\le\Big\|\sum_{n=1}^{2N+1}\xi_n f_{c,n,N}\Big\|_{H^s_\alpha(\Gamma_c)}\le C\sqrt{\sum_{n=1}^{2N+1}(1+|\lambda_{c,n,N}|^2)^s|\xi_n|^2}.$$

Lemma 8.3. *Assume one of the two conditions on k_c (cf. the assumptions on k in Sect. 3):*

a1) The function k_c is twice continuously differentiable and, for each $x_1\in[0,2\pi]$, there hold the relations $\Re e\,k_c(x_1)>0$ and $\Im m\,k_c(x_1)\ge0$. The basis $f_{c,n,N},\ n=1,\cdots,2N+1$ satisfies the Riesz property in $L^2=H^0_\alpha$ uniformly.

a2) The function k_c is piecewise twice continuously differentiable and there exists a positive constant c_k s.t. $k_c\ge c_k$.

Then, for $-1\le s\le1$, the system of eigenfunctions $f_{c,n,N},\ n=1,\cdots,2N+1$ forms a uniform Riesz basis in $\operatorname{im}\mathcal{P}_N\subset H^s_\alpha$. The strong convergences $[D_tN_{c,N}^+]^{\pm1}\mathcal{P}_N\to[D_tN_c^+]^{\pm1}\in\mathcal{L}(H^{\pm1/2}_\alpha,H^{\mp1/2}_\alpha)$ hold. Hence, the convergence $[D_tN_{c,N}^+]\mathcal{P}_N\to[D_tN_c^+]$ is stable.

Proof. At first, we consider the case of **Assumption a2)**. Similarly to the Riesz property of the basis $f_{c,n}$, $n\in\mathbb{Z}_+$, we get the uniform Riesz property in H^s_α for the bases $f_{c,n,N}$, $n=1,\cdots,2N+1$ and the Sobolev index $s\in[-1,1]$. Indeed, for any $\tilde g_N\in\operatorname{im}\mathcal{P}_N$, the selfadjointness of $\mathcal{P}_N$ and (7.3) leads us to

$$\begin{aligned}\left\langle\partial_{x_1}\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}\partial_{x_1}f_{c,n,N}+f_{c,n,N},\tilde g_N\right\rangle &= -\lambda^2_{c,n,N}\left\langle\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}f_{c,n,N},\tilde g_N\right\rangle,\\ \left\langle\partial_{x_1}k^{-2}\partial_{x_1}f_{c,n,N}+f_{c,n,N},\tilde g_N\right\rangle &= -\lambda^2_{c,n,N}\left\langle k^{-2}f_{c,n,N},\tilde g_N\right\rangle.\end{aligned}$$

Setting $h_{c,n,N}:=k^{-1}f_{c,n,N}\in\operatorname{im}[k^{-1}\mathcal{P}_N]$ and $g_N:=k^{-1}\tilde g_N\in\operatorname{im}[k^{-1}\mathcal{P}_N]$, we arrive at

$$\left\langle k\partial_{x_1}k^{-2}\partial_{x_1}kh_{c,n,N}+k^2h_{c,n,N},g_N\right\rangle = -\lambda^2_{c,n,N}\left\langle h_{c,n,N},g_N\right\rangle. \tag{8.3}$$

Consequently, the eigenfunctions $h_{c,n,N}$ of a selfadjoint operator are orthogonal. Thus we get the uniform Riesz property of the $f_{c,n,N}$ in L^2. For $h_N:=\sum_n\xi_nh_{c,n,N}$, we get (cf. (3.12))

$$\begin{aligned}a(h_N,h_N) &= -\sum_{n,m}\xi_n\bar\xi_m\left\langle[k\partial_{x_1}k^{-2}\partial_{x_1}kI+k^2I]h_{c,n,N},h_{c,m,N}\right\rangle\\ &= \sum_{n,m}\xi_n\bar\xi_m\lambda^2_{c,n,N}\left\langle h_{c,n,N},h_{c,m,N}\right\rangle = \sum_n\lambda^2_{c,n,N}|\xi_n|^2.\end{aligned}$$

Now follow the arguments in the proof of Lemma 3.3: Using the L^2 Riesz property and the last estimate, we get the uniform Riesz property in H^1_α. By interpolation and duality arguments we get the uniform Riesz property in H^s_α for $-1\le s\le1$.

Obviously, the uniform Riesz property of the $f_{c,n,N}$ in $H^{\pm1/2}_\alpha$ implies the uniform boundedness of the operators $[D_tN^+_{c,N}]^\pm$. It remains to prove the strong convergence on a dense subset. We shall prove the convergence on the set of basis functions $\{f_{c,m}:m\in\mathbb{Z}_+\}$. Fix an m and the corresponding $f_{c,m}$. Due to the orthonormality of the functions $k^{-1}f_{c,n,N}=:h_{c,n,N}$ and due to the Riesz property

$$\begin{aligned}\mathcal{P}_Nf_{c,m} &= \sum_{n=1}^{2N+1}\xi_{n,N}f_{c,n,N},\ \xi_{n,N}:=\left\langle k^{-1}\mathcal{P}_Nf_{c,m},h_{c,n,N}\right\rangle=\left\langle k^{-2}\mathcal{P}_Nf_{c,m},k^{-2}f_{c,n,N}\right\rangle,\\ 1\sim\|f_{c,m}\|^2_{H^s_\alpha} &\sim \sum_{n=1}^{2N+1}(1+|\lambda_{c,n,N}|^2)^s|\xi_{n,N}|^2.\end{aligned}$$

