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 (revision: June 25, 2025)

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 a domain decomposition method is applied, coupling the solutions in neighbour slices over the common interface via equating scattering data. A clever recursion is set up to compute an approximate operator, mapping incoming waves into outgoing ones. Combining this Scattering Matrix Algorithm with numerical schemes inside the slices, methods like Rigorous Coupled Wave Analysis and Fourier Modal Method were designed. The key for the analysis is the scattering problem over the slices. These are scattering problems with a radiation condition generalized to inhomogeneous cover and substrate materials and were first analyzed in [7]. Now suppose there exists a slicing s.t. the optical index function in each slice is independent of the direction perpendicular to the interfaces of the slicing. Then the slicing of the Scattering Matrix Algorithm should be fixed to this slicing, i.e. a refinement of the slicing is unnecessary. For such a fixed slicing and for Transverse Electric polarization combined with exact solvers over the subdomains (no full discretization), it was proved in [7] that the Scattering Matrix Algorithm leads to the exact solution of the scattering problem. In this paper 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].
- Speaking 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 the solution over each subdomain (slice), various realizations of the SMA are possible. Namely:

DOI 10.20347/WIAS.PREPRINT.3081 Berlin, December 22, 2023/rev. June 25, 2025

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. (5.5)).

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 numerical scheme like Finite Difference Method (FDM), Runge-Kutta Method or Linear Multistep Method.

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 list 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 Subsect. 8.1. 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 and its version (3.1).

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 in the other. For the classical diffraction, the propagation direction of the plane wave incident to the surface is orthogonal to the surface 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 reduce to the 2D Helmholtz equation, i.e., to $\Delta u + k^2 u = 0$ for TE and to $\nabla \cdot k^{-2} \nabla u + u = 0$ 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 direction 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 truncated 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. [7] and 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. 8.11 and the remarks following it), showing that the operators defined by the discretized EVDs converge strongly to the operators of the continuous level. For the inverse operators involved in the formulas, we show 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 piecewise twice continuously differentiable for the TM case and piecewise continuous for TE polarization. We believe the convergence probably holds in many more cases. So there remain many open problems around the assumptions imposed in Sect. 8 (cf. the end of Sect. 9).

Now consider the case of surface structures, which are not the union of slices with wavefunctions independent of the vertical direction. For this case of wavefunctions depending on the vertical and horizontal coordinates, 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 numerical schemes for ODEs, 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 classical BVP for the scattering by gratings under TM polarization. However, the solution of the scattering problem over the subdomains (slices) requires the notion of a general BVP, where the homogeneous materials of cover and substrate are replaced by special inhomogeneous materials. To prepare the corresponding generalized radiation conditions, we shall discuss an EVD of a one-dimensional ODE derived from the elliptic PDE in Sect. 3. In Sect. 4 we shall define the generalized radiation condition for the 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. Note that there exist several versions of the SMA, and we present the one which seems to be the best. Namely, we compute the accumulated S-matrix, using the actual S-matrix of the slice and not the T-matrix. Also, for slices defined by $h_{j-1} < x_2 \le h_j$, we do not use the additional splitting into the slice $h_{j-1} < x_2 < h_j$ and the infinitesimal slice $h_j - 0 < x_2 \le h_j + 0$, which simplifies some of the formulas but requires an additional accumulation step. 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. 8.11 on the convergence of

the RCWA for fixed slicing and 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

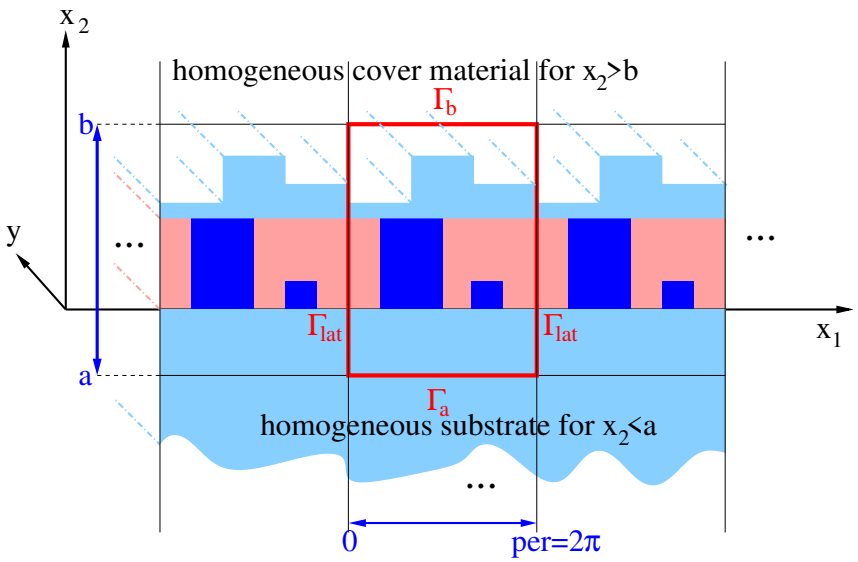

Figure 1: Geometry settings for homogeneous cover material and substrate.

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 $\{(x_1,y,x_2)^\top\in\mathbb{R}^3\colon\ a\le x_2\le b\}$ (cf. Fig. 1), which is 2π-periodic in x_1 direction and invariant w.r.t. shifts in y-direction. The incident plane wave is defined as $u_b^{\rm inc}(x_1,y,x_2):=u_b^{\rm inc}(x_1,x_2):=e^{{\rm i}\vec\alpha\cdot(x_1,x_2)}$ with $\vec\alpha=(\alpha,-\sqrt{[k^+]^2-\alpha^2}\,)^\top$, $0\le\alpha<k^+$. 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 $\omega:=2\pi/\lambda_{\rm inc}>0$ is the frequency of the incoming light of wavelength $\lambda_{\rm inc}$, where ε_0 and μ_0, respectively, are the electric permittivity and the magnetic permeability in vacuum, and where $\mathbf{n}_+$ is the refractive (optical) index of the material. The angle of incidence $\theta^{\rm inc}$ is connected to $\vec\alpha$ by $\vec\alpha:=k^+(\sin\theta^{\rm inc},-\cos\theta^{\rm inc})^\top$ and $\alpha:=k^+\sin\theta^{\rm inc}$. Similarly, there is a constant wavenumber $k^-=\omega\sqrt{\mu_0\varepsilon_0}\,\mathbf{n}_-$ for the half space with $x_2<a$.

The function $u_b^{\rm inc}$ is α-quasiperiodic, i.e. the function $e^{-{\rm i}\alpha x_1}u_b^{\rm inc}(x_1,x_2)$ 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 $H^1_\alpha(\Omega)$ and 1D Sobolev spaces $H^{1/2}_\alpha(\Gamma_c)=H^{1/2}_\alpha(0,2\pi)$, respectively, i.e. in spaces of α-quasiperiodic H^s-functions with $s=1$ and $s=1/2$, respectively. We can even admit a general incident field $u_b^{\rm inc}(x_1,x_2)$ for $x_2\ge b$ (cf. the subsequent (2.3)) if only the restriction $u_b^{\rm inc}|_{\Gamma_b}$ 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 $u_a^{\rm inc}(x_1,x_2)$ 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 scaled y-component of the magnetic field vector $u(x_1,y,x_2):=\sqrt{\mu_0/\varepsilon_0}H_y(x_1,y,x_2)$, which is independent of y. So the harmonic 3D Maxwell equations turns into the 2D equation for the function $u(x_1,y,x_2)=u(x_1,x_2)$. Altogether, the wave u is the

solution over the domain $\Omega := [0, 2\pi] \times [a, b]$ satisfying

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

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

$$\begin{aligned} u(x_1, x_2) - u_b^{\mathrm{inc}}(x_1, x_2) &= \sum_{l \in \mathbb{Z}} e^{\mathbf{i}(\alpha + l)x_1} \left\{ c_{b,l}^+ e^{\mathbf{i}\beta_l^b (x_2 - b)} + c_{b,l}^- e^{-\mathbf{i}\beta_l^b (x_2 - b)} \right\}, \quad x_2 \geq b\,, \\ &\beta_l^b := \sqrt{[k^+]^2 - [\alpha + l]^2}, \quad c_{b,l}^\pm \in \mathbb{C}. \end{aligned} \tag{2.2}$$

Here the argument $z := [k^+]^2 - [\alpha + l]^2$ of the square root is in the half plane $\{z \in \mathbb{C} \colon \Im m\, z \geq 0\}$. The square root is defined such that $\Re e\, \sqrt{z} \geq 0$ and $\Im m\, \sqrt{z} \geq 0$. The radiation condition on Γ_b requires coefficients $c_{b,l}^- = 0$ for all coefficients of downgoing modes $e^{\mathbf{i}(\alpha + l)x_1} e^{-\mathbf{i}\beta_l^b (x_2 - b)}$, 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 $\beta_l^b = 0$ (i.e. if $l = -\alpha \pm k^+$ is an integer), 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 the formula $\{c_{b,l}^+ \big(1 + (x_2 - b)\big) + c_{b,l}^- \big(1 - (x_2 - b)\big)\}$, 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 $\beta_l^a = 0$.

The portion of energy from an incoming plane wave radiated into the direction of the lth propagating outgoing wave mode is called efficiency. If the incoming field is defined as $u_a^{\mathrm{inc}}(x_1, x_2) := 0$ and $u_b^{\mathrm{inc}}(x_1, x_2) := \sqrt{\mathbf{n}_+} e^{\mathbf{i}\alpha x_1} e^{-\mathbf{i}\beta_0^b (x_2 - b)}$, then the efficiency e_l^b of the lth reflected plane-wave mode and the efficiency e_l^a of the lth transmitted plane-wave mode are given by

$$e_l^b = \frac{\beta_l^+}{\beta_0^+} \frac{|c_{b,l}^+|^2}{\mathbf{n}_+}, \quad \text{and} \quad e_l^a = \frac{\beta_l^-}{\beta_0^+} \frac{[k^+]^2}{[k^-]^2} \frac{|c_{a,l}^-|^2}{\mathbf{n}_+}, \tag{2.5}$$

respectively.

3 Eigenfunctions of the ODE for the reformulation of the "Helmholtz" equation

3.1 The ODE equivalent to the wave equation

For the scattering matrix algorithm, we have to generalize the radiation conditions (2.3) and (2.4) to model inhomogeneous cover and substrate materials as well. 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 function $\mathbf{n}$, 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

$$\nabla\cdot k(x_1,x_2)^{-2}\nabla u(x_1,x_2)+u(x_1,x_2) \;=\; 0,\quad (x_1,x_2)^\top\in\mathbb{R}^2. \tag{3.1}$$

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 $k(x_1,x_2)=k_+(x_1)$ for $x_2>b$ and $k(x_1,x_2)=k_-(x_1)$ for $x_2<a$. Equ. (3.1), restricted to cover and substrate, is equivalent to the operator valued ODE w.r.t. the variable x_2 of the form

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

The solution of this equation can be represented by expanding the function w.r.t. x_1 in a series of α-quasiperiodic eigenfunctions of the operator k^2L.

We reduce this second-order ODE (3.2) to a first-order ODE. Setting $v:=\partial_{x_2}u$ and $\vec{u}:=(u,v)^\top$, 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 the space of univariate and α-quasiperiodic 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, $k^2Lf_\lambda=\lambda^2 f_\lambda$. For the eigenvalues $\pm\lambda$ of M, we obtain the eigenvector $(f_\lambda,\pm\lambda f_\lambda)^\top$ with f_λ satisfying

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

3.2 EVD for twice continuously differentiable wavenumber function

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 $f=kh$ with an α-quasiperiodic h.

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

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

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

whereas, for the Helmholtz equation in the TE case,

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

The underlying Sobolev space of the variational forms is $H^1_\alpha(\Omega)$ in both cases. Consequently, the Dirichlet data $u|_{\Gamma_a}$, $c=a,b$ is in the trace space $H^{1/2}_\alpha(\Gamma_c)$ for TE and TM. For TE, the dual spaces $H^{-1/2}_\alpha(\Gamma_b)$ and $H^{-1/2}_\alpha(\Gamma_b)$ are the spaces of the traces of the normal derivatives $\partial_2 u|_{\Gamma_b}$ and $-\partial_2 u|_{\Gamma_a}$, respectively. For TM, however, the spaces $H^{-1/2}_\alpha(\Gamma_b)$ and $H^{-1/2}_\alpha(\Gamma_b)$ are the spaces of the traces of the co-normal derivatives $k^{-2}\partial_2 u|_{\Gamma_b}$ and $-k^{-2}\partial_2 u|_{\Gamma_a}$, respectively. So we get similar formulas for the PDE $\nabla\cdot k^{-2}\nabla u+u=0$ and h as for $\Delta u+k^2u=0$ and f. In any case, we can use the results collected in [7, Lemma 4.5].

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

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

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 $-2\le s\le 2$. Due to the boundedness of the operators of multiplication by k and k^{-1}, we also have a Riesz basis $(1+n^2)^{-s/2}f_n=(1+n^2)^{-s/2}kh_n\in H^s_\alpha,\ n\in\mathbb{Z}$ for $-2\le s\le 2$. Additionally, we have a Riesz basis $(1+n^2)^{-s/2}k^{-2}f_n\in H^s_\alpha,\ n\in\mathbb{Z}$ for $-2\le s\le 2$. We shall suppose throughout this paper that all eigenfunctions are of rank one. If rank-greater-than-one eigenfunctions occur, then the subsequent SMA must be adapted to that case. 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 all the eigenfunctions of selfadjoint operators have rank one.

Lemma 3.2. *For $|n|\to\infty$, we have the asymptotics*

$$h_n(t) = \begin{cases} e^{\mathbf{i}\lambda_n t}-\frac{1}{\lambda_n}\frac{e^{\mathbf{i}\lambda_n t}}{2\mathbf{i}}\int_0^t\tilde k^2(\tau)\mathrm{d}\tau+\mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right) & \text{if } \alpha\neq 0,\frac12 \text{ and } h_n(0)\neq 0\\ e^{\mathbf{i}\lambda_n t}+\mathcal{O}\left(\frac{1}{|\lambda_n|^{-1/2}}\right) & \text{if } \alpha=0,\frac12 \text{ and } h_n(0)\neq 0\\ \sin(\lambda_n t)-\frac{\cos(\lambda_n t)}{2\lambda_n}\int_0^t\tilde k^2(\tau)\mathrm{d}\tau+\mathcal{O}\left(\frac{1}{|\lambda_n|^2}\right) & \text{if } h_n(0)=0\end{cases}.$$

3.3 EVD for positive and piecewise twice continuously differentiable wavenumber function

In a second case, suppose that the wave number $k(x_1)$ is only piecewise twice continuously differentiable and that $k(x_1)\ge c_k>0$. Still we get

Lemma 3.3. *The spectrum σ_{k^2L} of operator k^2L is a discrete set of eigenvalues $\{\lambda_n^2 : n\in\mathbb{N}\}\subset\mathbb{R}$ with $\lambda_n^2\to\infty$, $|n|\to\infty$. Each eigenfunction f_n corresponding to λ_n^2 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 $(1+|\lambda_n|^2)^{-s/2}f_n$, $n\in\mathbb{N}$ of the differential operator k^2L form a Riesz basis in H^s_α for $0\le s\le 1$. The functions $(1+|\lambda_n|^2)^{-s/2}k^{-2}f_n$, $n\in\mathbb{N}$ form a Riesz basis in H^s_α for $-1\le s\le 0$.*

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 $\operatorname{dom}_{\tilde L} := k^{-1}H^1_\alpha$. This is the operator of the variational form $(h,g)\mapsto a(h,g)$ with

$$a(h,g) := \int_0^{2\pi}\Big\{k^{-2}(x_1)\,\partial_{x_1}[k(x_1)h(x_1)]\,\overline{\partial_{x_1}[k(x_1)g(x_1)]} - k^2(x_1)h(x_1)\overline{g(x_1)}\Big\}\,\mathrm{d}x_1. \quad (3.6)$$

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, $n\in\mathbb{N}$ 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 $h=\sum_{n\in\mathbb{N}}\xi_n h_n$ with $\xi_n\in\mathbb{C}$, we obtain

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

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 $\langle f_{n_1},k^{-2}f_{n_2}\rangle=\delta_{n_1,n_2}$ and by duality arguments, we get the Riesz property of $(1+|\lambda_n|^2)^{-s/2}\,k^{-2}f_n$ in H^s_α for $-1\le s\le 0$.

