# Weierstraß–Institut für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633


# Finite Element Solution of Conical Diffraction Problems


Johannes Elschner[1], Rainer Hinder[2], Gunther Schmidt[1]



submitted: March 15th 2001



[1] Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, D – 10117 Berlin, Germany
email: elschner@wias-berlin.de, schmidt@wias-berlin.de

[2] Forschungsinstitut für Optronik und Mustererkennung,
Gutleuthausstr. 1, D – 76275 Ettlingen, Germany
email: hinder@fom.fgan.de



Preprint No. 649

Berlin 2001


---


*2000 Mathematics Subject Classification.* 78A45, 78M10, 65N30

*Key words and phrases.* Conical diffraction, system of Helmholtz equations, transmission problem, strongly elliptic variational formulation, finite element solution

**Abstract.** This paper is devoted to the numerical study of diffraction by periodic structures of plane waves under oblique incidence. For this situation Maxwell's equations can be reduced to a system of two Helmholtz equations in $\mathbf{R}^2$ coupled via quasiperiodic transmission conditions on the piecewise smooth interfaces between different materials. The numerical analysis is based on a strongly elliptic variational formulation of the differential problem in a bounded periodic cell involving nonlocal boundary operators. We obtain existence and uniqueness results for discrete solutions and provide the corresponding error analysis.

## 1. Introduction

We consider a time–harmonic electromagnetic plane wave incident on a general periodic structure in $\mathbf{R}^3$, which is assumed to be infinitely wide and constant in one spatial direction, say $x_3$. The periodic structure separates two regions with constant dielectric coefficients. Inside the structure, the dielectric coefficient is supposed to be a piecewise constant function. The illuminating wave is given by

$$\mathbf{E}^i = \mathbf{p}\, e^{i\alpha x_1 - i\beta x_2 + i\gamma x_3} e^{-i\omega t}\,, \quad \mathbf{H}^i = \mathbf{s}\, e^{i\alpha x_1 - i\beta x_2 + i\gamma x_3} e^{-i\omega t}\,, \tag{1.1}$$

and will be diffracted by the structure. The far field pattern consists of a finite number of outgoing plane waves propagating in directions which lie on the surface of a cone. Therefore in optics this problems is known as conical diffraction.

Recently the analytic properties of conical diffraction were studied in [4]. It was shown that Maxwell's equations for conical diffraction can be reduced to a system of two–dimensional equations which are closely connected with the classical TE and TM diffraction, where the wave vector of the incident field is orthogonal to the $x_3$–direction. Under certain assumptions on the grating materials that have a reasonable physical interpretation and are satisfied for any relevant practical application the conical diffraction problem admits a strongly elliptic variational formulation. This was used to prove general existence and uniqueness results and to study the asymptotics and regularity near edges of the grating surface.

In the present paper we provide the numerical analysis for the finite element solution of conical diffraction problems. We give a variational formulation suitable for FE discretizations and study their convergence. Due to nonlocal boundary conditions and in general complex–valued material coefficients the resulting discrete system has a nonsymmetric matrix with fully populated blocks. For the technologically important devices where the periodic structure is incorporated into a multilayer stack, we describe how the domain of the FE discretization can be reduced by incorporating the layer system into the nonlocal boundary operators. By using generalized FE discretizations which are especially adapted to preserve the behavior of oscillating solutions, we obtain efficient solution methods for conical diffraction.

For the numerical solution of these problems a few methods have been proposed, extending the known engineering methods based on a system of first order differential equations (RCWA, differential equation method) which are used for TE and TM problems (cf. [8]). In [9] an integral equation method was proposed which solves the transmission problem described in Section 3. To our knowledge, no rigorous results on the convergence of these numerical methods are known.

## 2. Preliminaries

Suppose that the whole space is filled with non-magnetic material with a permittivity function $\varepsilon$, which in Cartesian coordinates $(x_1, x_2, x_3)$ does not depend on $x_3$, is periodic in $x_1$, and homogeneous above and below certain interfaces. In practice, the period $d$ of optical gratings under consideration is comparable with the wavelength $\lambda = 2\pi c/\omega$ of incoming plane optical waves, where $c$ denotes the speed of light. For notational convenience we will change the length scale by a factor of $2\pi/d$, so that the grating becomes $2\pi$–periodic: $\varepsilon(x_1 + 2\pi, x_2) = \varepsilon(x_1, x_2)$. Note that this is equivalent to multiplying the frequency $\omega$ by $d/2\pi$.