Consequently, for a fixed n_0 independent of N, we get, for the coefficients $\xi_{n,N}$ of $\mathcal{P}_Nf_m$,

$$\begin{aligned}\mathcal{P}_Nf_m &= \sum_{n=1}^{2N+1}\xi_{n,N}f_{c,n,N},\\ [D^+_{c,N}]^{\pm1}\left(\sum_{n=1}^{2N+1}\xi_{n,N}f_{c,n,N}\right) &= \sum_{n=1}^{2N+1}\lambda^{\pm1}_{c,n,N}\xi_{n,N}f_{c,n,N},\qquad(8.4)\\ \left\|[D^+_{c,N}]^{\pm1}\mathcal{P}_Nf_{c,m}-\lambda^{\pm1}_{c,m}f_{c,m}\right\|^2_{H^{\mp1/2}_\alpha} &\sim \left\|\sum_{n=1}^{n_0}\lambda^{\pm1}_{c,n,N}\xi_{n,N}f_{c,n,N}-\lambda^{\pm1}_{c,m}f_{c,m}\right\|^2_{H^{\mp1/2}_\alpha}\\ &\quad+\sum_{n=n_0+1}^{2N+1}(1+|\lambda_{c,n,N}|^2)^{\pm1/2}\,|\xi_{n,N}|^2.\qquad(8.5)\end{aligned}$$

Here the squared Sobolev norm on the right-hand side tends to zero, since we have $\lambda_{c,n,N}\to\lambda_{c,n}$ and $f_{c,n,N}\to f_{c,n}$, which implies $\xi_{n,N}\to\left\langle k^{-2}f_{c,m},f_{c,n}\right\rangle=\left\langle h_{c,m},h_{c,n}\right\rangle=\delta_{n,m}$. Note that the convergence

of eigenvalues and eigenfunctions follows by applying Thm. 8.2. For this, $E+F$ is the operator of the variational form in (3.12) mapping $H=k^{-1}H^1_\alpha$ into $H^*=kH^{-1}_\alpha$. Since the variational form is strongly elliptic, we have a splitting into the sum of an invertible operator E plus a compact operator F. Choosing $\mathbb{P}_N$ to be the L^2 orthogonal projection onto the space $\operatorname{im} k^{-1}\mathcal{P}_N$, the approximate operators are the Galerkin approximations (cf. (8.3)). The assumption in (8.1) guarantees that the error of the EVD algorithm does not disturb the convergence of the eigenvalues and eigenfunctions.

The sum on the right-hand side of (8.5) can be estimated as

$$\begin{aligned}\sum_{n=n_0+1}^{2N+1}(1+|\lambda_{c,n,N}|^2)^{\pm1/2}\,|\xi_{n,N}|^2 &\le \sum_{n=n_0+1}^{2N+1}(1+|\lambda_{c,n,N}|^2)\,|\xi_{n,N}|^2 \sup_{n_0<n\le 2N+1}(1+|\lambda_{c,n,N}|^2)^{-1}\\ &\le \sup_{n_0<n\le 2N+1}(1+|\lambda_{c,n,N}|^2)^{-1}\|f_{c,m}\|^2_{H^1_\alpha},\end{aligned}\tag{8.6}$$

which is small if n_0 is fixed such that the supremum is small. We have to show that the supremum is less than any prescribed $\varepsilon>0$ provided n_0 is large enough. Indeed, suppose the contrary. W.l.o.g. we assume an ordering of the eigenvalues $|\lambda_{c,n,N}|\le|\lambda_{c,n+1,N}|$ and $|\lambda_{c,n}|\le|\lambda_{c,n+1}|$ for $n=1,\cdots$. Suppose there is a $C>0$ s.t. $|\lambda_{c,n_N,N}|\le C$ for an $n_N\le 2N+1$ with $n_N\to\infty$. Due to the convergence $\lambda_{c,n}\to\infty$, there is an $\tilde n$ such that $\lambda_{c,\tilde n}>C+1$. However, from Thm. 8.2 we get a sequence N_k such that $\lambda_{c,n,N_k}\to\lambda_{c,n}$, $k\to\infty$ for all $n\le\tilde n$. For $n_{N_k}\ge\tilde n$, this leads to the contradiction $|\lambda_{c,\tilde n,N_k}|\le|\lambda_{c,n_{N_k},N_k}|\le C<C+1\le\lambda_{c,\tilde n}$.

Now consider **Assumption a1)**. The proof is similar to that for a2). We only mention how to get the uniform Riesz estimate and how to derive the convergence of the eigenvalues and eigensolutions. For the latter, we observe that the main part of the differential operator for the EVD and its discretization admit the following splitting

$$\begin{aligned}k^2\partial_{x_1}k^{-2}\partial_{x_1}f &= \partial^2_{x_1}f-2k^2k^{-3}k'\partial_{x_1}f,\ f\in H^{1/2}_\alpha,\\ [\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\partial_{x_1}&[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}f_N\\ &= [\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\mathcal{P}_N\partial_{x_1}k^{-2}\partial_{x_1}f_N\\ &= [\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}\partial^2_{x_1}f_N\\ &\quad -2[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\mathcal{P}_Nk^{-3}k'\partial_{x_1}f_N\\ &= \partial^2_{x_1}f_N-2[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}[\mathcal{P}_Nk^{-3}k'I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}f_N,\\ &\quad f_N\in\operatorname{im}\mathcal{P}_N.\end{aligned}$$

The compact convergence of $2[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}[\mathcal{P}_Nk^{-3}k'|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}\mathcal{P}_N$ to the compact operator $2k^2k^{-3}k'\partial_{x_1}\in\mathcal{L}(H^{1/2}_\alpha,H^{-1/2}_\alpha)$ enables the application of Thm. 8.2, and the convergence of the eigenvalues and eigenfunctions follows.