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 $h_n''+\tilde k^2h_n+\lambda_n^2h_n=0$ holds (cf. (3.5)). Consequently, the solution is twice continuously differentiable over the closed interval. On the other hand, the equation $k^2Lf_n=\lambda_nf_n$ 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 $\partial_{x_1}[k^{-2}f_n']$. □

3.4 General assumptions for EVD

We just formulate a general assumption, which holds e.g. for the second case considered above and, if (3.7) is true, also for the first case. In other words, we assume that, for k^2L, there is a system of univariate eigenfunctions f_n, $n\in\mathbb{N}$ with the corresponding eigenvalues λ_n^2 such that:

- $k^2Lf_n=\lambda_n^2f_n$ and all λ_n^2 are eigenvalues of rank one. (3.7)
- $|\lambda_n|^2\to\infty$ and $|\Im m\,\lambda_n|\le C_0+C_0|\Re e\,\lambda_n|$ for a fixed positive constant C_0. (3.8)
- $(1+|\lambda_n|^2)^{-s/2}f_n$ form a Riesz basis in H^s_α for $0\le s\le 1$ and $(1+|\lambda_n|^2)^{-s/2}k^{-2}f_n$ form a Riesz basis in H^s_α for $-1\le s\le 0$. (3.9)

Note that, in comparison to the above indices, we have changed the index set $\mathbb{Z}$ to $\mathbb{N}$. Furthermore, we order the eigenvalues and eigenfunctions such that $|\lambda_n| \leq |\lambda_{n+1}|$ for all $n \in \mathbb{N}$. For the square roots λ_n of the $\lambda_n^2 \neq 0$, we suppose either that $\Re e\, \lambda_n > 0$ or that $\Re e\, \lambda_n = 0, \Im m\, \lambda_n < 0$. Clearly, the $\pm\lambda_n$ are eigenvalues of M in (3.3). Depending on the choice of wavenumber function k as $k(x_1) = k^+(x_1) = k_b(x_1) = k(x_1, b+0)$ or as $k(x_1) = k^-(x_1) = k_a(x_1) = k(x_1, a-0)$, we write L_b or L_a for the differential operator L, $f_{b,n}$ or $f_{a,n}$ for the eigenfunction f_n, and $\lambda_{b,n}$ or $\lambda_{a,n}$ for the eigenvalue λ_n. Moreover, if the wavenumber function $k(x_1) = k(x_1, b \pm 0)$ depends on whether the trace is taken from above or below, we even write $L_{b\pm0}$, $f_{b\pm0,n}$, and $\lambda_{b\pm0,n}$.

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

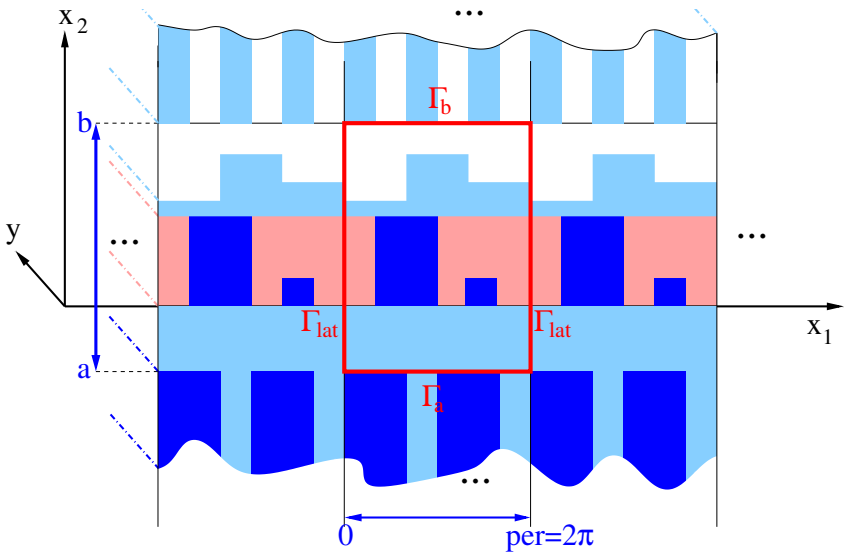

Figure 2: Geometry settings for inhomogeneous cover material and substrate.

Suppose $c = a$ or $c = b$. More precisely, $c = a-0$ or $c = b+0$. The eigenvalues λ of M in (3.3) are the square roots $\pm\sqrt{\lambda^2}$ of the eigenvalues λ^2 in (3.4). For definiteness, we choose the square root $\lambda = \sqrt{\lambda^2}$ s.t. either $\Re e\, \lambda > 0$ or that $\Re e\, \lambda = 0$, $\Im m\, \lambda \leq 0$, and consider the values $\pm\lambda$ as square root of λ^2. 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 $-$. Clearly, the $u^+_{c,n}$ are solutions of (3.1) for $x_2 > c$ and the $u^-_{c,n}$ solutions for $x_2 < c$. In the special case $\lambda_{c,n} = 0$, we define these waves by the modified formula $u^{\pm}_{c,n}(x_1, x_2) := (1 \mp (x_2 - c)) f_{c,n}(x_1)$. The general representation of the "Helmholtz" solutions in the inhomogeneous cover material close to Γ_c is

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

The expansion (2.2) for k is a special case of (4.1) for $c = b$, 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 still to be changed into $\mathbb{N}$. 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 $\{(x_1,x_2)^\top\colon x_2\geq b\}$ is said to satisfy the upper radiation condition if u admits the expansion $u(x_1,x_2)=u_b^{\rm inc}(x_1,x_2)+\sum_{n\in\mathbb{N}}c_{b,n}^+u_{b+0,n}^+(x_1,x_2)$ for a sequence of coefficients $c_{b,n}^+\in\mathbb{C}$. Similarly, an α-quasiperiodic solution u of $\nabla\cdot k^{-2}\nabla u-u=0$ over the lower half-space $\{(x_1,x_2)^\top\colon x_2\leq a\}$ is said to satisfy the lower radiation condition if it admits the expansion $u(x_1,x_2)=u_a^{\rm inc}(x_1,x_2)+\sum_{n\in\mathbb{N}}c_{a,n}^-u_{a-0,n}^-(x_1,x_2)$ for a sequence of coefficients $c_{a,n}^-\in\mathbb{C}$. The sums in the expansions converge in $H^1_{\rm loc}$.*

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 $u(x_1,x_2)=\sum_n c_ne^{-\lambda_nx_2}f_n(x_1)$ with the discrete $H^{1/2}_\alpha$ norm $\|(c_n)_n\|_{\ell^{1/2}}:=\left\{\sum_n(1+|\lambda_n|^2)^{1/2}|c_n|^2\right\}^{1/2}<\infty$, and note that $\|(c_n)_n\|_{\ell^{1/2}}\leq C\|u\|_{H^{1/2}_\alpha}$ by the Riesz property (3.9) for the scaled functions $(1+|\lambda_n|^2)^{-1/4}f_n$ in $H^{1/2}_\alpha$. Then we get

$$\begin{aligned}
\|\partial_{x_2}u\|^2_{L^2}&=\int_0^1\int_0^{2\pi}\left|\sum_n\lambda_nc_nf_n(x_1)e^{-\lambda_nx_2}\right|^2{\rm d}x_1{\rm d}x_2\leq C\int_0^1\sum_n\left|\lambda_nc_ne^{-\lambda_nx_2}\right|^2{\rm d}x_2\\
&\leq\sum_n|\lambda_nc_n|^2\int_0^1e^{-2\Re e\,\lambda_nx_2}{\rm d}x_2\leq C\sum_n|\lambda_n|\,|c_n|^2\leq C\|(c_n)_n\|^2_{\ell^{1/2}},\\
\|\partial_{x_1}u\|^2_{L^2}&=\int_0^1\int_0^{2\pi}\left|\partial_{x_1}\sum_nc_nf_n(x_1)e^{-\lambda_nx_2}\right|^2{\rm d}x_1{\rm d}x_2\leq C\int_0^1\sum_n(1+|\lambda_n|^2)\left|c_ne^{-\lambda_nx_2}\right|^2{\rm d}x_2\\
&\leq\sum_n(1+|\lambda_n|^2)\,|c_n|^2\int_0^1e^{-2\Re e\,\lambda_nx_2}{\rm d}x_2\leq C\sum_n(1+|\lambda_n|^2)^{1/2}\,|c_n|^2\leq C\|(c_n)_n\|^2_{\ell^{1/2}},
\end{aligned}\tag{4.2}$$

where we have used $|\Im m\,\lambda_n|\leq C_0+C\Re e\lambda_n$ 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 $u_a^{\rm inc}$ and $u_b^{\rm inc}$ by the grating with homogeneous cover material and substrate to the BVP for a grating with inhomogeneous super- and substrate. We obtain:

Theorem 4.2. *Suppose:*

i) For $c=a-0,b+0$ and the corresponding $k^2L=k^2L_c$, the assumptions (3.7)-(3.9) are fulfilled.

ii) Any solution of the scattering problem (2.1) with incident waves $u_a^{\rm inc}\equiv0$ and $u_b^{\rm inc}\equiv0$ is zero.

Then, for any given $u_b^{\rm inc}|_{\Gamma_b}\in H^{1/2}_\alpha(\Gamma_b)$ and $u_a^{\rm inc}|_{\Gamma_a}\in H^{1/2}_\alpha(\Gamma_a)$, there is a unique solution $u\in H^1_\alpha(\Omega)$ 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, cf. Fig. 3)

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

Proof. The argumentation is almost the same as that in the proof of [7, Theorem 5.7]. We define the Dirichlet-to-Neumann operators $D_tN_a^-$ and $D_tN_b^+$ for the basis function of the radiation conditions by $u_n^a\mapsto D_tN_a^-u_n^a:=-\partial_{x_2}u_n^a$ and by $u_n^b\mapsto D_tN_b^-u_n^b:=\partial_{x_2}u_n^b$, respectively. Then the sesquilinear form for (2.1) is

$$a(u,v):=\int_\Omega\left\{-k^{-2}\nabla u\overline{\nabla v}+u\bar v\right\}+\int_{\Gamma_b}k_{b+0}^{-2}D_tN_b^+u\bar v+\int_{\Gamma_a}k_{a-0}^{-2}D_tN_a^-u\bar v,\quad u,v\in H^1_\alpha(\Omega),\tag{4.3}$$

and the Riesz basis properties imply the boundedness of the second and third term on the right-hand side. It is sufficient to prove that the operator corresponding to the form a is strongly elliptic, i.e. that there exist a complex number θ, a positive real ε, and a compact form b with

$$\Re e\,[a(u,\theta u)-b(u,\theta u)] \;\;\geq\;\; \varepsilon\|u\|^2_{H^1_\alpha(\Omega)},\ \forall u\in H^1_\alpha(\Omega).$$

By the assumptions on k in Sect. 3, the numbers $k^{-2}(x_1,x_2)$ are contained in a compact subset of $\{z\in\mathbb{C}\colon \Im m\, z<0 \text{ or } \Im m\, z=0, \Re e\, z>0\}$. Consequently, there exists a $\theta\in\mathbb{C}$ s.t. $\Re e\,\theta k^{-2}\geq\varepsilon$, and the form defined by the first integral on the right-hand side of (4.3) is strongly elliptic.

For the second integral on the right-hand side of (4.3), the function u can be extended to a solution u^e of (3.1) over $[0,2\pi]\times[b,\infty)$ by the Rayleigh expansion w.r.t. the modes $u^+_{b,n}$, and we get (cf. (4.2))

$$u^e(x_1,x_2)=\sum_n c^+_{b,n}u^+_{b,n}(x_1,x_2),\ x_2\geq b,\ \ \|u^e\|_{H^1_\alpha([0,2\pi]\times[b,b+1])}\sim\|u\|_{H^{1/2}_\alpha(\Gamma_b)}\leq C\|u\|_{H^1_\alpha(\Omega)}.$$

Moreover, the mapping $u|_{\Gamma_b}\mapsto u^e|_{\Gamma_{b+1}}$ is compact since this maps by $u_{b,n}(\cdot,b)\mapsto e^{-\lambda_{b,n}}u_{b,n}(\cdot,b)$ with $\Re e\,\lambda_{b,n}\to\infty$. Indeed, $\Re e\,\lambda_{b,n}\to\infty$ follows from $|\lambda_{b,n}|^2\to\infty$ and $|\Im m\,\lambda_{b,n}|\leq C_0+C_0|\Re\lambda_{b,n}|$ (cf. the general-case assumptions (3.7)-(3.9) at the end of Sect. 3) as well as from $\Re e\,\lambda_{b,n}\geq 0$ (recall the choice of λ at the beginning of Sect. 4). Now the second integral takes the form

$$\int_{\Gamma_b}k^{-2}_{b+0}D_tN^+_b u\bar v=\int_{[0,2\pi]\times[b,b+1]}\left\{-k^{-2}_{b+0}\nabla u^e\overline{\nabla v^e}+u^e\bar v^e\right\}+\int_{\Gamma_{b+1}}k^{-2}_{b+0}D_tN^+_{b+1}u^e\bar v^e,$$

where the second integral on the right-hand side corresponds to a compact sesquilinear form and the first integral can be treated as the first integral term on the right-hand side of (4.3). The third integral on the right-hand side of (4.3) can be treated analogously. □

For assumption i), we refer to the two cases discussed in Sect. 3. In particular, for real-valued wavefunctions $k^\pm=k_{b+0},k_{a-0}$, there is no eigenfunction of rank greater than one 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. Then the Riesz-basis property is almost known at least for twice continuously differentiable k. 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 RCWA is needed. Such an adaption might be difficult since it is not clear, which eigenfunction has a 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 geometries Fig. 2 with trivial solutions of the scattering problem with zero incoming waves, 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 is possible but should be an exceptional case.

Remark 4.3. *For the solution theory of boundary value problems in Thm. 4.2, the choice of the radiation conditions Def. 4.1 based on the trace functions $x_1\mapsto k(x_1,b+0)$ and $x_1\mapsto k(x_1,a-0)$ is natural. However, for the SMA, we shall always consider the scattering matrices based on the trace functions $x_1\mapsto k(x_1,b+0)$ and $x_1\mapsto k(x_1,a+0)$. The corresponding solution theory follows from Thm. 4.2 if the continuity $k(x_1,a+0)=k(x_1,a-0)$ is supposed.*

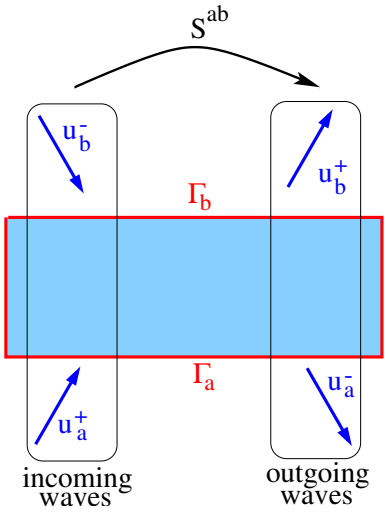

Figure 3: Scattering matrix.

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

5.1 Projections in the space of boundary functions

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 functions, i.e. in the space of Dirichlet and Neumann data. For the eigenvalues $\lambda_{b,n}$ and eigenfunctions $f_{b,n}$ defined with $k(x_1)=k_b(x_1):=k(x_1,b+0)$ as in Sect. 3, we consider upgoing waves u_b^+ and downgoing waves u_b^- together with their boundary data on Γ_b:

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

Due to the Riesz-basis property, each trace of $u_b^\pm\in H_\alpha^{1/2}(\Gamma_b)$ has a unique continuation $u_b^\pm$ to the upper and lower half space, respectively, such that $\nabla\cdot k_b^{-2}\nabla u_b^\pm+u_b^\pm=0$ over the half space and that $k^{-2}D_tN^\pm u_b^\pm=k_b^{-2}\partial_{x_2}u_b^\pm|_{\Gamma_b}\in H_\alpha^{-1/2}(\Gamma_b)$. In this sense, the Dirichlet traces $u_b^\pm$ of the upgoing and downgoing waves can be embedded into the H_α^1 space of wave solutions above and below Γ_b, respectively. Switching from the extensions to the boundary traces on Γ_b, the Dirichlet traces $u_b^\pm$ can be embedded into the space of boundary data consisting of couples of Dirichlet and modulated Neumann data (data of co-normal derivatives). We identify

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

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 $k_b^{-2}\partial_{x_2}u$ 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 $H_\alpha^{1/2}(\Gamma_b)$ and $H_\alpha^{-1/2}(\Gamma_b)$, respectively.

Lemma 5.2. *If the Assumptions (3.7)-(3.9) hold for $k=k_b$, then the space for Dirichlet and modulated Neumann data (data of co-normal derivatives) 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 $H_\alpha^{1/2}(\Gamma_b)\times H_\alpha^{-1/2}(\Gamma_b)$ onto $[H_\alpha^{1/2}\times H_\alpha^{-1/2}]_\pm(\Gamma_b)$ parallel to the space $[H_\alpha^{1/2}\times H_\alpha^{-1/2}]_\mp(\Gamma_b)$ are bounded. In particular, we get $\operatorname{im} P_b^\pm = [H_\alpha^{1/2}\times H_\alpha^{-1/2}]_\pm(\Gamma_b)$.