The intersection of the upper grating surface with the $(x_1, x_2)$–plane is denoted in the sequel by $\Lambda_0$, the intersection of the lower interface with the $(x_1, x_2)$–plane will be denoted by $\Lambda_1$. We assume that the curves $\Lambda_0$ and $\Lambda_1$ are simple and $2\pi$–periodic and that $\Lambda_0 > \Lambda_1$ pointwise, i.e., if $(x_1, y_0) \in \Lambda_0$, $(x_1, y_1) \in \Lambda_1$ then $y_0 > y_1$. The material in the region $G^+ \subset \mathbf{R}^3$ above the grating surface $\Lambda_0 \times \mathbf{R}$ has the constant dielectric coefficient $\varepsilon = \varepsilon_+$, whereas the medium in $G^-$ below $\Lambda_1 \times \mathbf{R}$ is homogeneous with $\varepsilon = \varepsilon_-$. The medium in the region $G_0$ between $\Lambda_0 \times \mathbf{R}$ and $\Lambda_1 \times \mathbf{R}$ is inhomogeneous with $\varepsilon = \varepsilon^0(x_1, x_2)$, and we assume that the function $\varepsilon_0$ is piecewise constant with jumps at certain interfaces $\Lambda_j$, $j = 2, \ldots, \ell$.

The grating is illuminated by a plane wave of the form (1.1) at oblique incidence. This wave will be diffracted by the grating, and the total fields satisfy the time–harmonic Maxwell equations

$$\nabla \times \mathbf{E} = i\omega\mu\mathbf{H} \quad \text{and} \quad \nabla \times \mathbf{H} = -i\omega\varepsilon\mathbf{E}\,, \tag{2.1}$$

where $\mu$ is the permeability of the free space. Additionally the tangential components of the fields are continuous when crossing an interface $\Lambda \times \mathbf{R}$ between two homogeneous media

$$\mathbf{n} \times (\mathbf{E}^{(1)} - \mathbf{E}^{(2)}) = 0 \quad \text{and} \quad \mathbf{n} \times (\mathbf{H}^{(1)} - \mathbf{H}^{(2)}) = 0 \quad \text{on} \ \ \Lambda \times \mathbf{R}\,, \tag{2.2}$$

where $\mathbf{n}$ is the unit normal to the interface $\Lambda \times \mathbf{R}$. We look for vector fields satisfying (2.1) and (2.2) and possessing locally a finite energy, that is

$$\mathbf{E}\,,\ \mathbf{H}\,,\ \nabla \times \mathbf{E}\,,\ \nabla \times \mathbf{H} \in \left(L^2_{loc}(\mathbf{R}^3)\right)^3 . \tag{2.3}$$

The incident plane wave $(\mathbf{E}^i, \mathbf{H}^i)$ has to satisfy (2.1). Therefore the constant amplitude vector $\mathbf{p}$ must be perpendicular to the wave vector $\mathbf{k} = (\alpha, -\beta, \gamma)$, $\mathbf{p}\cdot\mathbf{k} = 0$, further $\mathbf{k}\cdot\mathbf{k} = \omega^2\mu\varepsilon_+$ and $\mathbf{s} = (\omega\mu)^{-1}\mathbf{k} \times \mathbf{p}$. The wave vector can be expressed in terms of the angles of incidence $\theta, \phi \in (-\pi/2, \pi/2)$ as

$$\mathbf{k} = \omega(\mu\varepsilon_+)^{1/2}(\sin\theta\cos\phi, -\cos\theta\cos\phi, \sin\phi)\,.$$