For the Riesz property, we derive from the eigenvalue equation (7.3) that

$$\begin{aligned}[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\big[\partial_{x_1}[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}-I_N\big]f_{c,n,N}&\\ = -2[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{-1}f_{c,n,N}-\lambda_{c,n,N}f_{c,n,N}&\end{aligned}\tag{8.7}$$

The theory on the approximate solution to variational equations provides us with the N-uniform estimates

$$\begin{aligned}\big\|[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]^{\pm1}\big\|_{\mathcal{L}(H^0_\alpha,H^0_\alpha)} &\le C,\\ \Big\|\big[\partial_{x_1}[\mathcal{P}_Nk^{-2}I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}-I_N\big]^{\pm1}\Big\|_{\mathcal{L}(H^{\pm1}_\alpha,H^{\mp1}_\alpha)} &\le C.\end{aligned}$$

Again this theory together with the differentiability of k_c as well as the approximation and the inverse property of the spaces $\operatorname{im}\mathcal{P}_N$, yields

$$\begin{aligned}\left\|[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{\pm 1}\right\|_{\mathcal{L}(H^2_\alpha,H^2_\alpha)} &\le C,\\ \left\|\left[\partial_{x_1}[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}-I_N\right]\right\|_{\mathcal{L}(H^2_\alpha,H^0_\alpha)} &\le C,\\ \left\|\left[\partial_{x_1}[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}-I_N\right]^{-1}\right\|_{\mathcal{L}(H^0_\alpha,H^2_\alpha)} &\le C.\end{aligned}$$

We consider a general function $f_N=\sum_n \xi_n f_{c,n,m}\in\operatorname{im}\mathcal{P}_N$ and arrive at

$$\begin{aligned}\|f_N\|_{H^2_\alpha} &\sim \left\|[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\left[\partial_{x_1}[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}-I_N\right] f_N\right\|_{L^2}\\ &\sim \left\|\left[[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\partial_{x_1}[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}+[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\right] f_N\right\|_{L^2}\\ &\quad +\mathcal{O}(\|f_N\|_{L^2})\\ &\sim \left\|\sum_{n=1}^{2N+1}\xi_n\left[[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\partial_{x_1}[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\partial_{x_1}+[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\right] f_{c,n,N}\right\|_{L^2}\\ &\quad +\mathcal{O}(\|f_N\|_{L^2})\,.\end{aligned}$$

Using (8.7) and the uniform Riesz basis property in L^2, we continue

$$\|f_N\|_{H^2_\alpha} \sim \left\|\sum_{n=1}^{2N+1}\lambda^2_{c,n,N}\xi_n f_{c,n,N}\right\|_{L^2}+\mathcal{O}(\|f_N\|_{L^2}) \sim \sqrt{\sum_{n=0}^{2N+1}(1+|\lambda_{c,n,N}|^2)^2|\xi_n|^2},$$

which is the uniform Riesz property in H^2_α. By interpolation and duality arguments, we get the uniform Riesz property in H^s_α for $-2\le s\le 2$. □

In accordance with the last proof, the approximations $[D_t N^+_{c,N}]^{\pm 1}$ (cf. (8.4)) of the Dirichlet-to-Neumann mapping and their inverses converge if the Riesz estimate on the discrete level is uniform and if the discretized eigenvalues and eigenfunctions converge to those of the continuous level in the sense of Thm. 8.2.

Remark 8.4. *For the case of the TE polarization (cf. (7.4)), the assertion of Lemma 8.3 holds for any piecewise twice continuously differentiable k_c with $\Re e\, k_c>\varepsilon>0$ and $\Im m\, k_c\ge 0$ over $[0,2\pi]$.*

Remark 8.5. *If we have a uniform asymptotics of the form $\lambda_{c,n,N}=n\pm\alpha+c_a/n+\mathcal{O}(1/n^2)$, then it is not hard to prove the norm convergence $\|D^+_{c\pm 0,N}-\mathcal{P}_N D^+_{c\pm 0}|_{\operatorname{im}\mathcal{P}_N}\|\to 0$ for $N\to\infty$.*

Unfortunately, we need more. We need to have a stable convergence of the sum of two Dirichlet-to-Neumann operators $D^+_{c\pm 0,N}$ defined with different restrictions $k(x_1)=k(x_1,c\pm 0)$ of the wavenumber function. We guess that this is true. On the continuous level, both operators are strongly elliptic in the same manner s.t. also the sum is strongly elliptic, and together with a trivial null space for the sum operator the invertibility of the sum follows. In this spirit, if we could split the operators on the discretization level into strongly elliptic operators plus a compactly converging remainders, then we would obtain stable convergence for the sum. Unfortunately, we could not show this. We can only prove

Lemma 8.6. *Suppose the real-valued wavenumber function $k_{c\pm 0}$ are Hölder continuous and piecewise twice continuously differentiable and that there is a positive constant $c_k>0$ s.t. $k_{c\pm 0}\ge c_k$. If*

the sum operator $[k_{c+0}^{-2}D_tN_{c+0}^+ + k_{c-0}^{-2}D_tN_{c-0}^+] \in \mathcal{L}(H_\alpha^{1/2}, H_\alpha^{-1/2})$ *is invertible, then the approximate operators* $\{[\mathcal{P}_N k_{c+0}^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c+0,N} + [\mathcal{P}_N k_{c+0}^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c-0,N}\}$ *are invertible for* N *sufficiently large. The sequence* $\{[\mathcal{P}_N k_{c+0}^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c+0,N} + [\mathcal{P}_N k_{c+0}^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c-0,N}\}^{-1}\mathcal{P}_N$ *converges strongly to the operator* $[k_{c+0}^{-2}D_tN_{c+0}^+ + k_{c-0}^{-2}D_tN_{c-0}^+]^{-1}$. *In other words, the convergence of the approximate operators* $\{[\mathcal{P}_N k_{c+0}^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c+0,N} + [\mathcal{P}_N k_{c+0}^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c-0,N}\}\mathcal{P}_N$ *to the operator* $[k_{c+0}^{-2}D_tN_{c+0}^+ + k_{c-0}^{-2}D_tN_{c-0}^+]$ *is stable.*