Proof. From (5.1) we see $D_tN_b^- = -D_tN_b^+ \colon H_\alpha^{1/2}(\Gamma_b)\to k_b^2H_\alpha^{-1/2}(\Gamma_b)$. 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

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

with $[D_tN_b^+]^{-1} = [k_b^{-2}D_tN_b^+]^{-1}k_b^{-2}I \colon k_b^2H_\alpha^{-1/2}(\Gamma_b)\to H_\alpha^{1/2}(\Gamma_b)$. Using (5.1) and the Riesz property (3.9) of the $f_{b,n}$, $n\in\mathbb{N}$, we get the boundedness of the $k_b^{-2}D_tN^+$ and its inverse. The last formula proves the continuity of the projections. □

Note that in the TE case (cf. [7, Lemma 6.1]), the k_b^{-2} factors disappear in (5.3).

Analogously to the projections $P_b^\pm$ in $H_\alpha^{1/2}\times H_\alpha^{-1/2}$ over Γ_b based on the eigenfunctions $f_{b,n}$ for $k_b^2L_b$ with $k_b(x_1)=k(x_1,b+0)$, we have the projections $P_a^\pm$ in $H_\alpha^{1/2}\times H_\alpha^{-1/2}$ over Γ_a based on $k_a(x_1)=k(x_1,a+0)$. In the following we will need a formula (cf. the subsequent (5.4) and (5.5)) for the projections $P_b^\pm$ onto upgoing and downgoing waves over Γ_b applied to upgoing and downgoing waves $u_a^\pm\in\operatorname{im}P_a^\pm$ over Γ_a. 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

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

By the identification (5.2), the restricted projections $P_b^\pm\colon \operatorname{im}P_a^+\to\operatorname{im}P_b^\pm$ and $P_b^\pm\colon\operatorname{im}P_a^-\to\operatorname{im}P_b^\pm$ can be identified by the operators $\mathbf{P}_{a,b}^{\pm+}\colon H_\alpha^{1/2}(\Gamma_a)\to H_\alpha^{1/2}(\Gamma_b)$ and $\mathbf{P}_{a,b}^{\pm-}\colon H_\alpha^{1/2}(\Gamma_a)\to H_\alpha^{1/2}(\Gamma_b)$, respectively, where

$$\mathbf{P}_{a,b}^{\pm+}[u_a^+] = \frac{1}{2}\left[u_a^+ \pm [D_tN_b^+]^{-1}\frac{k_b^2}{k_a^2}D_tN_a^+u_a^+\right],\; \mathbf{P}_{a,b}^{\pm-}[u_a^-] = \frac{1}{2}\left[u_a^- \mp [D_tN_b^+]^{-1}\frac{k_b^2}{k_a^2}D_tN_a^+u_a^-\right]. \tag{5.5}$$

In this sense, we arrive at bounded operators $\mathbf{P}_{a,b}^{\pm,-}\colon H_\alpha^{1/2}\to H_\alpha^{1/2}$ and $\mathbf{P}_{a,b}^{\pm,+}\colon H_\alpha^{1/2}\to H_\alpha^{1/2}$. Though $P_b^\pm$ in (5.4) is a projection, the identified operators $\mathbf{P}_{a,b}^{\pm+}$ and $\mathbf{P}_{a,b}^{\pm-}$ on the right-hand sides of (5.5) are not. The formulas of (5.5) have been derived knowing that the functions $u_a^\pm$ are the Dirichlet traces of outgoing and downgoing waves, respectively. Note that, for the TE case, (5.5) holds with the factor k_b^2/k_a^2 deleted.

5.2 Structure of the S-matrix

In the S-matrix (cf. Fig. 3) the functions $u_a^\pm$ are to be identified with the corresponding pair of trace functions $(u_a^\pm, \pm k_a^{-2} D_t N_a^+ u_a^\pm)$ and $u_b^\pm$ with $(u_b^\pm, \pm k_b^{-2} D_t N_b^+ u_b^\pm)$. Thus the S-matrix maps trace space $[H_\alpha^{1/2} \times H_\alpha^{-1/2}]_+(\Gamma_a) \oplus [H_\alpha^{1/2} \times H_\alpha^{-1/2}]_-(\Gamma_b)$ into $[H_\alpha^{1/2} \times H_\alpha^{-1/2}]_+(\Gamma_b) \oplus [H_\alpha^{1/2} \times H_\alpha^{-1/2}]_-(\Gamma_a)$. So the S-matrix consists of the 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 $\mathbf{S}_{++}^{ab}$, $\mathbf{S}_{+-}^{ab}$, $\mathbf{S}_{-+}^{ab}$, and $\mathbf{S}_{--}^{ab}$ in $H_\alpha^{1/2}$ (cf. (5.2)). Its action corresponds to a system of two linear equations.

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

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

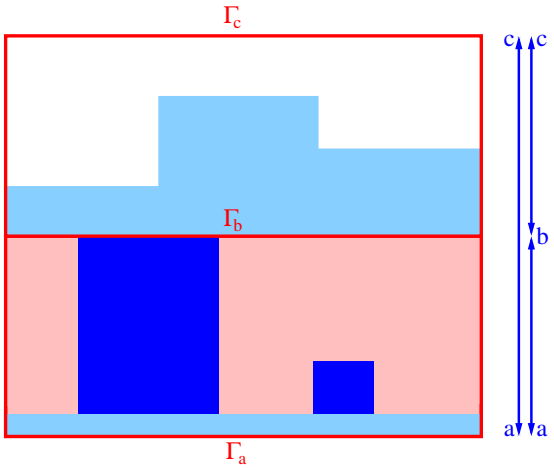

Figure 4: Step from two slices to their union.

5.3 SMA for two adjacent slices

We start the derivation of the algorithm with the case of a grating consisting of two adjacent slices (cf. Fig. 4)). Suppose the S-matrices $\mathbf{S}^{ab}$ and $\mathbf{S}^{bc}$ 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 $\mathbf{S}^{ac}$ between Γ_a and Γ_c looks like? In other words, we know

$$u_c^+ = \mathbf{S}_{++}^{bc} u_b^+ + \mathbf{S}_{+-}^{bc} u_c^- \tag{5.6}$$
$$u_b^- = \mathbf{S}_{-+}^{bc} u_b^+ + \mathbf{S}_{--}^{bc} u_c^- \tag{5.7}$$
$$u_b^+ = \mathbf{S}_{++}^{ab} u_a^+ + \mathbf{S}_{+-}^{ab} u_b^- \tag{5.8}$$
$$u_a^- = \mathbf{S}_{-+}^{ab} u_a^+ + \mathbf{S}_{--}^{ab} u_b^-, \tag{5.9}$$

and we look for

$$u_c^+ = \mathbf{S}_{++}^{ac} u_a^+ + \mathbf{S}_{+-}^{ac} u_c^-, \qquad u_a^- = \mathbf{S}_{-+}^{ac} u_a^+ + \mathbf{S}_{--}^{ac} u_c^-. \tag{5.10}$$

We set the traces $u_b^\pm = u_b^\pm(\cdot, b+0)$ of the functions $u_b^\pm(\cdot,\cdot)$ in the slice between Γ_b and Γ_c (cf. (5.6) and (5.7)) to the corresponding upper output and input functions $u_\pm^b$ of the S-matrix for the slice

between Γ_a and Γ_b (cf. (5.8) and (5.9)). Both traces are in the trace space $H_\alpha^{1/2}(\Gamma_{b+0})$. Automatically, we set the $H_\alpha^{-1/2}(\Gamma_{b+0})$ traces of the co-normal derivatives $k_{b+0}^{-2}\partial_{x_2}u_b^\pm = k^{-2}(\cdot, b+0)D_t N_{b+0}^\pm u_b^\pm$ of the upgoing and downgoing function from the slice between Γ_b and Γ_c to those of the functions between Γ_a and Γ_b. Then we eliminate these unknown functions from the linear system (5.6)–(5.9). Defining $D := (I - \mathbf{S}_{-+}^{bc}\mathbf{S}_{+-}^{ab})$, we arrive at the linear system (5.10) with the operator coefficients

$$\mathbf{S}^{ac} = \begin{pmatrix} \mathbf{S}_{++}^{ac} & \mathbf{S}_{+-}^{ac} \\ \mathbf{S}_{-+}^{ac} & \mathbf{S}_{--}^{ac} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{++}^{bc}\left[I + \mathbf{S}_{+-}^{ab}D^{-1}\mathbf{S}_{-+}^{bc}\right]\mathbf{S}_{++}^{ab} & \mathbf{S}_{+-}^{bc} + \mathbf{S}_{++}^{bc}\mathbf{S}_{+-}^{ab}D^{-1}\mathbf{S}_{--}^{bc} \\ \mathbf{S}_{-+}^{ab} + \mathbf{S}_{--}^{ab}D^{-1}\mathbf{S}_{-+}^{bc}\mathbf{S}_{++}^{ab} & \mathbf{S}_{--}^{ab}D^{-1}\mathbf{S}_{--}^{bc} \end{pmatrix}. \quad (5.11)$$

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 $\mathbf{S}^{ab}$, $\mathbf{S}^{ac}$, and $\mathbf{S}^{bc}$ exist. Furthermore, suppose the wavenumber function $k(x_1, x_2)$ is independent of x_2 in the slices between Γ_a and Γ_b and between Γ_b and Γ_c. Finally, suppose $k_a(x_1) := k(x_1, a)$ and $k_c(x_1) := k(x_1, c)$ are piecewise twice continuously differentiable w.r.t. x_1. Then the operator $D := (I - \mathbf{S}_{-+}^{bc}\mathbf{S}_{+-}^{ab})$ is invertible and Formula (5.11) is correct.*

Proof. First we observe that $\mathbf{S}_{-+}^{bc}$ is compact. Indeed, if the scattering problem is uniquely solvable for the grating between Γ_b and Γ_c, then, for a small $\varepsilon > 0$, the scattering problem is uniquely solvable for the grating between $\Gamma_{b+\varepsilon}$ and Γ_c. Between Γ_b and $\Gamma_{b+\varepsilon}$ we get $k_b = k_{b+\varepsilon}$ and the corresponding scattering matrix is diagonal, i.e. $\mathbf{S}_{\pm\mp}^{b,b+\varepsilon} = 0$ and $\mathbf{S}_{\pm\pm}^{b,b+\varepsilon}$ is diagonal w.r.t. the eigenmodes of the radiation condition. Due to diagonal entries $e^{-\varepsilon\lambda_{b,n}}$ decaying for $n \to \infty$, the $\mathbf{S}_{\pm\pm}^{b,b+\varepsilon}$ are compact. Formula (5.11) for $a = b$, $b = b+\varepsilon$, $c = c$ holds with $D = I$ and provides us with $\mathbf{S}_{-+}^{bc} = \mathbf{S}_{--}^{b,b+\varepsilon}\mathbf{S}_{-+}^{b+\varepsilon,c}\mathbf{S}_{++}^{b,b+\varepsilon}$, which is compact as well.

Now, due to the definition D and due to the compactness of $\mathbf{S}_{-+}^{bc}$, the operator D is a Fredholm operator of index zero. It remains to prove that the codimension of $\operatorname{im} D \subseteq H_\alpha^{1/2}(\Gamma_c)$ 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. (5.7) and (5.8) lead to the system

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

Multiplying the last equation by $\mathbf{S}_{-+}^{bc}$ and adding the result to the first, we arrive at $Du_b^- = \mathbf{S}_{--}^{bc}u_c^-$. From the subsequent Eqns. (6.8) and (6.9), we observe that the image space of D is dense. □

Remark 5.4. *For general slices with x_2-dependent wavenumber function, we conjecture that the product operator $\mathbf{S}_{-+}^{bc}\mathbf{S}_{+-}^{ab}$ is compact as well. Then a violation of the invertibility of D seems to be an exceptional case. If D is not invertible, then a general solver can be applied to (5.7)-(5.8) in order to express $u_b^\pm$ w.r.t. the functions u_a^+ and u_c^-. Substituting these expressions into (5.6) and (5.9), we get an alternative formula for $\mathbf{S}^{ac}$.*

5.4 SMA for all slices

Next we 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 between $\Gamma_{h_{j-1}}$ and Γ_{h_j} by $\mathbf{S}^j$ and the Dirichlet boundary

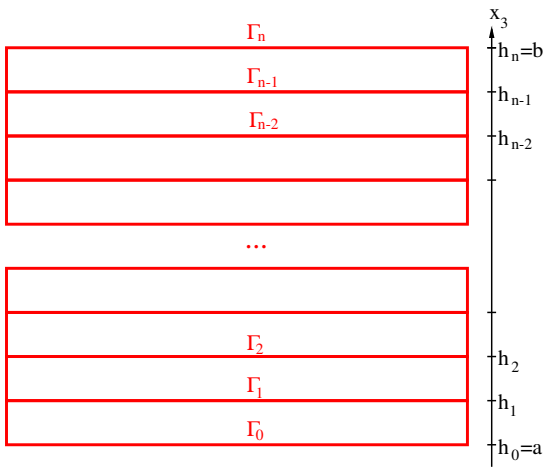

Figure 5: Step from many small slices to their union.

values on the slice boundaries $\Gamma_{h_j} := \Gamma_{h_j+0}$ by $u_j^{\pm}$. Furthermore, we introduce the accumulated S-matrix $\mathbb{S}^j$ over the union over all slices between Γ_{h_0} and Γ_{h_j}. In other words, we have

$$u_j^{\pm} := u_{h_j}^{\pm} \leftrightarrow P_{h_j+0}^{\pm}\big(u|_{\Gamma_{h_j+0}}, k_{h_j+0}^{-2}\partial_{x_2}u|_{\Gamma_{h_j+0}}\big),\ \mathbf{S}^j := \mathbf{S}^{h_{j-1}h_j},\ \mathbb{S}^j := \mathbf{S}^{h_0h_j},\ j=1,\cdots,n\ .$$

Suppose, we can compute the S-matrices $\mathbf{S}^j,\ j=1,\cdots,n$, which requires a solver for the BVP (2.1) between $\Gamma_{h_{j-1}}$ and Γ_{h_j} (cf. the subsequent Sect. 6). With this we get the

Scattering matrix algorithm. (5.12)

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

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

  i)-ii) Iteration for j running from 2 to n:
  Compute $\mathbf{S}^j$ (cf. Sect. 6).
  Apply the two-step formula (5.11) with $\mathbf{S}^{ab}=\mathbb{S}^{j-1}$, $\mathbf{S}^{bc}=\mathbf{S}^j$, and $\mathbf{S}^{ac}=\mathbb{S}^j$
  to compute $\mathbb{S}^j$ from $\mathbb{S}^{j-1}$ and $\mathbf{S}^j$.

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 $u_0^- = \mathbb{S}_{-+}^n u_0^+ + \mathbb{S}_{--}^n u_n^-$.
Compute the Rayleigh coefficients $c_{b,n}^+$ of u_n^+ and $c_{a,n}^-$of u_0^-.
Compute the squared moduli $|c_{b,n}^+|^2$ and $|c_{a,n}^-|^2$ and, by simple scaling (cf. (2.5)), 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 5.5. *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 $x_2 > h_j$ or $x_2 < h_j$, they use this condition to determine the S-matrices for the RCWA.*

Remark 5.6. *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 5.7. *If we are interested in the solution over the slices, then we can go backwards. Using the two-step equations (5.8) and (5.7) for the two slices $\mathbb{S}^{n-1}$ and $\mathbf{S}^n$, we compute $u^{\pm}_{n-1}$. Using (5.8) and (5.7) for the two slices $\mathbb{S}^{n-2}$ and $\mathbf{S}^{n-1}$, we compute $u^{\pm}_{n-2}$. Using (5.8) and (5.7) for the two slices $\mathbb{S}^{n-3}$ and $\mathbf{S}^{n-2}$, we compute $u^{\pm}_{n-3}$. Going recursively up to 1, we get $u_1^{\pm}$ from (5.8) and (5.7) for the two slices $\mathbb{S}^1$ and $\mathbf{S}^2$. 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 $\mathbf{S}^j$ (cf. Sect. 6). Knowing the boundary data u^+_{j-1} and u^-_j, 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

6.1 Auxiliary operators for the representation of the S-matrix

Clearly, the scattering problem over the slice is equivalent to a variational formulation (cf. the corresponding sesquilinear form in (4.3)), which can be solved numerically by FEM combined with a discretization of the nonlocal boundary operators $D_t N_c^+$, $c=a,b$. 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. the subsequent (6.7) and (6.8)), which reduces the scattering problem to the solution of a problem for an operator valued ODE (cf. the subsequent (6.1) and (6.2)).