Since the grating is invariant with respect to any translation parallel to the $x_3$–axis, in view of (1.1) we assume the representation

$$\begin{aligned}\mathbf{E}(x_1, x_2, x_3) &= (E_1, E_2, E_3)(x_1, x_2)\, e^{i\gamma x_3},\\ \mathbf{H}(x_1, x_2, x_3) &= (H_1, H_2, H_3)(x_1, x_2)\, e^{i\gamma x_3},\end{aligned}$$

with $E_i, H_i : \mathbf{R}^2 \to \mathbf{C}$. Then (2.1) leads to the relations

$$\begin{aligned} i\omega\mu\,(H_1, H_2, H_3) &= (\partial_2 E_3 - i\gamma E_2, i\gamma E_1 - \partial_1 E_3, \partial_1 E_2 - \partial_2 E_1)\,,\\ -i\omega\varepsilon\,(E_1, E_2, E_3) &= (\partial_2 H_3 - i\gamma H_2, i\gamma H_1 - \partial_1 H_3, \partial_1 H_2 - \partial_2 H_1)\,. \end{aligned} \tag{2.4}$$

Consequently, in the regions where $\varepsilon$ is constant one obtains

$$\begin{aligned} \omega^2\mu\varepsilon E_3 &= i\gamma(\partial_1 E_1 + \partial_2 E_2) - \Delta E_3 \,, \qquad \partial_1 E_1 + \partial_2 E_2 = -i\gamma E_3 \,, \\ \omega^2\mu\varepsilon H_3 &= i\gamma(\partial_1 H_1 + \partial_2 H_2) - \Delta H_3 \,, \qquad \partial_1 H_1 + \partial_2 H_2 = -i\gamma H_3 \,. \end{aligned} \tag{2.5}$$

For the following we introduce the piecewise constant function

$$k = \sqrt{\omega^2\varepsilon\mu}\,, \tag{2.6}$$

where the branch of the square root is chosen such that $k > 0$ for positive real arguments $\omega^2\varepsilon\mu$ and its branch-cut is $(-\infty, 0)$. Note that $k$ can be expressed by the optical index $\nu = c\,(\mu\varepsilon)^{1/2} = (\varepsilon/\varepsilon_0)^{1/2}$ of the corresponding material, here $\varepsilon_0$ denotes the permitivitty of free space. In the new length scale used in the computations we have

$$k = \frac{d}{\lambda}\nu \,.$$

By (2.5) the functions $E_3$ and $H_3$ satisfy therefore the Helmholtz equations with piecewise constant wave numbers

$$(\Delta + k^2 - \gamma^2)\, E_3 = (\Delta + k^2 - \gamma^2)\, H_3 = 0\,.$$

Denoting $k_\gamma^2 := k^2 - \gamma^2$ one obtains after some algebraic manipulations from (2.4)

$$\begin{aligned} k_\gamma^2 E_1 &= i\,(\omega\mu\partial_2 H_3 + \gamma\partial_1 E_3)\,, \quad k_\gamma^2 H_1 = i\,(-\omega\varepsilon\partial_2 E_3 + \gamma\partial_1 H_3)\,, \\ k_\gamma^2 E_2 &= i\,(-\omega\mu\partial_1 H_3 + \gamma\partial_2 E_3)\,, \; k_\gamma^2 H_2 = i\,(\omega\varepsilon\partial_1 E_3 + \gamma\partial_2 H_3)\,. \end{aligned} \tag{2.7}$$

These relations show that the transverse components $(E_1, E_2)$, $(H_1, H_2)$ can be computed from the third components $E_3, H_3$ of the electric and the magnetic field if $k^2 \neq \gamma^2$. Note that the condition of locally finite energy (2.3) is satisfied only if $E_3$ and $H_3$ are $H^1$–regular.

We shall assume throughout the paper that the material parameters of the grating fulfill the following conditions

$$\begin{gathered} k^2 - \gamma^2 \neq 0\,, \\ k_+ > 0\,,\ \operatorname{Re} k_- > 0\,,\ \operatorname{Im} k_- \geq 0\,, \\ \operatorname{Re} k_0(x_1, x_2) > 0\,,\ \operatorname{Im} k_0(x_1, x_2) \geq 0\,. \end{gathered} \tag{2.8}$$

Since $\mathbf{n} = (n_1, n_2, 0)$ one has at the interfaces

$$\begin{aligned} \mathbf{n} \times \mathbf{E} &= \Big(n_2 E_3, -n_1 E_3, n_1 \frac{\partial_1 H_3 - i\gamma H_1}{i\omega\varepsilon} - n_2 \frac{i\gamma H_2 - \partial_2 H_3}{i\omega\varepsilon}\Big)\, e^{i\gamma x_3}, \\ \mathbf{n} \times \mathbf{H} &= \Big(n_2 H_3, -n_1 H_3, n_1 \frac{i\gamma E_1 - \partial_1 E_3}{i\omega\mu} - n_2 \frac{\partial_2 E_3 - i\gamma E_2}{i\omega\mu}\Big)\, e^{i\gamma x_3}. \end{aligned}$$