Proof. The eigenvalue equation (7.3) implies that the functions $[\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]^{1/2} f_{c,n,N}$ are the orthogonal eigenvalues of the selfadjoint operator

$$[\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1/2}\partial[\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]\partial[\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1/2} + [\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1}.$$

Consequently, the functions $f_{c,n,N}$ form an orthogonal basis in $\mathrm{im}\,\mathcal{P}_N$ w.r.t. the weighted L^2 scalar product $\langle k_c^{-2}\cdot,\cdot\rangle = \langle [\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]\cdot,\cdot\rangle$. So we arrive at (cf. (8.4))

$$\left\langle [\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c,N}^+ \sum_{n=1}^{2N+1} \xi_n f_{c,n,N}, \sum_{n=1}^{2N+1} \xi_n f_{c,n,N} \right\rangle \;=\; \sum_{n}^{2N+1} \lambda_{c,n,N}|\xi_n|^2.$$

For a fixed $\varepsilon>0$, there are $n_1, N_1 \in \mathbb{Z}_+$ s.t. $\lambda_{c,n,N} \geq \varepsilon$ for $n \leq n_1$ and all $N \geq N_1$. Thus the operator $[\mathcal{P}_N k_c^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c,N}^+$ splits into an operator of rank less or equal to n_1 and an operator with positive definite real part greater or equal to constant times εI. The first operators corresponding to the $n \leq n_1$ converge compactly to an operator of rank less or equal to n_1, and the second operators converge to an operator with positive definite real part greater or equal to constant times εI. Since all these second operators have an inverse of norm less than constant times ε^{-1}, the second convergence is stable. Summing $c = c+0, c-0$ and applying Thm. 8.2, we get the assertion. □

Remark 8.7. *If the operator* $D_{c-0,c+0}^- := P_{c+0}^-|_{\mathrm{im}\,P_{c-0}^-}\colon H_\alpha^{1/2} = \mathrm{im}\,P_{c-0}^- \to \mathrm{im}\,P_{c+0}^- = H_\alpha^{1/2}$ *is invertible (cf. Lemma 6.1), then the operator* $[k_{c+0}^{-2}D_{c+0}^+ + k_{c-0}^{-2}D_{c-0}^+]$ *is invertible. Indeed, by (5.6) we have*

$$D_{c-0,c+0}^- = \frac{1}{2}\left[I + [D_tN_{c+0}^+]^{-1}\frac{k_{c+0}^2}{k_{c-0}^2}D_tN_{c-0}^+\right] = \frac{1}{2}[D_tN_{c+0}^+]^{-1}k_{c+0}^2\left[k_{c+0}^{-2}D_{c+0}^+ + k_{c-0}^{-2}D_{c-0}^+\right].$$

Now we consider the RCWA. Recall that this is the SMA of (5.13) applied with the operators $S^{h_{j-1}h_j}$ replaced by $S_N^{h_{j-1}h_j}$, which are computed by (7.7). For the ingredients of (7.7), we solve the EVD (7.3) for TM polarization resp. (7.4) for TE polarization to get the eigenvalues $\lambda_{h_j,n,N}$ and the corresponding eigenfunctions $f_{h_j,n,N} \in \mathrm{im}\,\mathcal{P}_N$. We get the operators $T_{h_{j-1}h_j,N}^+$ and $T_{h_jh_{j-1},N}^-$ by their matrix $(e^{\mathbf{i}(h_j-h_{j-1})\lambda_{h_{j-1},n,N}}\delta_{n,m})_{n,m=1}^{2N+1}$ w.r.t. the basis $f_{h_{j-1},n,N}$, $n=1,\cdots,2N+1$. We get the Dirichlet-to-Neumann maps by (7.5), the projections $P_{h_j,N}^\pm$ by (7.9) (cf. the subsequent (8.8)), and the $D_{h_{j-1}h_j,N}^-$ by the subsequent (8.9). Finally (cf. item ii) of (5.13)), the reflected and transmitted waves are obtained by applying $S_N^{h_0h_n}$ to the approximate boundary data of the incident waves $u_{h_0,N}^+ := \mathcal{P}_N^{h_0}u_{h_0}^+$ and $u_{h_n,N}^- := \mathcal{P}_N^{h_n}u_{h_n}^-$, respectively.

Theorem 8.8. *For the SMA (5.13) discretized as the RCWA defined in Sect. 7, suppose*

- *i) The slicing is fixed s.t., for* $j=1,\cdots,n$, *the wavenumber functions* $k(x_1,x_2)$ *are independent of* x_2 *in* $h_{j-1} \leq x_2 < h_j$.
- *ii) For the* x_2*-coordinates* $c = h_j$, $j=0,\cdots,n$, *the scaled eigenfunctions* $(1+|\lambda_{c,n,N}|^2)^s f_{c,n,N}$ *of* $[\mathcal{P}_N^c k^{-2} I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1}L_{c,N}$ *(cf. (7.8)), and the scaled eigenfunctions* $(1+|\lambda_{c,n}|^2)^s f_{c,n}$ *of* k^2L_c *(cf. (3.2)) form Riesz bases in the spaces* H_α^s *for* $-1/2 \leq s \leq 1/2$ *with* N*-independent constants in the corresponding norm estimates (cf. Lemma 8.3). The numerically computed eigenfunctions* $f_{c,n,N}$ *and eigenvalues* $\lambda_{c,n,N}$ *satisfy (8.1).*

iii) The discretized operators $[D_tN_{c,N}]^{\pm 1}\mathcal{P}_N$ (cf. (7.5)), defined for sufficiently large N, converge strongly to $[D_tN_c]^{\pm 1}$ (cf. (5.1)) for $c=h_j,\ j=0,\cdots,n$ (cf. Lemma 8.3).

iv) For $c=h_j,\ j=1,\cdots,n$, all the sums $[k_{c+0}^{-2}D_tN_{c+0}+k_{c-0}^{-2}D_tN_{c-0}]$ are invertible (cf. Lemma 5.3), and the $\{[\mathcal{P}_Nk_{c+0}^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c+0,N}+[\mathcal{P}_Nk_{c+0}^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]D_tN_{c-0,N}\}^{-1}\mathcal{P}_N$, exist for sufficiently large N and converge strongly to $[k_{c+0}^{-2}D_tN_{c+0}+k_{c-0}^{-2}D_tN_{c-0}]^{-1}$ (cf. Lemma 8.6).

v) For $j=1,\cdots,n$, all the S-matrices $S^{h_{j-1}h_j}$ and $S^{h_0h_j}$ are bounded operators (cf. Thm. 4.2).