To prepare the formulas of the S-matrix based on the ODEs (cf. the subsequent (6.7),(6.9), and (6.11)), 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 $P_b^{\pm}$ in $H^{1/2}\times H^{-1/2}$ over Γ_b based on the eigenfunctions $f_{b,n}$ for the differential operator $k^2 L$ 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 $T^{\pm}_{ab}$

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

The operator T^+_{ab} maps $(u^+_a, v^+_a := k_a^{-2} D_t N_a u^+_a)^\top$ (cf. (5.1)) to the vector $(u(\cdot,b-0), v(\cdot,b-0))^\top$, where $(u,v)^\top$ is the solution of the initial value problem (cf. (3.1)-(3.4))

$$\begin{aligned} a)\quad \partial_{x_2}\begin{pmatrix} u(x_1,x_2)\\ v(x_1,x_2)\end{pmatrix} &= \begin{pmatrix} 0 & k^2 I\\ L & 0\end{pmatrix}\begin{pmatrix} u(x_1,x_2)\\ v(x_1,x_2)\end{pmatrix},\\ b)\qquad \begin{pmatrix} u(x_1,a)\\ v(x_1,a)\end{pmatrix} &= \begin{pmatrix} u^+_a(x_1)\\ v^+_a(x_1)\end{pmatrix},\ 0\le x_1\le 2\pi, \end{aligned} \tag{6.1}$$

Note that the operator valued first order ODE system in a) is the order reduction of the operator valued second order ODE $\partial_{x_2} k^{-2}\partial_{x_2} u = Lu$ equivalent to the "Helmholtz" equation (cf. (3.1) and (3.2)). We identify $[P^{\pm}_{b-0}T^+_{ab}]$ (cf. (5.3)) with the bounded operator $\mathbf{PT}^{\pm+}_{ab}: H^{1/2}_\alpha(\Gamma_a)\to H^{1/2}_\alpha(\Gamma_{b-0})$ mapping u^+_a to the wave functions $u^{\pm}_b$ s.t. $u^{\pm}_b \leftrightarrow P^{\pm}_{b-0}\big(u(\cdot,b-0), v(\cdot,b-0)\big)$.

Similarly, the transition operator T^-_{ba} maps $(u^-_{b-0}, v^-_{b-0} := -k^{-2}_{b-0} D_t N_{b-0} u^-_{b-0})^\top$ to the boundary-value couple $(u(\cdot, a+0), v(\cdot, a+0))^\top$, where $(u,v)^\top$ is the solution of the problem

$$\begin{aligned} a) \quad \partial_{x_2} \begin{pmatrix} u(x_1,x_2) \\ v(x_1,x_2) \end{pmatrix} &= \begin{pmatrix} 0 & k^2 I \\ L & 0 \end{pmatrix} \begin{pmatrix} u(x_1,x_2) \\ v(x_1,x_2) \end{pmatrix}, \\ b) \qquad \begin{pmatrix} u(x_1,b) \\ v(x_1,b) \end{pmatrix} &= \begin{pmatrix} u^-_{b-0}(x_1) \\ v^-_{b-0}(x_1) \end{pmatrix}, \ 0 \le x_1 \le 2\pi. \end{aligned} \tag{6.2}$$

Using this, we identify $[P^\pm_{a+0} T^-_{ba}]$ with the bounded operator $\mathbf{PT}^{\pm-}_{ba}: H^{1/2}_\alpha(\Gamma_{b-0}) \to H^{1/2}_\alpha(\Gamma_a)$ mapping u^-_{b-0} to $u^\pm_{a+0}$ s.t. $u^\pm_{a+0} \leftrightarrow P^\pm_{a+0}\big(u(\cdot,a+0), v(\cdot,a+0)\big)$.

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 $P^\pm_{b-0} = P^\pm_a := P^\pm_{a+0}$. The transition operators $\mathbf{PT}^{++}_{ab}$ and $\mathbf{PT}^{--}_{ba}$ are given by (cf. (4.1))

$$\mathbf{PT}^{++}_{ab}[u^+_{a,n}(\cdot,a)] \ = \ u^+_{a,n}(\cdot,b), \qquad \mathbf{PT}^{--}_{ba}[u^-_{a,n}(\cdot,b)] \ = \ u^-_{a,n}(\cdot,a), \tag{6.3}$$

and $\mathbf{PT}^{-+}_{ab} = 0 = \mathbf{PT}^{+-}_{ba}$. W.r.t. the basis $f_{a,n}$, $n \in \mathbb{N}$, both transition operators $\mathbf{PT}^{++}_{ab}$ and $\mathbf{PT}^{--}_{ba}$ have the same diagonal matrix $(e^{-\lambda_{a,n}[b-a]}\delta_{m,n})_{m,n\in\mathbb{N}}$ and are bounded.

6.2 Representation of the S-matrix

Next we derive the formula for the S-matrix. The boundary value functions v^+_a and $v^-_a = 0$ over the straight line Γ_a lead to a solution of the scattering problem with the Γ_b boundary value functions $v^+_b \leftrightarrow P^+_b T^+_{ab} v^+_a$ and $v^-_b \leftrightarrow P^-_b T^+_{ab} v^+_a$ over Γ_b. Clearly, using the identification (5.2) and the operators $\mathbf{P}^{\pm+}_{b-0,b}$ and $\mathbf{P}^{\pm-}_{b-0,b}$ of (5.5), the operators $P^+_b T^+_{ab}$ and $P^-_b T^+_{ab}$ are identified by the operators $\mathbf{PPT}^{++}_{ab} := [\mathbf{P}^{++}_{b-0,b}\mathbf{PT}^{++}_{ab} + \mathbf{P}^{+-}_{b-0,b}\mathbf{PT}^{-+}_{ab}]$ and $\mathbf{PPT}^{-+}_{ab} := [\mathbf{P}^{-+}_{b-0,b}\mathbf{PT}^{++}_{ab} + \mathbf{P}^{--}_{b-0,b}\mathbf{PT}^{-+}_{ab}]$, respectively. In other words, we get $v^+_b = \mathbf{PPT}^{++}_{ab} v^+_a$ as well as $v^-_b = \mathbf{PPT}^{-+}_{ab} v^+_a$ over Γ_b. We arrive at

$$\mathbf{S}^{ab}: \begin{pmatrix} v^+_a \\ \mathbf{PPT}^{-+}_{ab} v^+_a \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{PPT}^{++}_{ab} v^+_a \\ 0 \end{pmatrix}. \tag{6.4}$$

On the other hand, take a downgoing v^-_{b-0} over Γ_b. Then the boundary values $v^+_b \leftrightarrow P^+_b v^-_{b-0}$ and $v^-_b \leftrightarrow P^-_b v^-_{b-0}$ over Γ_b lead us to the other boundary values $v^+_a \leftrightarrow P^+_a T^-_{ba} v^-_{b-0}$ and $v^-_a \leftrightarrow P^-_a T^-_{ba} v^-_{b-0}$ over Γ_a. Recall that the operators $P^+_a T^-_{ba}$ and $P^-_a T^-_{ba}$ are identified by $\mathbf{PT}^{+-}_{ba}$ and $\mathbf{PT}^{--}_{ba}$, respectively. So we get $v^+_a = \mathbf{PT}^{+-}_{ba} v^+_{b-0}$ as well as $v^-_a = \mathbf{PT}^{--}_{ba} v^+_{b-0}$ over Γ and arrive at

$$\mathbf{S}^{ab}: \begin{pmatrix} \mathbf{PT}^{+-}_{ba} v^-_{b-0} \\ \mathbf{P}^{--}_{b-0,b} v^-_{b-0} \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{P}^{+-}_{b-0,b} v^-_{b-0} \\ \mathbf{PT}^{--}_{ba} v^-_{b-0} \end{pmatrix}. \tag{6.5}$$

For the functions $u^+_a = v^+_a + \mathbf{PT}^{+-}_{ba} v^-_{b-0}$ and $u^-_b = \mathbf{PPT}^{-+}_{ab} v^+_a + \mathbf{P}^{--}_{b-0,b} v^-_{b-0}$, Eqns. (6.4) and (6.5) yield

$$\begin{aligned} \begin{pmatrix} u^+_a \\ u^-_b \end{pmatrix} &= \begin{pmatrix} I|_{\operatorname{im}\mathbf{P}^+_a} & \mathbf{PT}^{+-}_{ba} \\ \mathbf{PPT}^{-+}_{ab} & \mathbf{P}^{--}_{b-0,b} \end{pmatrix} \begin{pmatrix} v^+_a \\ v^-_{b-0} \end{pmatrix}, \\ \mathbf{S}^{ab} \begin{pmatrix} u^+_a \\ u^-_b \end{pmatrix} &= \begin{pmatrix} \mathbf{PPT}^{++}_{ab} & \mathbf{P}^{+-}_{b-0,b} \\ 0 & \mathbf{PT}^{--}_{ba} \end{pmatrix} \begin{pmatrix} v^+_a \\ v^-_{b-0} \end{pmatrix}. \end{aligned} \tag{6.6}$$

Assuming that the determinant operator $D_{ab}^{-}:=\{\mathbf{P}_{b-0,b}^{--}-\mathbf{PPT}_{ab}^{-+}\mathbf{PT}_{ba}^{+-}\}\colon H_\alpha^{1/2}(\Gamma_{b-0})\to H_\alpha^{1/2}(\Gamma_b)$ of the first matrix in (6.6) is invertible, we arrive at

$$\begin{aligned}\mathbf{S}^{ab} &= \begin{pmatrix}\mathbf{PPT}_{ab}^{++} & \mathbf{P}_{b-0,b}^{+-}\\ 0 & \mathbf{PT}_{ba}^{--}\end{pmatrix}\begin{pmatrix} I|_{\operatorname{im}\mathbf{P}_a^+}+\mathbf{PT}_{ba}^{+-}[D_{ab}^{-}]^{-1}\mathbf{PPT}_{ab}^{-+} & -\mathbf{PT}_{ba}^{+-}[D_{ab}^{-}]^{-1}\\ -[D_{ab}^{-}]^{-1}\mathbf{PPT}_{ab}^{-+} & [D_{ab}^{-}]^{-1}\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{PPT}_{ab}^{++}-\left[\mathbf{P}_{b-0,b}^{+-}-\mathbf{PPT}_{ab}^{++}\mathbf{PT}_{ba}^{+-}\right][D_{ab}^{-}]^{-1}\mathbf{PPT}_{ab}^{-+} & \left[\mathbf{P}_{b-0,b}^{+-}-\mathbf{PPT}_{ab}^{++}\mathbf{PT}_{ba}^{+-}\right][D_{ab}^{-}]^{-1}\\ -\mathbf{PT}_{ba}^{--}[D_{ab}^{-}]^{-1}\mathbf{PPT}_{ab}^{-+} & \mathbf{PT}_{ba}^{--}[D_{ab}^{-}]^{-1}\end{pmatrix}.\end{aligned} \tag{6.7}$$

Note that, for the case of x_2-invariant wavenumber $k(x_1,x_2)=k(x_1,a)$ in the slice $[0,2\pi]\times[a,b)$, the formula (6.7) simplifies to $\mathbf{S}^{ab}\colon H_\alpha^{1/2}(\Gamma_a)\times H_\alpha^{1/2}(\Gamma_b)\to H_\alpha^{1/2}(\Gamma_b)\times H_\alpha^{1/2}(\Gamma_a)$ with

$$\mathbf{S}^{ab}=\begin{pmatrix}\left\{\mathbf{P}_{b-0,b}^{++}-\mathbf{P}_{b-0,b}^{+-}[D_{ab}^{-}]^{-1}\mathbf{P}_{b-0,b}^{-+}\right\}\mathbf{PT}_{ab}^{++} & \mathbf{P}_{b-0,b}^{+-}[D_{ab}^{-}]^{-1}\\ -\mathbf{PT}_{ba}^{--}[D_{ab}^{-}]^{-1}\mathbf{P}_{b-0,b}^{-+}\mathbf{PT}_{ab}^{++} & \mathbf{PT}_{ba}^{--}[D_{ab}^{-}]^{-1}\end{pmatrix}, \tag{6.8}$$

where $D_{ab}^{-}=\mathbf{P}_{ab}^{--}\colon H_\alpha^{1/2}(\Gamma_a)\to H_\alpha^{1/2}(\Gamma_b)$ (cf. (5.5)).

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 $a\le x_2<b$. In the substrate suppose $k(x_1,x_2)=k_a(x_1)$ for $x_2\le a$ and in the cover material $k(x_1,x_2)=k_b(x_1)$ for $b\le x_2$. Furthermore suppose that k_a and k_b are piecewise twice continuously differentiable w.r.t. x_1. Finally suppose the BVP (2.1) over this grating is uniquely solvable (cf. Thm. 4.2), i.e., that there exists the bounded S-matrix S^{ab}. Then the operator $D_{ab}^{-}=\mathbf{P}_{ab}^{--}$ is invertible and formula (6.8) holds true.*

Proof. By the piecewise differentiability of k_a and k_b and by (5.5), we get $\mathbf{PPT}_{b-0,b}^{\pm,+}v_a^+\in H_\alpha^{1/2}(\Gamma_b)$, $\mathbf{P}_{b-0,b}^{\pm,-}v_{b-0}^-\in H_\alpha^{1/2}(\Gamma_b)$, and the boundedness of the linear operator $\mathbf{P}_{ab}^{--}$.

If $v_a^+=0$ and $\mathbf{P}_{ab}^{--}v_{b-0}^-=0$ and if u_a^+ and u_b^- are defined by the first equation of (6.6), then, due to $\mathbf{PT}_{ba}^{+-}=0$, we get $u_a^+=0$ and $u_b^-=0$ s.t. the left-hand side of the second equation of (6.6) satisfies $\mathbf{S}^{ab}(u_a^+,u_b^-)^\top=(0,0)^\top$. In particular, $\mathbf{PT}_{ba}^{--}v_{b-0}^-=0$, and by (6.3) we obtain $v_{b-0}^-=0$. In other words, the null space of the operator $\mathbf{P}_{ab}^{--}\colon H_\alpha^{1/2}(\Gamma_a)\to H_\alpha^{1/2}(\Gamma_b)$ is trivial.

Now we shall show that the image of D_{ab}^{-} coincides with the space $H_\alpha^{1/2}(\Gamma_b)$. Consider the scattering problem (2.1) with the incoming functions $u_a^+=0$ from below and an arbitrary $u_b^-\in H_\alpha^{1/2}(\Gamma_b)$ from above. Then, for the scattering problem, there exists a unique wave solution $u\in H_\alpha^1([0,2\pi]\times[a,b])$ and a unique solution pair $u_b^+\in H_\alpha^{1/2}(\Gamma_b)$ and $u_a^-\in H_\alpha^{1/2}(\Gamma_a)$. Clearly, $u_{b-0}^{\pm}=\mathbf{P}_{b,b-0}^{\pm+}u_b^+ +\mathbf{P}_{b,b-0}^{\pm-}u_b^-$ s.t. we get $u_{b-0}^+=\mathbf{PT}_{ab}^{++}u_a^+=0$ and $u_b^++u_b^-=u_{b-0}^++u_{b-0}^-=u_{b-0}^-$. Hence, we arrive at the formula $u_b^-=\mathbf{P}_{b-0,b}^{--}u_{b-0}^-$. In other words, all functions $u_b^-\in H_\alpha^{1/2}(\Gamma_b)$ are in the image of $\mathbf{P}_{b-0,b}^{--}$. Moreover, $u_{b-0}^-=u_b^-+u_b^+$ leads to the inverse $[D_{ab}^{-}]^{-1}=I+\mathbf{S}_{+-}^{ab}$. □

The matrix $\mathbf{S}_m^{ab}$ of $\mathbf{S}^{ab}$ w.r.t. the four bases, namely with $f_{a,n}\in H_\alpha^{1/2}(\Gamma_a), n\in\mathbb{N}$ in the sense that $f_{a,n}\leftrightarrow(f_{a,n},k_a^{-2}\lambda_{a,n}f_{a,n})$, with $f_{a,n}\in H_\alpha^{1/2}(\Gamma_a), n\in\mathbb{N}$ s.t. $f_{a,n}\leftrightarrow(f_{a,n},-k_a^{-2}\lambda_{a,n}f_{a,n})$, with the basis $f_{b,n}\in H_\alpha^{1/2}(\Gamma_b), n\in\mathbb{N}$ such that $f_{b,n}\leftrightarrow(f_{b,n},k_b^{-2}\lambda_{b,n}f_{b,n})$, and with $f_{b,n}\in H_\alpha^{1/2}(\Gamma_b), n\in\mathbb{N}$ s.t. $f_{b,n}\leftrightarrow(f_{b,n},-k_b^{-2}\lambda_{b,n}f_{b,n})$ is

$$\mathbf{S}_m^{ab}=\begin{pmatrix}\left(\Theta_{++}^{ab}-\Theta_{+-}^{ab}[\Theta_{--}^{ab}]^{-1}\Theta_{-+}^{ab}\right)\mathbf{T}^{ab} & \Theta_{+-}^{ab}[\Theta_{--}^{ab}]^{-1}\\ -\mathbf{T}^{ab}[\Theta_{--}^{ab}]^{-1}\Theta_{-+}^{ab}\mathbf{T}^{ab} & \mathbf{T}^{ab}[\Theta_{--}^{ab}]^{-1}\end{pmatrix},\ \mathbf{T}^{ab}=\left(e^{-(b-a)\lambda_{a,n}}\delta_{n,m}\right)_{n,m\in\mathbb{N}} \tag{6.9}$$

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 basis $(f_{a,n},(-1)^l k_a^{-2}\lambda_{a,n}f_{a,n}),\ n\in\mathbb{N},\ l=0,1$ to basis $(f_{b,n},(-1)^l k_b^{-2}\lambda_{a,n}f_{b,n}),\ n\in\mathbb{N},\ l=0,1$.

6.3 Alternative representation of the S-matrix

Finally, we shall present a formula for the S-matrix alternative to (6.7), which contains unbounded transition operators. Together with a truncation of the infinite series, this formula is frequently used and yields, in most cases, almost the same computational 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 (6.7). However, the unbounded operators make the analysis of the RCWA difficult. So we mention only the formula. For simplicity, 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 (6.8) of (6.7)).

Define the transition operator T_{ab} by (6.1) but with general initial values (u_a, v_a) instead of (u_a^+, v_a^+). Then the T-matrix $\mathbb{T}_{ab}$ is defined by

$$\begin{pmatrix} u_b^+ \\ u_b^- \end{pmatrix} \;=\; \mathbb{T}_{ab} \begin{pmatrix} u_a^+ \\ u_a^- \end{pmatrix} \;:=\; \begin{pmatrix} [P_b^+ \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^+}] & [P_b^+ \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^-}] \\ [P_b^- \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^+}] & [P_b^- \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^-}] \end{pmatrix} \begin{pmatrix} u_a^+ \\ u_a^- \end{pmatrix}. \tag{6.10}$$

Supposing the existence of the inverse of $E_{ab}^- := [P_b^- \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^-}]$, writing the vector equation (6.10) 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

$$S^{ab} = \begin{pmatrix} [P_b^+ \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^+}] - [P_b^+ \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^-}][E_{ab}^-]^{-1}[P_b^- \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^+}] & [P_b^+ \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^-}][E_{ab}^-]^{-1} \\ -[E_{ab}^-]^{-1}[P_b^- \mathbb{T}_{ab}|_{\mathrm{im}\, P_a^+}] & [E_{ab}^-]^{-1} \end{pmatrix}. \tag{6.11}$$

If this formula is used for the SMA of (5.12), 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 (5.11), replacing (5.6)–(5.7) by (6.10).

7 Discretization used by RCWA and FMM

7.1 Discretization by truncated Fourier series

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 u 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. Here we use the finite-section operator $\mathcal{P}_N$ acting on univariate functions depending on x_1 and use the same symbol $\mathcal{P}_N$ for the operator $\mathcal{P}_N\otimes I$ on functions depending on $(x_1,x_2)^\top$, which truncate the Fourier series w.r.t. x_1 only (cf. (7.1)). Setting $v:=k^{-2}\partial_{x_2}u$ and $\vec{u}:=(u,v)^\top$, the partial differential equation $\nabla\cdot k^{-2}\nabla u+u=0$ is equivalent (compare (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 (cf. (6.1) and (6.2)),

$$M_{x_2}:=\begin{pmatrix}0 & k^2(\cdot,x_2)I\\ L_{x_2} & 0\end{pmatrix},\quad L_{x_2}u:=-\partial_{x_1}k^{-2}(\cdot,x_2)\partial_{x_1}u-u,$$

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

$$\begin{aligned}M_{x_2,N} &:= \begin{pmatrix}0 & [\mathcal{P}_Nk^{-2}(\cdot,x_2)I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1}\\ L_{x_2,N} & 0\end{pmatrix},\\ L_{x_2,N}u_N &:= [\mathcal{P}_NL_{x_2}|_{\mathrm{im}\,\mathcal{P}_N}]\,u_N \;=\; -\partial_{x_1}\left[\mathcal{P}_Nk^{-2}(\cdot,x_2)I|_{\mathrm{im}\,\mathcal{P}_N}\right]\partial_{x_1}u_N-u_N.\end{aligned}\tag{7.2}$$

Note that the matrix of $[\mathcal{P}_Nk^{-2}(\cdot,x_2)I|_{\mathrm{im}\,\mathcal{P}_N}]$ w.r.t. the basis functions $x_1\mapsto e^{\mathbf{i}(\alpha+l)x_1}$, $-N\le l\le N$ is a truncated Toeplitz matrix and that of operator $[\mathcal{P}_N\partial_{x_1}|_{\mathrm{im}\,\mathcal{P}_N}]=\partial_{x_1}|_{\mathrm{im}\,\mathcal{P}_N}$ 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 $[\mathcal{P}_Ng^{-1}I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1}u$ instead of $[\mathcal{P}_NgI|_{\mathrm{im}\,\mathcal{P}_N}]u$ improves the approximation (cf. [10]) if gu is smoother than u. Additionally, in (7.2) the inverse matrix of the Galerkin approximation appears naturally from the reduction of the second-order differential equation $\partial_{x_2}[\mathcal{P}_Nk^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]\partial_{x_2}u_N=-\partial_{x_1}[\mathcal{P}_Nk^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]\partial_{x_1}u_N-u_N$ to a system of two first-order equations.*

Starting from (7.2), we define the eigenfunction, 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 $H_\alpha^{\pm1/2}$ with a norm uniformly bounded w.r.t. the index 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 $\mathrm{im}\,\mathcal{P}_{N,N}^c$. Due to the assumptions $\Re e\,k>0$ and $\Im m\,k\ge0$, there is a $\zeta\in\mathbb{C}$ s.t. $\Re e\,[\zeta k_c^{-2}]\ge\varepsilon>0$ s.t. the real part of the operator of multiplication by $[\zeta k_c^{-2}]$ is positive definite. Hence, to the bounded multiplication operators $k_c^{-2}I\in\mathcal{L}(H_\alpha^s),\ s\in\mathbb{R}$, the Galerkin method applies, the operators $\mathcal{P}_Nk_c^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}$ are invertible, and the $[\mathcal{P}_Nk_c^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1}$ are bounded uniformly w.r.t. the index N. If k_c is piecewise twice continuously differentiable, then $k_cI\in\mathcal{L}(H_\alpha^s)$ for $|s|<1/2$. If, additionally, it is continuous, then $k_cI\in\mathcal{L}(H_\alpha^s)$ for $|s|<3/2$. 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))

$$[\mathcal{P}_Nk_c^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]^{-1}L_{c,N}f_{c,n,N}+\lambda_{c,n,N}^2f_{c,n,N}=0,\ L_{c,N}u_N:=\partial\,[\mathcal{P}_Nk_c^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]\partial\,u_N+u_N.\tag{7.3}$$

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 $\Delta u+k^2u=0$ in part a) of (2.1) and, instead of (7.3), we have the simpler discretized eigenvalue equation

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

To approximate (5.1), we set

$$D_tN_{c,N}^{\pm}\Big\{\sum_{n=1}^{2N+1}\xi_nf_{c,n,N}\Big\}=\pm\sum_{\substack{n=1\\|\lambda_{c,n,N}|>\varepsilon_{EVD}}}^{2N+1}\xi_n\lambda_{c,n,N}f_{c,n,N}\pm\sum_{\substack{n=1\\|\lambda_{c,n,N}|\le\varepsilon_{EVD}}}^{2N+1}\xi_nf_{c,n,N},\quad\xi_n\in\mathbb{C},\tag{7.5}$$

where $\varepsilon_{EVD}>0$ is chosen sufficiently small. In fact this threshold number is to be chosen such that the eigenvalues $\lambda_{c,n,N}$ approximating the values $\lambda_{c,n}=0$ are caught by the condition $|\lambda_{c,n,N}|\le\varepsilon_{EVD}$. Similarly to Equ. (5.2), we can identify the Dirichlet data $u^{\pm}_{c,N}=\sum_{n=1}^{2N+1}\xi_n f_{c,n,N}$ for upgoing and downgoing waves, respectively, with couples of discrete Dirichlet and Neumann data s.t.

$$\begin{aligned} u^{\pm}_{c,N} &\leftrightarrow (u^{\pm}_{c,N},[\mathcal{P}^c_N k_c^{-2}|_{\mathrm{im}\,\mathcal{P}^c_N}]D_t N^{\pm}_{c,N}u^{\pm}_{c,N}),\ k_c:=k(\cdot,c+0),\\ H^{1/2}_\alpha(\Gamma_c)\supset \mathrm{im}\,\mathcal{P}^c_N &\leftrightarrow \left[\mathrm{im}\,\mathcal{P}^c_{N,N}\right]_{\pm}(\Gamma_c)\subseteq \mathrm{im}\,\mathcal{P}^c_{N,N}\subset H^{1/2}_\alpha(\Gamma_c)\times H^{-1/2}_\alpha(\Gamma_c). \end{aligned} \tag{7.6}$$

Like in Lemma 5.2 the full space $\mathrm{im}\,\mathcal{P}^c_{N,N}$ is the direct sum of the two subspaces $[\mathrm{im}\,\mathcal{P}^c_{N,N}]_{\pm}$, and we denote the projection $\mathrm{im}\,\mathcal{P}^c_{N,N}\to[\mathrm{im}\,\mathcal{P}^c_{N,N}]_+$ onto the upgoing waves parallel to the downgoing waves by $P^+_{c,N}$. We set $P^-_{c,N}:=I-P^+_{c,N}$, which projects $\mathrm{im}\,\mathcal{P}^c_{N,N}\to[\mathrm{im}\,\mathcal{P}^c_{N,N}]_-$. So the space of Dirichlet data $\mathrm{im}\,\mathcal{P}^c_N$ of upgoing waves is identified with the space $\mathrm{im}\,P^+_{c,N}\subset\mathrm{im}\,\mathcal{P}^c_{N,N}$. The space of Dirichlet data $\mathrm{im}\,\mathcal{P}^c_N$ of downgoing waves is identified with $\mathrm{im}\,P^-_{c,N}\subset\mathrm{im}\,\mathcal{P}^c_{N,N}$. Analogously to (5.5), we set

$$\begin{aligned} \mathbf{P}^{\pm+}_{ab,N}[u^+_{a,N}] &= \frac{1}{2}\Big[u^+_{a,N}\pm[D_tN^+_{b,N}]^{-1}\left[\mathcal{P}_N k_b^{-2}|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\mathcal{P}_N k_a^{-2}|_{\mathrm{im}\,\mathcal{P}_N}\,D_tN^+_{a,N}u^+_{a,N}\Big],\\ \mathbf{P}^{\pm-}_{ab,N}[u^-_{a,N}] &= \frac{1}{2}\Big[u^-_{a,N}\mp[D_tN^+_{b,N}]^{-1}\left[\mathcal{P}_N k_b^{-2}|_{\mathrm{im}\,\mathcal{P}_N}\right]^{-1}\mathcal{P}_N k_a^{-2}|_{\mathrm{im}\,\mathcal{P}_N}\,D_tN^+_{a,N}u^-_{a,N}\Big]. \end{aligned} \tag{7.7}$$

Now, replacing the "Helmholtz" equation $\nabla\cdot k^{-2}\nabla u+u=0$ by the discretized operator valued ODE $\partial_{x_2}[\mathcal{P}_N k^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]\partial_{x_2}u_N=-\partial_{x_1}[\mathcal{P}_N k^{-2}I|_{\mathrm{im}\,\mathcal{P}_N}]\partial_{x_1}u_N-u_N$ and using the just mentioned discretized splitting into upgoing and downgoing waves, we can consider the discretized version of the BVP (2.1) over the full grating and over each slice of Sect. 5. We get a discretized solution operator (scattering matrix) S^{ab}_N mapping $\mathrm{im}\,P^+_{a,N}\otimes\mathrm{im}\,P^-_{b,N}$ into $\mathrm{im}\,P^+_{b,N}\otimes\mathrm{im}\,P^-_{a,N}$ s.t.

$$S^{ab}_N = \begin{pmatrix} S^{ab}_{++,N} & S^{ab}_{+-,N}\\ S^{ab}_{-+,N} & S^{ab}_{--,N}\end{pmatrix},$$

with the entries $S^{ab}_{++,N}:=P^+_b S^{ab}_N|_{\mathrm{im}\,P^+_a}$, $S^{ab}_{+-,N}:=P^+_b S^{ab}_N|_{\mathrm{im}\,P^-_b}$, $S^{ab}_{-+,N}:=P^-_a S^{ab}_N|_{\mathrm{im}\,P^+_a}$ as well as $S^{ab}_{--,N}:=P^-_a S^{ab}_N|_{\mathrm{im}\,P^-_b}$. This S^{ab}_N is identified by the operator $\mathbf{S}^{ab}_N$ mapping $\mathrm{im}\,\mathcal{P}^a_N\otimes\mathrm{im}\,\mathcal{P}^b_N$ into $\mathrm{im}\,\mathcal{P}^b_N\otimes\mathrm{im}\,\mathcal{P}^a_N$ (compare (6.7)).

$$\begin{aligned} \mathbf{S}^{ab}_N &= \begin{pmatrix} \mathbf{S}^{ab}_{++,N} & \mathbf{S}^{ab}_{+-,N}\\ \mathbf{S}^{ab}_{-+,N} & \mathbf{S}^{ab}_{--,N}\end{pmatrix},\\ &\mathbf{S}^{ab}_{++,N}:=\mathbf{PPT}^{++}_{ab,N}-\left[\mathbf{P}^{+-}_{b-0,b,N}-\mathbf{PPT}^{++}_{ab,N}\mathbf{PT}^{+-}_{ba,N}\right][D^-_{ab,N}]^{-1}\mathbf{PPT}^{-+}_{ab,N},\\ &\mathbf{S}^{ab}_{+-,N}:=\left[\mathbf{P}^{+-}_{b-0,b,N}-\mathbf{PPT}^{++}_{ab,N}\mathbf{PT}^{+-}_{ba,N}\right][D^-_{ab,N}]^{-1},\\ &\mathbf{S}^{ab}_{-+,N}:=-\mathbf{PT}^{--}_{ba,N}[D^-_{ab,N}]^{-1}\mathbf{PPT}^{-+}_{ab,N},\\ &\mathbf{S}^{ab}_{--,N}:=\mathbf{PT}^{--}_{ba,N}[D^-_{ab,N}]^{-1}. \end{aligned} \tag{7.8}$$

Here the operator ingredients are defined as follows: The discretized transitions $T^+_{ab,N}$ and $T^-_{ba,N}$ are

the solution operators of the initial value problems for the discretized ODE

$$\begin{aligned} a) \quad \partial_{x_2}\begin{pmatrix} u_N(x_1,x_2) \\ v_N(x_1,x_2)\end{pmatrix} &= M_{x_2,N}\begin{pmatrix} u_N(x_1,x_2) \\ v_N(x_1,x_2)\end{pmatrix}, \\ b) \quad \begin{pmatrix} u_N(x_1,a) \\ v_N(x_1,a)\end{pmatrix} &= \begin{pmatrix} u^+_{a,N}(x_1) \\ v^+_{a,N}(x_1)\end{pmatrix},\ 0\le x_1\le 2\pi, \end{aligned}$$

and

$$\begin{aligned} a) \quad \partial_{x_2}\begin{pmatrix} u_N(x_1,x_2) \\ v_N(x_1,x_2)\end{pmatrix} &= M_{x_2,N}\begin{pmatrix} u_N(x_1,x_2) \\ v_N(x_1,x_2)\end{pmatrix}, \\ b) \quad \begin{pmatrix} u_N(x_1,b) \\ v_N(x_1,b)\end{pmatrix} &= \begin{pmatrix} u^-_{b-0,N}(x_1) \\ v^-_{b-0,N}(x_1)\end{pmatrix},\ 0\le x_1\le 2\pi, \end{aligned}$$

corresponding to the initial value problems (6.1) and (6.2), respectively. We identify $[P^{\pm}_{b-0,N}T^+_{ab,N}]$ with the bounded transition operator $\mathbf{PT}^{\pm+}_{ab,N}\colon \operatorname{im}\mathcal{P}^a_N \to \operatorname{im}\mathcal{P}^{b-0}_N$ mapping the truncated Fourier series $u^+_{a,N}$ to $u^{\pm}_{b,N}$ with $u^{\pm}_{b,N} \leftrightarrow P^{\pm}_{b-0,N}\big(u_N(\cdot,b-0),v_N(\cdot,b-0)\big)$. Similarly, we identify $[P^{\pm}_{a,N}T^+_{ba,N}]$ with the bounded transition operator $\mathbf{PT}^{\pm+}_{ba,N}\colon \operatorname{im}\mathcal{P}^{b-0}_N \to \operatorname{im}\mathcal{P}^a_N$ mapping the truncated Fourier series $u^+_{b-0,N}$ to $u^{\pm}_{a,N}$ with $u^{\pm}_{a,N} \leftrightarrow P^{\pm}_{a,N}\big(u_N(\cdot,a+0),v_N(\cdot,a+0)\big)$. Further, the operators $P^{\pm}_{b,N}T^+_{ab,N}$ are identified by the operators $\mathbf{PPT}^{\pm+}_{ab,N} := [\mathbf{P}^{\pm+}_{b-0,b,N}\mathbf{PT}^{++}_{ab,N} + \mathbf{P}^{\pm-}_{b-0,b,N}\mathbf{PT}^{-+}_{ab,N}]$. Finally, the discretization of operator D^-_{ab} is defined as $D^-_{ab,N} := \{\mathbf{P}^{--}_{b-0,b,N} - \mathbf{PPT}^{-+}_{ab,N}\mathbf{PT}^{+-}_{ba,N}\}\colon \operatorname{im}\mathcal{P}^{b-0}_N \to \operatorname{im}\mathcal{P}^b_N$. In particular, for x_2-independent wavenumbers $k(x_1,x_2)=k(x_1,a)$, $0\le x_1<2\pi$, the representation (7.8) turns into (compare (6.8))

$$\mathbf{S}^{ab}_N = \begin{pmatrix} \{\mathbf{P}^{++}_{b-0,b,N} - \mathbf{P}^{+-}_{b-0,b,N}[D^-_{ab,N}]^{-1}\mathbf{P}^{-+}_{b-0,b,N}\}\mathbf{PT}^{++}_{ab,N} & \mathbf{P}^{+-}_{b-0,b,N}[D^-_{ab,N}]^{-1} \\ -\mathbf{PT}^{--}_{ba,N}[D^-_{ab,N}]^{-1}\mathbf{P}^{-+}_{b-0,b,N}\mathbf{PT}^{++}_{ab,N} & \mathbf{PT}^{--}_{ba,N}[D^-_{ab,N}]^{-1}\end{pmatrix}, \quad (7.9)$$
$$D^-_{ab,N} = \mathbf{P}^{--}_{b-0,b,N}\colon \operatorname{im}\mathcal{P}^{b-0}_N \to \operatorname{im}\mathcal{P}^b_N.$$

We shall discuss the computation of the discretization $\mathbf{PT}^{++}_{ab,N}$ and $\mathbf{PT}^{--}_{ba,N}$ in Subsect. 7.2. If this is done, then Equation (7.8) enables us to compute $\mathbf{S}^{ab}_N$. Approximating $\mathbf{S}^j = \mathbf{S}^{h_{j-1}h_j}$ by the discretized $\mathbf{S}^j_N = \mathbf{S}^{h_{j-1}h_j}_N$ and the incoming waves u^-_n and u^+_0 by their truncations $u^-_{n,N} := \mathcal{P}^{h_n}_N u^-_n$ and $u^+_{0,N} := \mathcal{P}^{h_0}_N u^+_0$, we can perform the SMA (5.12) on the discrete level.