Choose any pair $(u_{h_0}^+,u_{h_n}^-)$ of incoming waves with $u_{h_0}^+\in H_\alpha^{1/2}(\Gamma_{h_0})$ and $u_{h_n}^-\in H_\alpha^{1/2}(\Gamma_{h_n})$. Then there is a threshold N_0 s.t., for any $N>N_0$, the iterative SMA method (5.13) can be applied on the discrete level without any problem of inverting a noninvertible matrix. The resulting discrete solutions $u_{h_n,N}^+$ and $u_{h_0,N}^-$ tend to the true solutions of the scattering problem, i.e., $\|u_{h_n,N}^+-u_{h_n}^+\|_{H_\alpha^{1/2}(\Gamma_{h_n})}\to 0$ and $\|u_{h_0,N}^--u_{h_0}^-\|_{H_\alpha^{1/2}(\Gamma_{h_0})}\to 0$.

Proof. The plan of proof is as follows. The strong and stable convergence $[D_tN_{c,N}^\pm]^{\pm 1}\to[D_tN_c^\pm]^{\pm 1}$ will imply the strong convergence $P_{c,N}^\pm\to P_c^\pm$. Furthermore, we shall show the strong and stable convergences $T_{ab,N}^+\to T_{ab}^+$, $T_{ba,N}^-\to T_{ba}^-$, and $D_{ab,N}^-\to D_{ab}^-$. Consequently, we shall obtain the strong convergence $S_N^{ab}\to S^{ab}$. For the two-step computation (5.12), define $D_N:=I-S_{-+,N}^{bc}S_{+-,N}^{ab}$. We shall get $D_N\to D$ and the strong and stable convergence $D_N^{-1}\to D^{-1}$. Hence, we shall obtain the strong convergence $S_N^{ac}\to S^{ac}$ for the matrices computed by the two-step algorithm. Applying these arguments in the finitely many steps of Algorithm (5.13), we get the strong convergence of the corresponding operators $\mathbb{S}_N^n\to\mathbb{S}^n$, and the RCWA is shown to be convergent.

So look at the projections $P_{b,N}^\pm$ applied to functions in the images of $P_{a,N}^\pm$ for the case $k(\cdot,b)\not\equiv k(\cdot,a)$. The splitting $u_N=u_{b,N}^++u_{b,N}^-=u_{a,N}^++u_{a,N}^-$ means (cf. (5.2))

$$\left(u_{b,N}^+,D_tN_{b,N}^+u_{b,N}^+\right)+\left(u_{b,N}^-,D_tN_{b,N}^-u_{b,N}^-\right)=\left(u_{a,N}^+,D_tN_{a,N}^+u_{a,N}^+\right)+\left(u_{a,N}^-,D_tN_{a,N}^-u_{b,N}^-\right).$$

Using $D_tN_{c,N}^-=-D_tN_{c,N}^+$, we easily conclude (compare the continuous version (5.6))

$$\begin{aligned}
u_{b,N}^\pm &= \frac{1}{2}\left[u_{a,N}^++u_{a,N}^-\right]\\
&\quad\pm\frac{1}{2}[D_tN_{b,N}^+]^{-1}\left[\mathcal{P}_Nk_b^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\left[\mathcal{P}_Nk_a^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]D_tN_{a,N}^+\left[u_{a,N}^+-u_{a,N}^-\right],\\
P_{b,N}^\pm u_N &= \frac{1}{2}\left[P_{a,N}^+u_N+P_{a,N}^-u_N\right]\\
&\quad\pm\frac{1}{2}[D_tN_{b,N}^+]^{-1}\left[\mathcal{P}_Nk_b^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\left[\mathcal{P}_Nk_a^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]D_tN_{a,N}^+\left[P_{a,N}^+u_N-P_{a,N}^-u_N\right],
\end{aligned}\tag{8.8}$$

where, assuming $\lambda_{c,n,N}\neq 0,\ c=a,b$ for simplicity of presentation, we have (8.4). In view of (8.8), for the convergence of the projections $P_{c,N}^\pm,\ c=a,b$ to $P_c^\pm$, we only need the strong convergences $[D_tN_{c,N}^\pm]\to[D_tN_c^\pm]$ and $[D_tN_{c,N}^\pm]^{-1}\to[D_tN_c^\pm]^{-1}$. This however follows from Lemma 8.3. For the current theorem, the stable convergence is one of the assumptions.