7.2 Discretization of the ODE over the slices

To get a first method 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, the numerical methods are not stable, and the integration error blows up for large widths $h_j - h_{j-1}$ of the slice. To overcome this problem (analysis still open), the widths $h_j - h_{j-1}$ are reduced by increasing the number of slices n, and with thin slices a stable algorithm is achieved. Of course, the computing time increases with increasing n.

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 to the ODE solution as well. E.g. for $\mathbf{PT}^{++}_{ab,N}$, 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

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

for any $\xi_n \in \mathbb{C}$. A similar formula holds for $\mathbf{PT}^{--}_{ba,N}$. Note that these formulas reveal the importance of the decomposition of the waves into upgoing and downgoing ones. For improper decompositions, there would appear coefficients $\big[e^{\lambda_{a,n,N}(b-a)}\xi_n\big]$ blowing up with larger width $b-a$ and $n\to\infty$.

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 lower 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.

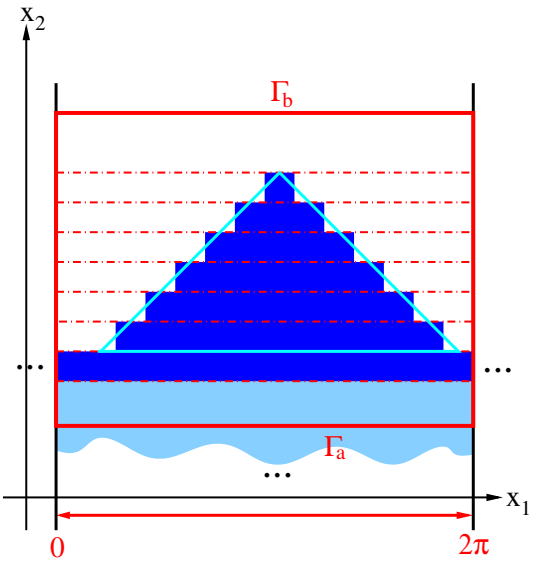

Figure 6: Staircase example for subslices to approximate an x_2-depending wavenumber function.

Altogether, the parameters of RCWA 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 next 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. So the last discretization parameter is the stepsize of such an algorithm.

8 Analysis of convergence

8.1 Results of Civiletti, Lakhtakia, and Monk [4]

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 (6.7) is used. This might be not acceptable for the FMM.
- On the common boundary between consecutive slices, the boundary data for the upgoing plus 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 (6.7) and that of D for the iteration step in (5.11).

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 empirical 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.

8.2 Convergence of the SMA on the continuous level

A first step of the RCWA is to approximate the wavenumber function by an approximate one, for which there is a slicing s.t. the wavenumber function is x_2-independent over each slice. To find such a slicing, the arguments used in [4] maybe helpful.

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 (6.8) and of the iteration (cf. Sect. 5), from Thm. 4.2, and from the Lemmata 5.3 and 6.1, we conclude

Theorem 8.1. *Suppose the slicing is fixed s.t., for $j = 1, \cdots, n$,*

- *The wavenumber function $k(x_1, x_2)$ in the slice $h_{j-1} \le x_2 < h_j$ is independent of x_2 and satisfies $\Re e\, k > 0$ as well as $\Im m\, k \ge 0$.*

- *For the wavenumber function k and with $k=k(\cdot,h_{j-1})$ and $k=k(\cdot,h_j)$, we assume that k is piecewise twice continuously differentiable and that the EVD of the operators k^2L with L defined in (3.2) satisfy the assumptions (3.7)–(3.9).*
- *The S-matrices $S^{h_{j-1}h_j}$ and $S^{h_0h_j}$ are bounded operators.*

Choose any pair of incoming wave functions $(u_a^+,u_b^-)\in H_\alpha^{1/2}(\Gamma_a)\times H_\alpha^{1/2}(\Gamma_b)$. Consider the iterative SMA method (5.12), where the S-matrices are defined by (6.9). Then the resulting pair of outgoing waves $(u_b^+,u_a^-)\in H_\alpha^{1/2}(\Gamma_b)\times H_\alpha^{1/2}(\Gamma_a)$ are the true solutions of the scattering problem (2.1).

Remark 8.2. *For the conditions (3.7)–(3.9), see the two cases discussed in Sect. 3, and, for the existence of bounded S-matrices $S^{h_{j-1}h_j}$ and $S^{h_0h_j}$, see Thm. 4.2. The analogous result for the case of TE polarization holds true (cf. [7, Thm. 6.3]).*

To prepare the proof on the convergence of the discretized RCWA (cf. Thm. 8.11), we recall a well-known result on the approximation of general eigenvalues (cf. e.g. [17, Sect. 4.2]) in Subsect. 8.3. We discuss two assumptions on the EVD of Sect. 3 in Subsect. 8.4. Moreover, we derive two lemmata on the stable convergence of the Dirichlet-to-Neumann maps in Subsect. 8.5. Finally, we present the convergence result in Subsect. 8.6.

8.3 Notation and results on abstract discrete convergence

First we recall some definitions and facts on discrete approximation from [17, Chapts. 1-2,4]. Consider an operator $A\in\mathcal{L}(E,F)$ between the Hilbert spaces E and F. Suppose, for integers $N>0$, there are finite dimensional approximate spaces E_N and F_N connected to the spaces E and F by linear injection operators operators $p_N\colon E\to E_N$ and $q_N\colon F\to F_N$ s.t. $\|p_Ne\|_{E_N}\to\|e\|_E$, $\forall e\in E$ and $\|q_Nf\|_{E_N}\to\|f\|_E$, $\forall f\in F$. For instance, if $\mathcal{P}_N$ is a projection in E strongly converging to the identity, then we can choose $E_N:=\operatorname{im}\mathcal{P}_N$, $\|\mathcal{P}_Ne\|_{E_N}:=\|\mathcal{P}_Ne\|_E$, and $p_Ne:=\mathcal{P}_Ne$. For the general setting of approximate spaces E_N, we say that the sequence $\{e_N\}$ with $e_N\in E_N$ converges discretely to $e\in E$ if $\|e_N-p_Ne\|_{E_N}\to 0$. We write $e_N\to e$.

Any sequence $\{e_N,\ N\in\mathbb{N}\}$ with $e_N\in E_N$ is called compact if, for any infinite subset $\mathbb{N}'\subset\mathbb{N}$, there exists an infinite subset $\mathbb{N}''\subset\mathbb{N}'$ and an $e\in E$ such that the sequence $\{e_N,\ N\in\mathbb{N}''\}$ converges to e. Equivalently, the sequence $\{e_N,\ N\in\mathbb{N}\}$ is compact if, for any $\varepsilon>0$ and any infinite $\mathbb{N}'\subset\mathbb{N}$, there is an infinite subset $\mathbb{N}''\subset\mathbb{N}'$ and an $e\in E$ such that $\|e_N-p_Ne\|\le\varepsilon$ for $N\in\mathbb{N}''$.

For approximate spaces E_N and F_N and an operator $A\in\mathcal{L}(E,F)$, we consider approximate operators $A_N\in\mathcal{L}(E_N,F_N)$. We say that the sequence $\{A_N,\ N\in\mathbb{N}\}$ converges discretely to A if, for any $e_N\to e$, we have $A_Ne_N\to Ae$. In this case, we write $A_N\to A$. Note that $\{A_N,\ N\in\mathbb{N}\}$ converges discretely to A if and only if $\sup_{N\in\mathbb{N}}\|A_N\|<\infty$ and if, for any e' in a dense subset E' of E, there holds $A_Np_Ne'\to Ae'$. Furthermore, we recall that the discrete converge $A_N\to A$ is called stable if, additionally to the discrete convergence, there is an $N_0\in\mathbb{N}$ s.t. $\sup_{N\ge N_0}\|A_N^{-1}\|_{\mathcal{L}(F_N,E_N)}<\infty$. If A is invertible, then $A_N\to A$ is stable if and only if $A_N^{-1}\to A^{-1}$. The convergence $A_N\to A$ is called compact if, for any bounded sequence $\{e_N,\ N\in\mathbb{N}\}$, $e_N\in E_N$, the closure of the sequence $\{A_Ne_N,\ N\in\mathbb{N}\}$, $N\in\mathbb{N}$ is compact.

With all these definition we have ([17, Chapt. 4])

Theorem 8.3. *Suppose we have two operators $A,B\in\mathcal{L}(E,F)$ with corresponding approximate operators $A_N,B_N\in\mathcal{L}(E_N,F_N)$ such that $A+B$ is invertible, that the convergence $A_N\to A$ is stable, and that $B_N\to B$ is compact. Then the convergence $A_N+B_N\to A+B$ is stable.*

Theorem 8.4. *Suppose we have three operators $A, B, C \in \mathcal{L}(E,F)$ with approximate operators $A_N, B_N, C_N \in \mathcal{L}(E_N, F_N)$ and with*

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

ii) There is a complex number $\lambda^{\#} \in \Lambda$ s.t. the operator $[A+B-\lambda^{\#}C] \in \mathcal{L}(E,F)$ is invertible.

iii) The convergence $A_N \to A$ is stable. The convergences $B_N \to B$ and $C_N \to C$ are compact.

Then there holds:

a) For a sequence λ_N of eigenvalues s.t. $[A_N+B_N-\lambda_N C_N]e_N = 0$ with $\lambda_N \to \lambda$, the limit λ is an eigenvalue s.t. $[A+B-\lambda C]e = 0$ with an eigenfunction $e \in E$.

b) For any eigenvalue λ and eigenfunction e s.t. $[A+B-\lambda C]e = 0$, there exist eigenvalues λ_{N_k} and eigenfunctions e_{N_k} s.t. $[A_{N_k}+B_{N_k}-\lambda_{N_k}C_{N_k}]e_{N_k} = 0$ with $\lambda_{N_k} \to \lambda$ and $e_{N_k} \to e$ for $k \to \infty$.

8.4 Assumptions on the EVD

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. Thm. 8.4), it seems natural to require the following property of the EVD algorithm applied to the SMA:

ASSUMPTION ON THE APPROXIMATION OF THE EVD:

For the operator $A := k^2 L \colon H^1_\alpha[0,2\pi] \to k^2 H^{-1}_\alpha[0,2\pi]$ and the approximate operators $A_N := [\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1} \mathcal{P}_N L|_{\operatorname{im}\mathcal{P}_N} \in \mathcal{L}\big(\operatorname{im}\mathcal{P}_N, \operatorname{im}\mathcal{P}_N\big)$ consider the EVDs with $Af_n = \lambda_n f_n,\ n \in \mathbb{N}$ and $A_N f_{n,N} = \lambda_{n,N} f_{n,N},\ n = 1, \cdots, 2N+1.$ (8.1)
Suppose that, for any $\varepsilon > 0$ and $n_1 \in \mathbb{N},$ there is a threshold $N_0 = N_0(\varepsilon, n_1),$ s.t. $|\lambda_n - \lambda_{N,n}| \le \varepsilon$ and $\|f_n - f_{n,N}\|_{H^1_\alpha} \le \varepsilon,\ 1 \le n \le \min\{n_1, 2N+1\}$ for any N with $N_0 \le N$.

Lemma 8.5. *Assume the wavenumber function $x_1 \mapsto k(x_1) := k(x_1, x_2)$ is piecewise twice continuously differentiable with $\Re e\, k > 0$ and $\Im m\, k \ge 0$. Suppose the EVD for the continuous level satisfies the conditions (3.7)–(3.9). Then Assumption (8.1) is fulfilled.*

Proof. The operators $A \colon H^1_\alpha \to k^2 H^{-1}_\alpha$ are approximated by the $A_N \colon H^1_{\alpha,N} \to K^2_N H^{-1}_{\alpha,N}$, where the finite dimensional $H^s_{\alpha,N}$ is the trigonometric function space $[\operatorname{im}\mathcal{P}_N]$ endowed with the norm of $H^s_\alpha[0,2\pi]$ and where $K_N := [\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1/2}$. Note that the value $[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1/2}$ of the reciprocal square root function at the operator $[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]$ can be defined by the Cauchy integral over a curve Γ surrounding the compact L^2 spectrum of operator $[\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]$ contained in $\{\langle \mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N} u_N, u_N\rangle \colon \|u_N\|_{L^2} = 1\}$, i.e. the integration curve Γ can be chosen in the set $\{\zeta \in \mathbb{C} \colon \Im m\, \zeta < 0 \text{ or } \Re e\, \zeta > 0\}$. We define the discrete convergence of functions $u_N \in H^{\pm 1}_{\alpha,N}$ to a function $u \in H^{\pm 1}_\alpha$ writing $u_N \to u$ if and only if $\|u_N - \mathcal{P}_N u\|_{H^{\pm 1}_\alpha} \to 0$. Similarly, we define the discrete convergence $K^2_N H^{-1}_{\alpha,N} \ni [K^2_N u_N] \to [k^2 u] \in k^2 H^{-1}_\alpha$ by the formula $q_N := K^2_N \mathcal{P}_N k^{-2}$ and by

$$\big\|[K^2_N u_N] - q_N[k^2 u]\big\|_{K^2_N H^{-1}_{\alpha,N}} \quad = \quad \|u_N - \mathcal{P}_N u\|_{H^{-1}_{\alpha,N}} \to 0.$$

Then A_N converges discretely to A in the sense that, for any $u_N \to u$, there holds $A_N u_N \to Au$. Indeed: The discrete convergence $K^2_N = [\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1} \to k^2 I$ follows by the definition of the discrete convergence in $K^2_N H^{-1}_{\alpha,N}$, and the discrete convergence $L_N|_{\operatorname{im}\mathcal{P}_N} \to L$ is a simple consequence of the strong convergence $L_N|_{\operatorname{im}\mathcal{P}_N}\mathcal{P}_N \to L$. Altogether, the operator product $A_N = K^2_N L_N$ converges discretely to the product $A = k^2 L$.