The uniform Riesz property of the bases $f_{c,n,N},\ n=1,\cdots,2N+1$ with $c=a,b$ implies the uniform boundedness of the $T_{ab,N}^+$. For a strong convergence, we need the convergence on a dense subset in $H_\alpha^{1/2}$. We show the convergence on the eigenfunctions. We split the operators by splitting the matrices with respect to the bases of eigenfunctions $f_{c,n,N}$ and $f_{c,n}$. Fixing an appropriate n_0 and setting

$$\begin{aligned}
T_{ab,N,n_0}^+ &:= \left(d_n\delta_{n,m}e^{-\lambda_{a,n,N}[b-a]}\right)_{m,n=1}^{2N+1},\quad d_n:=\begin{cases}1 & \text{if } n\le n_0\\ 0 & \text{else}\end{cases},\\
T_{ab,n_0}^+ &:= \left(d_n\delta_{n,m}e^{-\lambda_{a,n}[b-a]}\right)_{m,n\in\mathbb{Z}_+},
\end{aligned}$$

we get $T^+_{ab,N}=T^+_{ab,N,n_0}+[T^+_{ab,N}-T^+_{ab,N,n_0}]$ and $T^+_{ab}=T^+_{ab,n_0}+[T^+_{ab}-T^+_{ab,n_0}]$. Now, for any prescribed $\varepsilon>0$ there is an n_0 s.t. the estimate $\|[T^+_{ab,N}-T^+_{ab,N,n_0}]\mathcal{P}_N f_{a,m}-[T^+_{ab}-T^+_{ab,n_0}]f_{a,m}\|_{H^{1/2}_\alpha}\le\varepsilon$ holds for sufficiently large n_0 (cf. the arguments in (8.6)). By the same arguments, we even get the norm estimate $\|[T^+_{ab,N}-T^+_{ab,N,n_0}]\mathcal{P}_N-[T^+_{ab}-T^+_{ab,n_0}]\|_{\mathcal{L}(H^{1/2}_\alpha)}\le\varepsilon$. If $f_{a,m}=\sum_{n=1}^{2N+1}\xi_{n,N}f_{a,n,N}$, then

$$T^+_{ab,N,n_0}\mathcal{P}_N f_{a,m}-T^+_{ab,n_0}f_{a,m}=\sum_{n=1}^{n_0}\left[e^{-\lambda_{a,n,N}[b-a]}\xi_{n,N}-e^{-\lambda_{a,m}[b-a]}\xi_{n,N}\right]f_{a,n,N},$$

$$\left\|T^+_{ab,N,n_0}\mathcal{P}_N f_{a,m}-T^+_{ab,n_0}f_{a,m}\right\|^2_{H^{1/2}_\alpha}\sim\sum_{n=1}^{n_0}(1+|\lambda_{a,n,N}|^2)\left|e^{-\lambda_{a,n,N}[b-a]}-e^{-\lambda_{a,m}[b-a]}\right|^2|\xi_{n,N}|^2.$$

Similarly, from the convergence of the Dirichlet-to-Neumann mappings, we conclude

$$\left\|D_tN^+_{a,N,n_0}\mathcal{P}_N f_{a,m}-D_tN^+_{a,n_0}f_{a,m}\right\|^2_{H^{1/2}_\alpha}\sim\sum_{n=1}^{n_0}(1+|\lambda_{a,n,N}|^2)^{-1}|\lambda_{a,n,N}-\lambda_{a,m}|^2|\xi_{n,N}|^2\to0.$$

For fixed n_0, the last two formulas imply $T^+_{ab,N,n_0}\mathcal{P}_N f_{a,m}-T^+_{ab,n_0}f_{a,m}\to0$, and the strong convergence $T^+_{ab,N}\mathcal{P}_N\to T^+_{ab}$ is proved. Moreover, using the above splitting, we even get that the convergence $T^+_{ab,N}\mathcal{P}_N\to T^+_{ab}$ is compact. Indeed, we take $x_N\in H^{1/2}_\alpha$, $N\in\mathbb{Z}_+$ uniformly bounded, take any subsequence x_N, $N\in\mathbb{N}'\subset\mathbb{Z}_+$, and take any $\varepsilon>0$. Then, for a suitable fixed n_0, we obtain the estimates $\|T^+_{ab,N}-T^+_{ab,N,n_0}\|\le\varepsilon$ and $\|T^+_{ab}-T^+_{ab,n_0}\|\le\varepsilon$. Expanding the truncated Fourier series into the eigenfunction basis as $\mathcal{P}_N x_N=\sum_{n=1}^{2N+1}\xi_{N,n}f_{c,n,N}$, we can choose an infinite subset $\mathbb{N}''\subset\mathbb{N}'$ s.t. the $\xi_{N,n}$ is close to a limit $\xi_n\in\mathbb{C}$, i.e., $|\xi_{N,n}-\xi_n|<\varepsilon$ for $N\in\mathbb{N}''$. Using $f_{c,n,N}\to f_{c,n}$ and $\lambda_{c,n,N}\to\lambda_{c,n}$ and setting $x:=\sum_{n=1}^{n_0}\xi_n f_{c,n}$, we arrive at $\|T^+_{ab,N,n_0}\mathcal{P}_N x_N-T^+_{ab,n_0}x\|_{H^{1/2}_\alpha}\le C\varepsilon$ for sufficiently large N. In other words, for $N\in\mathbb{N}''$ sufficiently large $\|T^+_{ab,N}\mathcal{P}_N x_N-T^+_{ab}x\|_{H^{1/2}_\alpha}\le C\varepsilon$, and the convergence $T^+_{ab,N}\mathcal{P}^a_N x_N\to T^+_{ab}$ is really compact. Similarly, it can be shown that the strong convergence $T^-_{ba,N}\mathcal{P}^a_N x_N\to T^-_{ba}$ is compact.