Now our Lemma follows from Thm. 8.4 if we can prove that the convergence $A_N \to A$ is stable, i.e. if the norms $\|A_N^{-1}\|_{\mathcal{L}(K_N^2 H_\alpha^{-1}, H_\alpha^1)}$ are uniformly bounded for sufficiently large N. Since a shift of A_N and A by a constant multiple of the identity does not change the nature of the EVD, we only have to show the uniform boundedness of the inverses of the operators $L_{c,N} := K_N\{\mathcal{P}_N[\partial_{x_1} k^{-2}\partial_{x_1} - cI]|_{\operatorname{im}\mathcal{P}_N}\}$ for a fixed positive constant c. However, similarly to the variational-form estimates in (3.6), we get the uniform boundedness of the inverse of $K_N L_{c,N} K_N \colon K_N^{-1} H_\alpha^1 \to K_N H_\alpha^{-1}$ from the assumptions on function k. By definition, the operators $K_N \colon K_N H_\alpha^{-1} \to K_N^2 H_\alpha^{-1}$ and $K_N^{-1} \colon H_\alpha^1 \to K_N H_\alpha^1$ and their inverses are uniformly bounded. Consequently, the inverse operators of $K_N^2 L_{c,N} \colon H_\alpha^1 \to K_N^2 H_\alpha^{-1}$ are uniformly bounded w.r.t. N. □

Similarly to Assumptions (3.7) and (3.9), we need the corresponding condition on the discrete level.

ASSUMPTION ON THE RIESZ PROPERTY OF THE DISCRETIZED EVD:

For operator $A_N := [\mathcal{P}_N k^{-2} I|_{\operatorname{im}\mathcal{P}_N}]^{-1}\mathcal{P}_N L|_{\operatorname{im}\mathcal{P}_N} \in \mathcal{L}\big(\operatorname{im}\mathcal{P}_N\big)$ and the EVD $A_N f_{n,N} = \lambda_{n,N} f_{n,N}, n = 1, \cdots, 2N+1,$ we suppose that all eigenvectors are of rank one and that, for a constant $C > 0$ independent of N, there hold the uniform Riesz estimates (8.2)

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

$$\frac{1}{C}\Big\| \sum_{n=1}^{2N+1} c_{n,N} K_N^{-2} f_{n,N}\Big\|_{H^t_{\alpha,N}}^2 \le \sum_{n=1}^{2N+1} (1+|\lambda_{n,N}|^2)^t |c_{n,N}|^2 \le C\Big\| \sum_{n=1}^{2N+1} c_{n,N} K_N^{-2} f_{n,N}\Big\|_{H^t_{\alpha,N}}^2$$

for all coefficients $c_{n,N} \in \mathbb{C}$ and for the Sobolev indices $0 \le s \le 1$ and $-1 \le t \le 0$.

Lemma 8.6. *Suppose the wavenumber function $x_1 \mapsto k(x_1) := k(x_1, x_2)$ is piecewise twice continuously differentiable and that, for a fixed positive ε, there holds $k \ge \varepsilon$. Then Assumption (8.2) is fulfilled.*

Proof. The arguments of Lemma 3.3 apply with the operator of multiplication by k replaced by the discretized operator $K_N := [\mathcal{P}_N k^{-2}|_{\operatorname{im}\mathcal{P}_N}]^{-1/2}$. This leads to orthogonal basis functions $K_N f_{n,N}$ and to the uniform Riesz estimates. □

Remark 8.7. *For the case of the TE polarization (cf. (7.4)), Lemma 8.6 holds as well.*

8.5 Assumptions on the convergence of the discretized DtN operators

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 $|\lambda_{c,n,N}| > \varepsilon_{EVD}$ s.t. we get the Dirichlet-to-Neumann maps and their inverses by (cf. (7.5))

$$\begin{aligned}
[D_t N_c^+]^{\pm 1}\Big(\sum_{n\in\mathbb{N}} \xi_n f_{c,n}\Big) &:= \sum_{n\in\mathbb{N}} [\lambda_{c,n}]^{\pm 1} \xi_n f_{c,n},\\
[D_t N_{c,N}^+]^{\pm 1}\Big(\sum_{n=1}^{2N+1} \xi_n f_{c,n,N}\Big) &:= \sum_{n=1}^{2N+1} [\lambda_{c,n,N}]^{\pm 1} \xi_n f_{c,n,N}, \quad \xi_n \in \mathbb{C}.
\end{aligned}$$

Note that $D_t N_c^- = -D_t N_c^+$ and $D_t N_{c,N}^- = -D_t N_{c,N}^+$. So, for the convergence analysis, we only have to consider the Dirichlet-to-Neumann operators with plus sign. Due to the Riesz property (3.9)

and Assumption (8.2), we get the uniformly bounded operators $k^{-2}D_tN^+ \in \mathcal{L}(H_\alpha^{1/2}, H_\alpha^{-1/2})$ and $K_N^{-2}D_tN_{c,N}^+ \in \mathcal{L}(H_{\alpha,N}^{1/2}, H_{\alpha,N}^{-1/2})$.

Lemma 8.8. *Suppose the wave number k_c is twice continuously differentiable and, for the EVD with $k = k_c$, Assumptions (8.1) and (8.2) are satisfied. Then the operators $D_tN_{c,N}^+ \in \mathcal{L}(H_{\alpha,N}^{1/2}, K_{c,N}^2 H_{\alpha,N}^{-1/2})$ and their inverse operators $[D_tN_{c,N}^+]^{-1} \in \mathcal{L}(K_{c,N}^2 H_{\alpha,N}^{-1/2}, H_{\alpha,N}^{1/2})$ discretely converge to the operators $D_tN_c^+ \in \mathcal{L}(H_\alpha^{1/2}, k_c^2 H_\alpha^{-1/2})$ and $[D_tN_c^+]^{-1} \in \mathcal{L}(k_c^2 H_\alpha^{-1/2}, H_\alpha^{1/2})$, respectively. These discrete convergences are stable. Finally, the approximate operators $K_{c,N}^{-2}D_tN_{c,N}^+\mathcal{P}_N$ converge strongly to $k_c^{-2}D_tN_c^+ \in \mathcal{L}(H_\alpha^{1/2}, H_\alpha^{-1/2})$ and the $[D_tN_{c,N}^+]^{-1}K_{c,N}^2\mathcal{P}_N$ to $[D_tN_c^+]^{-1}k_c^2 I \in \mathcal{L}(H_\alpha^{-1/2}, H_\alpha^{1/2})$.*

Proof. Since the operators $[D_tN_{c,N}^+]^{\pm 1}$ are uniformly bounded by the Riesz properties and since the $K_{c,N}^{\pm 2}$ are uniformly bounded by the definition of the norm in the discrete spaces, it remains to prove the discrete convergence of the $D_tN_{c,N}^+$ and the strong convergence of the $K_{c,N}^{-2}D_tN_{c,N}^+\mathcal{P}_N$ over a dense subset.

For the $D_tN_{c,N}^+$, we shall prove the convergence on the set of basis functions $\{f_{c,m} : m \in \mathbb{N}\}$. Fix an m and the corresponding $f_{c,m}$. Due to the Riesz property for $0 \le s \le 1$ and $-1 \le t \le 0$, we get

$$\begin{aligned}
\mathcal{P}_N f_{c,m} &= \sum_{n=1}^{2N+1} \xi_{n,N} f_{c,n,N}, \quad K_{c,N}^{-2}\mathcal{P}_N f_{c,m} = \sum_{n=1}^{2N+1} \xi_{n,N} K_{c,N}^{-2} f_{c,n,N}, \\
1 \sim \|f_{c,m}\|_{H_\alpha^s}^2 &\sim \sum_{n=1}^{2N+1} (1 + |\lambda_{c,n,N}|^2)^s |\xi_{n,N}|^2, \\
1 \sim \|f_{c,m}\|_{k_c^{-2}H_\alpha^t}^2 &\sim \sum_{n=1}^{2N+1} (1 + |\lambda_{c,n,N}|^2)^t |\xi_{n,N}|^2, \\
D_tN_{c,N}^+ \left(\sum_{n=1}^{2N+1} \xi_{n,N} f_{c,n,N} \right) &= \sum_{n=1}^{2N+1} \lambda_{c,n,N} \xi_{n,N} f_{c,n,N}.
\end{aligned} \tag{8.3}$$

Consequently, for a fixed $n_0 > m$ independent of N, we get

$$\begin{aligned}
\left\| D_tN_{c,N}^+ \mathcal{P}_N f_{c,m} - \lambda_{c,m} \mathcal{P}_N f_{c,m} \right\|_{K_{c,N}^2 H_{\alpha,N}^{-1/2}}^2 &\sim \left\| \sum_{n=1}^{n_0} (\lambda_{c,n,N} - \lambda_{c,m}) \xi_{n,N} K_{c,N}^{-2} f_{c,n,N} \right\|_{H_{\alpha,N}^{-1/2}}^2 \\
&\quad + \mathcal{O}\Big(\sum_{n=n_0+1}^{2N+1} (1 + |\lambda_{c,n,N}|^2)^{1/2} |\xi_{n,N}|^2 \Big).
\end{aligned} \tag{8.4}$$

Thus, for the discrete convergence at the $f_{c,m}$, we have to show that the right-hand side tends to zero. The convergence of the projection $\mathcal{P}_N$ together with Assumption (8.1) leads to

$$\|\mathcal{P}_N f_{c,m} - f_{c,m,N}\|_{H_{\alpha,N}^1}^2 \sim \sum_{n=1}^{2N+1} (1 + |\lambda_{c,n,N}|^2) |\xi_{n,N} - \delta_{n,m}|^2 \to 0. \tag{8.5}$$

So the first term on the right-hand side of (8.4), the squared Sobolev norm tends to zero since we have $\lambda_{c,n,N} \to \lambda_{c,n}$ and (8.5). The second term on the right-hand side of (8.4) can be estimated as

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

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. Assuming the ordering of the eigenvalues $|\lambda_{c,n,N}|\le|\lambda_{c,n+1,N}|$ and $|\lambda_{c,n}|\le|\lambda_{c,n+1}|$ for $n=1,\cdots$, we 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. 8.4 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 $|\lambda_{c,\tilde n,N_k}|\le|\lambda_{c,n_{N_k},N_k}|\le C<C+1\le\lambda_{c,\tilde n}$.

To see the strong convergence $K_{c,N}^{-2}D_tN_{c,N}^+\mathcal{P}_N\to k_c^{-2}D_tN_c^+$, we shall prove the convergence on the same dense set of functions. We get

$$\begin{aligned}&\left\|K_{c,N}^{-2}D_tN_{c,N}^+\mathcal{P}_Nf_{c,m}-k_c^{-2}D_tN_c^+f_{c,m}\right\|_{H_\alpha^{-1/2}}=\left\|K_{c,N}^{-2}D_tN_{c,N}^+\mathcal{P}_Nf_{c,m}-\lambda_{c,m}k_c^{-2}f_{c,m}\right\|_{H_\alpha^{-1/2}}\\ &\le\ \left\|D_tN_{c,N}^+\mathcal{P}_Nf_{c,m}-\lambda_{c,m}\mathcal{P}_Nf_{c,m}\right\|_{K_{c,N}^2H_\alpha^{-1/2}}+\lambda_{c,m}\left\|K_{c,N}^{-2}\mathcal{P}_Nf_{c,m}-k_c^{-2}f_{c,m}\right\|_{H_\alpha^{-1/2}},\end{aligned}$$

where the zero convergence of the first term has been shown above. For the second term, we conclude

$$\left\|K_{c,N}^{-2}\mathcal{P}_Nf_{c,m}-k_c^{-2}f_{c,m}\right\|_{H_\alpha^{-1/2}}\ \le\ C\left\|\mathcal{P}_Nk_c^{-2}\mathcal{P}_Nf_{c,m}-k_c^{-2}f_{c,m}\right\|_{L^2}.$$

So the strong convergence $\mathcal{P}_N\to I\in\mathcal{L}(L^2)$ and the boundedness of the multiplication operator $k_c^{-2}I\in\mathcal{L}(L^2)$ implies the convergence $\|K_{c,N}^{-2}D_tN_{c,N}^+\mathcal{P}_Nf_{c,m}-k_c^{-2}D_tN_c^+f_{c,m}\|_{H_\alpha^{-1/2}}\to0$. □

Unfortunately, we need more. We need to have a stable convergence of the sum of the two operators $K_{c\pm0,N}^{-2}D_tN_{c\pm0,N}^+$ defined with the different restrictions $k_{c\pm0}(x_1):=k(x_1,c\pm0)$ of the wavenumber function:

ASSUMPTION ON THE STRONG CONVERGENCE OF THE SUM OF DTN'S:

For the restricted wavenumber functions $k_{c\pm0}(x_1):=k(x_1,c\pm0)$, consider the approximate operators $\Sigma_N:=\left\{[\mathcal{P}_Nk_{c+0}^{-2}I|_{\operatorname{im}\mathcal{P}_N}]D_tN_{c+0,N}+[\mathcal{P}_Nk_{c-0}^{-2}I|_{\operatorname{im}\mathcal{P}_N}]D_tN_{c-0,N}\right\}$ with $\Sigma_N\mathcal{P}_N$ converging to $\Sigma:=\left\{k_{c+0}^{-2}D_tN_{c+0}+k_{c-0}^{-2}D_tN_{c-0}\right\}\in\mathcal{L}(H_\alpha^{1/2},H_\alpha^{-1/2})$. (8.7)
Then we suppose that Σ is invertible and that the convergence $\Sigma_N\mathcal{P}_N\to\Sigma$ is stable, i.e. we suppose there is an $N_d>0$ s.t. Σ_N is invertible for $N>N_d$ and the operator norms $\|[\Sigma_N]^{-1}\mathcal{P}_N\|_{\mathcal{L}(H_\alpha^{-1/2},H_\alpha^{1/2})}$, $N>N_d$ are uniformly bounded.

We guess that this is true. On the continuous level, the operators $k_{c\pm0}^{-2}D_tN_{c\pm0}$ are strongly elliptic in the same manner s.t. also the sum Σ is strongly elliptic, and together with a trivial null space for Σ the invertibility of Σ follows. In this spirit, if we could split the operators on the discretization level into strongly elliptic operators plus a compactly converging remainder, then we would obtain stable convergence for the sum. Unfortunately, we could not show this. We can only prove

Lemma 8.9. *Suppose that operator Σ is invertible, that the real-valued wavenumber functions $k_{c\pm0}$ are piecewise twice continuously differentiable, and that there is a positive constant $c_k>0$ s.t. $k_{c\pm0}\ge c_k$. Then Assumption (8.7) is fulfilled.*

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

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

Consequently, the functions $f_{c,n,N}$ form an orthogonal basis in $\operatorname{im}\mathcal{P}_N$ w.r.t. the weighted L^2 scalar product $\langle k_c^{-2}\cdot,\cdot\rangle = \langle [\mathcal{P}_N k_c^{-2} I|_{\operatorname{im}\mathcal{P}_N}]\cdot,\cdot\rangle$. So we arrive at (cf. (8.3))

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

For a fixed $\varepsilon>0$, there are $n_1, N_1 \in \mathbb{N}$ s.t. $\lambda_{c,n,N} \geq \varepsilon$ for $n \leq n_1$ and all $N \geq N_1$. Thus the operator $[\mathcal{P}_N k_c^{-2} I|_{\operatorname{im}\mathcal{P}_N}] D_t N_{c,N}^+$ 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 up over $c = c+0, c-0$ and applying Thm. 8.3, we get the assertion. □

Remark 8.10. *If the operator $D_{c-0,c+0}^- := \mathbf{P}_{c+0}^- \colon H_\alpha^{1/2}(\Gamma_{c-0}) \to H_\alpha^{1/2}(\Gamma_{c+0})$ is invertible (cf. Lemma 6.1), then the operator $[k_{c+0}^{-2} D_t N_{c+0}^+ + k_{c-0}^{-2} D_t N_{c-0}^+]$ is invertible. Indeed, by (5.5) we have*

$$\begin{aligned} D_{c-0,c+0}^- &= \frac{1}{2}\left[I + [D_t N_{c+0}^+]^{-1} \frac{k_{c+0}^2}{k_{c-0}^2} D_t N_{c-0}^+ \right] \\ &= \frac{1}{2} [D_t N_{c+0}^+]^{-1} k_{c+0}^2 \left[k_{c+0}^{-2} D_t N_{c+0}^+ + k_{c-0}^{-2} D_t N_{c-0}^+ \right], \end{aligned}$$

where $[D_t N_{c+0}^+]^{-1} k_{c+0}^2$ is the inverse of the invertible $k_{c+0}^{-2} D_t N_{c+0}^+$ (cf. the Riesz properties (3.9) of our general assumption and (5.1)).

8.6 Convergence of the RCWA for $N \to \infty$

Now we consider the RCWA for wavenumbers k constant w.r.t. x_2 over the slices. Recall that this is the SMA of (5.12) applied with the operators $\mathbf{S}^{h_{j-1}h_j}$ replaced by $\mathbf{S}_N^{h_{j-1}h_j}$, which are computed by (7.8). For the ingredients of (7.8), 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 $f_{h_j,n,N} \in \operatorname{im}\mathcal{P}_N$. We get $\mathbf{PT}_{h_{j-1}h_j,N}^{+-} = 0 = \mathbf{PT}_{h_{j-1}h_j,N}^{-+}$ and the operators $\mathbf{PT}_{h_{j-1}h_j,N}^{++}$ and $\mathbf{PT}_{h_j h_{j-1},N}^{--}$ by their common 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 $f_{h_{j-1},n,N}$, $n = 1, \cdots, 2N+1$ (compare the arguments following (6.3)). We get the Dirichlet-to-Neumann maps by (7.5), the two projections $\mathbf{P}_{h_{j-1}h_j,N}^{\pm+}$ and $\mathbf{P}_{h_{j-1}h_j,N}^{\pm-}$ by (7.7) (cf. the subsequent (8.8)), and the $D_{h_{j-1}h_j,N}^-$ by the subsequent (8.9). Finally (cf. item ii) of (5.12)), the reflected and transmitted waves are obtained by applying $\mathbf{S}_N^{h_0 h_n}$ 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 8.11. *For the SMA (5.12) 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 $h_{j-1} \leq x_2 < h_j$ with $\Re e\, k > 0$ and $\Im m\, k \geq 0$.*
- *ii) For $k = k_{c\pm 0}$ with the x_2-coordinates $c = h_j$, $j = 0, \cdots, n$, we assume that the k are piecewise twice continuously differentiable. Furthermore, the assume that the Assumptions (3.7)–(3.9), (8.1), (8.2), and (8.7) are satisfied with these k (cf. Lemmas 8.5-8.6 and 8.8-8.9).*
- *iii) For $j = 1, \cdots, n$, all the S-matrices $S^{h_{j-1}h_j}$ and $S^{h_0 h_j}$ 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 (5.12) 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.,* $\|u^+_{h_n,N}-u^+_{h_n}\|_{H^{1/2}_\alpha(\Gamma_{h_n})}\to 0$ *and* $\|u^-_{h_0,N}-u^-_{h_0}\|_{H^{1/2}_\alpha(\Gamma_{h_0})}\to 0$.

Proof. The plan of proof is as follows. The strong and stable convergences of $K^{-2}_{c,N}[D_tN^\pm_{c,N}]\mathcal{P}_N$ to $k_c^{-2}[D_tN^\pm_c]$ and of $[D_tN^\pm_{c,N}]^{-1}K^2_{c,N}\mathcal{P}_N$ to $[D_tN^\pm_c]^{-1}k_c^2$ will imply the two strong convergences $\mathbf{P}^{\pm+}_{ab,N}\mathcal{P}_N\to\mathbf{P}^{\pm+}_{ab}$ and $\mathbf{P}^{\pm-}_{ab,N}\mathcal{P}_N\to\mathbf{P}^{\pm-}_{ab}$. Further, we shall show the strong and compact convergence $\mathbf{PT}^+_{ab,N}\mathcal{P}_N\to\mathbf{PT}^+_{ab}$ and the strong and stable convergence $D^-_{ab,N}\mathcal{P}_N\to D^-_{ab}$. Consequently, we shall obtain the strong convergence $\mathbf{S}^{ab}_N\mathcal{P}_{N,N}\to\mathbf{S}^{ab}$ (cf. (6.8) and (7.9)). For the two-step computation (5.11), define the determinant operator $D_N:=I-\mathbf{S}^{bc}_{-+,N}\mathbf{S}^{ab}_{+-,N}$. We shall get the strong and stable convergence $D_N^{-1}\mathcal{P}_N\to D^{-1}$. Hence, we shall obtain the strong convergence $\mathbf{S}^{ac}_N\mathcal{P}_{N,N}\to\mathbf{S}^{ac}$ for the matrices computed by the two-step algorithm. Applying these arguments in the finitely many steps of Algorithm (5.12), we get the strong convergence of the corresponding operators $\mathbb{S}^n_N\mathcal{P}_{N,N}\to\mathbb{S}^n$, and the RCWA is shown to be convergent.

So look at the projections $\mathbf{P}^{\pm+}_{b-0,b,N}$ and $\mathbf{P}^{\pm-}_{b-0,b,N}$ for the case $k(\cdot,b)\not\equiv k(\cdot,a)\equiv k(\cdot,b-0)$. The function splitting $u_N=u^+_{b,N}+u^-_{b,N}=u^+_{a,N}+u^-_{a,N}$ means (cf. (5.2))

$$\begin{aligned}&\left(u^+_{b,N},\left[\mathcal{P}_Nk_b^{-2}I\big|_{\operatorname{im}\mathcal{P}_N}\right]D_tN^+_{b,N}u^+_{b,N}\right)+\left(u^-_{b,N},\left[\mathcal{P}_Nk_b^{-2}I\big|_{\operatorname{im}\mathcal{P}_N}\right]D_tN^-_{b,N}u^-_{b,N}\right)\\ &=\left(u^+_{a,N},\left[\mathcal{P}_Nk_a^{-2}I\big|_{\operatorname{im}\mathcal{P}_N}\right]D_tN^+_{a,N}u^+_{a,N}\right)+\left(u^-_{a,N},\left[\mathcal{P}_Nk_a^{-2}I\big|_{\operatorname{im}\mathcal{P}_N}\right]D_tN^-_{a,N}u^-_{b,N}\right).\end{aligned}$$

Using $D_tN^-_{c,N}=-D_tN^+_{c,N}$, we easily conclude (compare the continuous version (5.5))

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

where, assuming $|\lambda_{c,n,N}|>\varepsilon_{EVD},\ c=a,b$ for simplicity of presentation, we have (8.3). In other words, for the convergence of the projections $\mathbf{P}^{\pm+}_{ab,N}$ to $\mathbf{P}^{\pm+}_{ab,}$, we only need the two strong convergences $K^{-2}_{c,N}[D_tN^\pm_{c,N}]\mathcal{P}_N\to k_c^{-2}[D_tN^\pm_c]$ and $[D_tN^\pm_{c,N}]^{-1}K^2_{c,N}\mathcal{P}_N\to[D_tN^\pm_c]^{-1}k_c^2$. These, however, follow from Lemma 8.8.

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 discretized operator $\mathbf{PT}^{++}_{ab,N}\in\mathcal{L}(\operatorname{im}\mathcal{P}^a_N,\operatorname{im}\mathcal{P}^{b-0}_N)=\mathcal{L}(H^{1/2}_{\alpha,N},H^{1/2}_{\alpha,N})$ and of $\mathbf{PT}^{++}_{ab,N}\mathcal{P}^a_N\in\mathcal{L}(H^{1/2}_\alpha,H^{1/2}_\alpha)$. For a strong convergence, we need the convergence on a dense subset in $H^{1/2}_\alpha$. 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

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

we get $\mathbf{PT}^{++}_{ab,N}=\mathbf{PT}^{++}_{ab,N,n_0}+[\mathbf{PT}^{++}_{ab,N}-\mathbf{PT}^{++}_{ab,N,n_0}]$ and $\mathbf{PT}^{++}_{ab}=\mathbf{PT}^{++}_{ab,n_0}+[\mathbf{PT}^{++}_{ab}-\mathbf{PT}^{++}_{ab,n_0}]$. So, for $\varepsilon>0$, there is an n_0 s.t. $\|[\mathbf{PT}^{++}_{ab,N}-\mathbf{PT}^{++}_{ab,N,n_0}]\mathcal{P}_Nf_{a,m}-[\mathbf{PT}^{++}_{ab}-\mathbf{PT}^{++}_{ab,n_0}]f_{a,m}\|_{H^{1/2}_\alpha}\le\varepsilon$ holds for sufficiently large n_0 (cf. the arguments in (8.6)). By the same arguments, we even get the norm

estimate $\|[\mathbf{PT}^{++}_{ab,N}-\mathbf{PT}^{++}_{ab,N,n_0}]\mathcal{P}_N-[\mathbf{PT}^{++}_{ab}-\mathbf{PT}^{++}_{ab,n_0}]\|_{\mathcal{L}(H^{1/2}_\alpha)}\le\varepsilon$. If the expansion of $f_{a,m}$ w.r.t. the basis $f_{a,n,N},\ n=1,\cdots,2N+1$ is $f_{a,m}=\sum_{n=1}^{2N+1}\xi_{n,N}f_{a,n,N}$, then

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

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

$$\begin{aligned}&\sum_{n=1}^{n_0}(1+|\lambda_{a,n,N}|^2)^{-1/2}\,|\lambda_{a,n,N}-\lambda_{a,m}|^2\,|\xi_{n,N}|^2\\ &\sim\quad\left\|K^{-2}_{a,N}D_tN^+_{a,N,n_0}\mathcal{P}_Nf_{a,m}-k_a^{-2}D_tN^+_{a,n_0}f_{a,m}\right\|^2_{H^{-1/2}_\alpha}\to 0.\end{aligned}$$

For fixed n_0, the last two formulas imply $\mathbf{PT}^{++}_{ab,N,n_0}\mathcal{P}_Nf_{a,m}-\mathbf{PT}^{++}_{ab,n_0}f_{a,m}\to 0$, and the strong convergence $\mathbf{PT}^{++}_{ab,N}\mathcal{P}_N\to\mathbf{PT}^{++}_{ab}$ is proved.

Moreover, using the above splitting, we even get that the convergence $\mathbf{PT}^{++}_{ab,N}\mathcal{P}_N\to\mathbf{PT}^{++}_{ab}$ is compact. Indeed, we take a sequence $x_N\in H^{1/2}_\alpha,\ N\in\mathbb{N}$ uniformly bounded, take any subsequence $x_N,\ N\in\mathbb{N}'\subset\mathbb{N}$, and take any $\varepsilon>0$. Then, for a suitable fixed n_0, we obtain the two estimates $\|\mathbf{PT}^{++}_{ab,N}-\mathbf{PT}^{++}_{ab,N,n_0}\|\le\varepsilon$ and $\|\mathbf{PT}^{++}_{ab}-\mathbf{PT}^{++}_{ab,n_0}\|\le\varepsilon$. Expanding the truncated Fourier series into the eigenfunction basis as $\mathcal{P}_Nx_N=\sum_{n=1}^{2N+1}\xi_{N,n}f_{c,n,N}$, 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 $x:=\sum_{n=1}^{n_0}\xi_nf_{c,n}$, we arrive at $\|\mathbf{PT}^{++}_{ab,N,n_0}\mathcal{P}_Nx_N-\mathbf{PT}^{++}_{ab,n_0}x\|_{H^{1/2}_\alpha}\le C\varepsilon$ for sufficiently large N. In other words, for numbers $N\in\mathbb{N}''$ sufficiently large, we get the estimate $\|\mathbf{PT}^{++}_{ab,N}\mathcal{P}_Nx_N-\mathbf{PT}^{++}_{ab}x\|_{H^{1/2}_\alpha}\le C\varepsilon$, showing that the convergence $\mathbf{PT}^{++}_{ab,N}\mathcal{P}^a_Nx_N\to\mathbf{PT}^{++}_{ab}$ is compact indeed. Similarly, it can be shown that the strong convergence $\mathbf{PT}^{--}_{ba,N}\mathcal{P}^a_Nx_N\to\mathbf{PT}^{--}_{ba}$ 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

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

The strong and stable convergence of the first factor $\frac{1}{2}[D_tN^+_{b,N}]^{-1}[\mathcal{P}_Nk_b^{-2}I\big|_{\mathrm{im}\,\mathcal{P}_N}]^{-1}$ on the last right-hand side follows from Lemma 8.8. The same for the second factor in brackets follows from Assumption (8.7) mentioned in Condition ii) of the current Theorem. So $D^-_{ab,N}\mathcal{P}_N$ converges strongly to D^-_{ab} and this convergence is stable. Unfortunately, with Assumption ii) we rely on the poor result of Lemma 8.9. Nevertheless, putting the strong and stable convergences together (cf. (6.8) and (7.9)), we get the strong convergence $\mathbf{S}^{ab}_N\mathcal{P}_{N,N}\to\mathbf{S}^{ab}$.

For the two-step computation in (5.11), we define $D_N:=I_N-\mathbf{S}^{bc}_{-+,N}\mathbf{S}^{ab}_{+-,N}$. 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 $\mathbf{S}^{bc}_{-+}$ is compact since $\mathbf{PT}^{++}_{ab}$

and $\mathbf{PT}^{--}_{ba}$ are compact (cf. (6.8)), which follows by the Riesz property, by the representation as a diagonal matrix $(e^{-\lambda_{a,n}[b-a]}\delta_{m,n})_{m,n\in\mathbb{N}}$ (recall the end of Sect. 6), and by the decay of the diagonal entries. Based on this fact, above we have shown the compact convergence $\mathbf{PT}^{--}_{cb,N}\mathcal{P}^b_N x_N \to \mathbf{PT}^{--}_{cb}$. Since $\mathbf{S}^{bc}_{-+}$ is equal to $\mathbf{PT}^{--}_{cb}$ multiplied by a bounded operator, we get the compact convergence $\mathbf{S}^{bc}_{-+,N}\mathbf{S}^{ab}_{+-,N}\mathcal{P}_N \to \mathbf{S}^{bc}_{-+}\mathbf{S}^{ab}_{+-}$. Consequently, the first assertion of Thm. 8.3 implies that the convergence $D_N\mathcal{P}_N \to D$ is stable. We finally obtain that the $\mathbf{S}^{ac}_N\mathcal{P}_{N,N}$ converge strongly to $\mathbf{S}^{ac}$ for $N\to\infty$.

Altogether, we have shown the strong convergence $\mathbf{S}^{ac}_N\mathcal{P}_{N,N}\to\mathbf{S}^{ac}$ for the matrices computed by the two-step algorithm if we know the strong convergences $\mathbf{S}^{ab}_N\mathcal{P}_{N,N}\to\mathbf{S}^{ab}$ and $\mathbf{S}^{bc}_N\mathcal{P}_{N,N}\to\mathbf{S}^{bc}$ and if the convergence $\mathbf{PT}^{--}_{cb,N}\mathcal{P}^b_N x_N \to \mathbf{PT}^{--}_{cb}$ is compact. Applying these arguments to the finitely many steps of Algorithm (5.12), we get the strong convergence $\mathbb{S}^n_N\mathcal{P}_{N,N}\to\mathbb{S}^n$. In other words, the RCWA is shown to be convergent. □

Remark 8.12. *Assumption iii) 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.*
If the surface structure of the grating admits a slicing like in i), then 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 (8.2) in Assumption ii) is of technical nature. Condition (8.1) of ii) should hold also for the numerical EVD and is designed to avoid that an inaccurate EVD computation spoils the convergence.

Remark 8.13. *Assumption (8.1) of ii) is technical, and in many cases (suppose the additional assumptions in Lemma 8.5 are not satisfied) it is unclear how to check this practically. So in many cases it cannot be excluded that Assumption (8.1) of ii) is 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 $N\to\infty$. At least, in the special case of TM polarization with piecewise twice continuously differentiable wavenumber functions $k_{h_j}\geq\varepsilon>0$, all technical assumptions of Thm. 8.11 are fulfilled.*

Remark 8.14. *Thm. 8.11, which is shown in the TM case, holds for the TE case as well. The proof is similar but simpler.*

9 Concluding remarks: Open problems, Area of application

Mathematically, there remain many interesting open problems.

- For the discretization with fixed slicing and with wavenumber function independent of x_2 over each slice, 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 case 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 w.r.t. the truncation of the Fourier series? If yes, then discretized norms other than the 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.
- For the discretized SMA with wavenumber function depending on x_2 and with $n\to\infty$ s.t. the width of slicing $\max\{h_j-h_{j-1}:\ j=1,\cdots,n\}$ tends to zero:

a) For a single slice, can the formula for the S-matrix $\mathbf{S}^{h_{j-1}h_j}$ with x_2-dependent k and with width $h_j - h_{j-1}$ 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 numerical methods of ODE systems. Inside the slice the method of the FMM might be explicit. Over the boundary of two slices an implicit step is used.

The area of applications for the general SMA are scattering problems over deep surface structures, i.e., gratings with period per in the size of λ_{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 (5.12) 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 numerical ODE algorithm performs well. For non-smoothness w.r.t. x_2, an adaptive choice of the slicing stepsize and of the stepsize for the numerical ODE algorithm 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, an abstract and general comparison of the methods is difficult. Recall that the RCWA/FMM is an SMA, i.e. a global domain-decomposition algorithm combined with a local basic discretization scheme in each slice, using truncated Fourier expansions combined with an EVD/numer. ODE integration. Similarly, a sophisticated FEM is a global solver like an iterative multigrid method with preconditioner and/or a domain-decomposition algorithm and/or an adaptive scheme with local error estimates. Locally, the FEM is a simple basic discretization scheme, which should be compared to the Galerkin approximation of the RCWA/FMM based on truncated Fourier expansions combined by EVD/numer. ODE integration. 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 efficiently by truncated Fourier series.

The SMA part of the RCWA/FMM should be compared to the global parts of the FEM, i.e. to domain decomposition, preconditioning, and iterative solvers. Looking at its nature, the RCWA should rather be compared to FEM combined with domain decomposition. In this sense, Assumption iii) in Thm. 8.11 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 it 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. Often they provide fast and acceptably good approximate solutions with small n and N.

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 Diskretisierungsmethoden, Teubner, Leipzig, 1976.