Next we have to show that the strong convergence $D^-_{ab,N}\mathcal{P}_N\to D^-_{ab}$ is stable, i.e., we have to prove $[D^-_{ab,N}]^{-1}\mathcal{P}_N\to[D^-_{ab}]^{-1}$. Choosing the sign $\pm$ as $-$ in (8.8) and setting $u^+_{a,N}=0$, we get

$$\begin{aligned}D^-_{ab,N}&=\frac{1}{2}\left\{I_N+[D_tN^+_{b,N}]^{-1}\left[\mathcal{P}_N k_b^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\left[\mathcal{P}_N k_a^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]D_tN^+_{a,N}\right\}\\&=\frac{1}{2}[D_tN^+_{b,N}]^{-1}\left[\mathcal{P}_N k_b^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\left\{\left[\mathcal{P}_N k_b^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]D_tN^+_{b,N}+\left[\mathcal{P}_N k_a^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]D_tN^+_{a,N}\right\}.\end{aligned}\tag{8.9}$$

The convergence of the approximate operators $\left[\mathcal{P}_N k_c^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}\right]$ to the strongly elliptic multiplication operator $k_c^{-2}I$ is stable in H^s_α for $-1/2\le s\le1/2$. By the Assumptions iii) and iv) of the current theorem we conclude that $D^-_{ab,N}\mathcal{P}_N$ converges strongly to D^-_{ab} and that this convergence is stable. Unfortunately, with Assumption iv) we rely on the poor result of Lemma 8.6. Nevertheless, putting the strong and stable convergences together, we get the strong convergence $S^{ab}_N\mathcal{P}_{N,N}\to S^{ab}$.

For the two-step computation in (5.12), we define $D_N:=I-S^{bc}_{-+,N}S^{ab}_{+-,N}$. Clearly, we get the strong convergence $D_N\mathcal{P}_N\to D$. However, we need a stable convergence since the operator D_N is inverted in the recursion step. We need $D_N^{-1}\mathcal{P}_N\to D^{-1}$. Fortunately, the block S^{bc}_{-+} is compact since T^+_{ab} and T^-_{ba} compact, which follows by the Riesz property, by the representation as a diagonal matrix $(e^{-\lambda_{a,n}[b-a]}\delta_{m,n})_{m,n\in\mathbb{Z}_+}$ (recall Sect. 6), and by the decay of the diagonal entries. Based on this fact, we have shown the compact convergence $T^-_{cb,N}\mathcal{P}^b_N x_N\to T^-_{cb}$. Since S^{bc}_{-+} is equal to T^-_{cb} multiplied by a bounded operator, we get the compact convergence $S^{bc}_{-+,N}S^{ab}_{+-,N}\mathcal{P}_N\to S^{bc}_{-+}S^{ab}_{+-}$. Thm. 8.2

implies the stable convergence $D_N \mathcal{P}_N \to D$. We finally obtain that $S_N^{ac}\mathcal{P}_{N,N}$ converge strongly to S^{ac} for $N \to \infty$.

Altogether, we have shown the strong convergence $S_N^{ac}\mathcal{P}_{N,N} \to S^{ac}$ for the matrices computed by the two-step algorithm if we know the strong convergences $S_N^{ab}\mathcal{P}_{N,N} \to S^{ab}$ and $S_N^{bc}\mathcal{P}_{N,N} \to S^{bc}$ and if the convergence $T^-_{cb,N}\mathcal{P}_N^b x_N \to T^-_{cb}$ is compact. Applying these arguments to finitely many steps of Algorithm (5.13), we get the strong convergence of the corresponding operators $\mathbb{S}_N^n \mathcal{P}_{N,N} \to \mathbb{S}^n$. In other words, the RCWA is shown to be convergent. □

Remark 8.9. *Assumption v) of the theorem is natural. If the problem over a complex domain is reduced to the solution of problems in subdomains, then the solution of the subdomain should exist and should be unique. Trapped modes in subdomains must be excluded. Assumption i) is natural as well. There is no reason to subdivide any domain with wavenumber function independent of x_2 since this means more work with no improvement. The Riesz basis property in Assumption ii) is of technical nature. The second part of ii), the condition of (8.1), is designed to avoid that an inaccurate EVD computation spoils the convergence.*

Remark 8.10. *The assumptions iii) and iv) are technical, and in many cases (cf. the additional assumptions in the cited Lemmata) it is unclear how to check these practically. So in many cases it cannot be excluded that these assumptions are violated. It may happen, that the algorithm breaks down due to a required inversion of an ill-behaved matrix block. Or the RCWA might not be convergent for $N \to \infty$. However, in the special case of TM polarization with Hölder continuous and piecewise twice continuously differentiable wavenumber functions $k_{h_j} \geq \varepsilon > 0$, the technical assumptions iii) and iv) are fulfilled. Similarly, in the special case of TE polarization with piecewise continuous wavenumber function $k_{h_j} \geq \varepsilon > 0$, the technical assumptions iii) and iv) are fulfilled.*

Remark 8.11. *If in the TM case the functions k_{h_j} are only piecewise smooth, then the splitting into upgoing and downgoing wave is not continuous, and a stable convergence of the RCWA seems not to be possible. Usually, this means that the convergence is not possible for all natural input values but for special inputs like plane-wave incidence it might be converging. However, there will be a loss in the convergence rate in comparison to the rate of approximation by the trial space.*

9 Concluding remarks: Open problems, Area of application

Mathematically, there remain many interesting **open problems**.

- For the discretization with fixed slicing, i.e., for fixed $h_0, h_1, \cdots, h_n$:
  a) Is there any situation s.t. an eigenfunction of rank greater one occurs? If yes, then a modification of the code is required. This is difficult to check since the rank does not depend continuously on the geometry and optical indices.
  b) Is there any situation s.t. the Riesz basis property is not satisfied or is not uniform? If yes, then other discretized norms than that of a weighted ℓ^2 norm may appear. An extension of the theoretical background might be necessary.
  c) How does the EVD looks like for more general k and the TM polarization. Is there still some kind of error analysis for the complete system of eigenvalues and eigenfunctions?
  d) What about the rates of convergence? In simple cases, this might not be too difficult.
  e) What about function $x_1 \mapsto k(x_1, x_2)$ with jump discontinuities? This leads to different trace spaces $k^{-2}(\cdot, c\pm 0)H_\alpha^{-1/2}$. Should the coupling over the Γ_{h_j} and, therewith, the SMA be modified?

- For the discretized SMA with wavenumber function depending on x_2 and with $n\to\infty$ s.t. the width of slices $\max\{h_j-h_{j-1}\colon\ j=1,\cdots,n\}$ tends to zero:
  a) For a single slice, can the formula for the S-matrix $S^{h_{j-1}h_j}$ with x_2-dependent k and with h_j-h_{j-1} be approximated by an S-matrix with frozen x_2-independent k s.t. the results proven for the independent case take over to the dependent case?
  b) How to analyze the recursion algorithm for $n\to\infty$? This reminds of the stability problems for FDM of ODE-systems. Inside the slice the method of the FMM might be explicit. Over the boundary of two slices it is implicit.

The **area of applications** for the general SMA are scattering problems over deep surface structures, i.e., gratings with period $\mathrm{per}\sim\lambda_{\mathrm{inc}}$ and with $h_n-h_0\geq\mathcal{O}(\mathrm{per})$. In particular, suppose the grating is the union of many slices but all these layers are shifted versions of two or three standard layers. In this case, the scattering matrices of the two or three standard slices are computed once and can be reused many times. Gratings similar to photonic crystals (cf. e.g. [2]) are of this structure. The SMA discretized as RCWA is efficient for gratings with a few slices of big size $h_j-h_{j-1}\geq\mathcal{O}(\mathrm{per})$, each with x_2-independent optical index. No additional slicing is needed for these. Furthermore, if the wave solution u is smooth w.r.t. the horizontal x_1-coordinate, then a small truncation index N is sufficient, and the discretized iteration (5.13) is fast. For an application of the FMM, no independence of the vertical x_2-coordinate is needed. However, the wave solution should have a certain degree of smoothness w.r.t. this vertical coordinate s.t. the FDM performs well. For non-smoothness w.r.t. x_2, an adaptive choice of the slicing stepsize and FDM stepsize would be helpful.

We conclude with a remark on the **comparison** of RCWA/FMM and FEM. Clearly, engineers and physicist prefer the RCWA/FMM, for it is based on eigenmode expansion, i.e., on physical intuition. Though a comparison of a code for RCWA/FMM and one for the FEM is possible, a comparison of the general methods is difficult. In general, the FEM is a basic discretization scheme and, therefore, should be compared to a Galerkin approximation based on truncated Fourier expansions. Here it is clear that, due to elaborated standard techniques, FEM is more suitable to approximate singularities. Surely, this requires adaptive FEM grids and error estimators. Corresponding adaptions on the side of the RCWA might be possible, but require to develop new codes. On the other hand, for special situations (cf. [2]), a smooth solution can be approximated very effectively by truncated Fourier series.

Looking at its nature, the RCWA should rather be compared to FEM combined with domain decomposition. In this sense, Assumption v) in Thm. 8.8 is common for both methods. If this is fulfilled, then FEM is guaranteed to converge for our elliptic PDE. For the RCWA/FMM, there still might occur problems with rank-two eigenfunctions, with ill conditioned systems of eigenfunction, and with the inversion of ill-behaved operators. Note that these are open problems, and is not clear whether such problems really occur. Besides, at least for real-valued k, numerical experiments and the successful applications over many years prove the RCWA/FMM to be reliable numerical schemes.

References

[1] N. Anttu and H.Q. Xu, Scattering matrix method for optical excitation of surface plasmons in metal films with periodic arrays of subwavelength holes, *Physical review B*, Vol. **83**, 165431, pp. 165431-1–165431-17, 2011.

[2] C. Brée et.al., Chirped photonic crystal for spatially filtered optical feedback to a broad-area laser, *Journal of Optics (IOP Publishing Journal)* **20**, issue 9 (2018), 095805.

[3] J. Bischoff, Improved diffraction computation with a hybrid C-RCWA method. *Advanced lithography* 134, 2009

[4] B.J. Civiletti, A. Lakhtakia, and P.B. Monk, Analysis of the Rigorous Coupled Wave Approach for p-polarized light in gratings, *J. Comp. Appl. Math.* **386** (2021), 113235.

[5] 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.

[6] 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.

[7] G. Hu, A. Rathsfeld, *Radiation conditions for the Helmholtz equation in a half plane filled by inhomogeneous periodic material*, WIAS Preprint **2726**, revised version, Berlin 2023.

[8] B.H. Kleemann, *Elektromagnetische Analyse von Oberflächengittern von IR bis XUV mittels einer parametrisierten Randintegralmethode: Theorie, Vergleich und Anwendungen*. Dissertation, TU Ilmenau, 2002, Mensch und Buch Verlag Berlin, 2003.

[9] P. Lalanne and G.M. Morris, Highly improved convergence of the coupled-wave method for TM-polarization, *JOSA A*, **13** (1996), pp. 779–784.

[10] L. Li, Use of Fourier series in the analysis of discontinuous periodic structures, *J. Opt. Soc. Am.*, **13**, No. 9, (1996), pp. 1870–1876.

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

[12] M.G. Moharam, E.B. Grann, D.A. Pommet, and T.K. Gaylord, Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach, *JOSA A*, **12** (1995), pp. 1077–1086.

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

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

[15] E. Popov (ed.), *Gratings: Theory and Numeric Applications*, Second revisited Edition, Aix Marseille Universite, CNRS, Centrale Marseille, Institut Fresnel UMR, 7249 (2014).

[16] A. Tavrov, M. Totzeck, N. Kerwien, H.J. Tiziani, Rigorous coupled-wave analysis calculus of submicrometer interference pattern and resolving edge position versus signal-to-noise ratio, *Opt. Eng.*, **41**(8) (2002), pp. 1886–1892.

[17] G. Vainikko, Funktionalanalysis der Diskkretisierungsmethoden, Teubner, Leipzig, 1976.