Hence the continuity of the tangential components (2.2) leads to the conditions

$$\begin{aligned} &E_3^{(1)} = E_3^{(2)}, \;\; \partial_n E_3^{(1)} - i\gamma(n_1 E_1^{(1)} + n_2 E_2^{(1)}) = \partial_n E_3^{(2)} - i\gamma(n_1 E_1^{(2)} + n_2 E_2^{(2)})\,, \\ &H_3^{(1)} = H_3^{(2)}, \; \frac{\partial_n H_3^{(1)} - i\gamma(n_1 H_1^{(1)} + n_2 H_2^{(1)})}{i\omega\varepsilon^{(1)}} = \frac{\partial_n H_3^{(2)} - i\gamma(n_1 H_1^{(2)} + n_2 H_2^{(2)})}{i\omega\varepsilon^{(2)}}\,, \end{aligned}$$

on $\Lambda$. From (2.7) we see that

$$n_1 E_1 + n_2 E_2 = \frac{i}{k^2 - \gamma^2}(\gamma\partial_n E_3 + \omega\mu\partial_t H_3)\,,$$

$$n_1 H_1 + n_2 H_2 = \frac{i}{k^2 - \gamma^2} (\gamma \partial_n H_3 - \omega \varepsilon \partial_t E_3) \,,$$

where $\partial_t = n_1 \partial_2 - n_2 \partial_1$ denotes the corresponding tangential derivative. Therefore

$$\partial_n E_3 - i\gamma(n_1 E_1 + n_2 E_2) = \frac{k^2 \, \partial_n E_3 + \gamma \omega \mu \, \partial_t H_3}{k^2 - \gamma^2} \,,$$

$$\partial_n H_3 - i\gamma(n_1 H_1 + n_2 H_2) = \frac{k^2 \, \partial_n H_3 - \gamma \omega \varepsilon \, \partial_t E_3}{k^2 - \gamma^2} \,,$$

which yields the transmission conditions at the interfaces $\Lambda$

$$[E_3]_\Lambda = 0, \; \Big[\frac{k^2 \, \partial_n E_3 + \gamma \omega \mu \, \partial_t H_3}{k^2 - \gamma^2}\Big]_\Lambda = 0 \,,$$
$$[H_3]_\Lambda = 0, \; \Big[\frac{k^2 \, \partial_n H_3 - \gamma \omega \varepsilon \, \partial_t E_3}{(k^2 - \gamma^2)\omega\varepsilon}\Big]_\Lambda = 0 \,,$$

where $[\cdot]_\Lambda$ denotes the jump across $\Lambda$.

In order to give the same physical dimension to the unknowns we introduce the field $(B_1, B_2, B_3) = Z\,(H_1, H_2, H_3)$, with the positive constant $Z = \sqrt{\mu/\varepsilon_+}$, and set $\mathbf{q} = Z\mathbf{s}$. Hence the functions $E_3$ and $B_3$ satisfy in $\mathbf{R}^2$ the equations

$$(\Delta + k^2 - \gamma^2)\, E_3 = (\Delta + k^2 - \gamma^2)\, B_3 = 0 \tag{2.9}$$

together with the transmission conditions

$$\begin{aligned} &[E_3]_\Lambda = 0, \; \Big[\frac{k^2 \, \partial_n E_3 + k_+ \gamma \, \partial_t B_3}{k^2 - \gamma^2}\Big]_\Lambda = 0 \,, \\ &[B_3]_\Lambda = 0, \quad \Big[\frac{k_+ \partial_n B_3 - \gamma \, \partial_t E_3}{k^2 - \gamma^2}\Big]_\Lambda = 0 \,. \end{aligned} \tag{2.10}$$

The periodicity of $\varepsilon$, together with the form of the incident wave, motivates to seek for physical solutions $\mathbf{E}$ and $\mathbf{H}$ which are $\alpha$–quasiperiodic in $x_1$, i.e., we look for solutions of (2.9), (2.10) satisfying

$$E_3(x_1 + 2\pi, x_2) = e^{2\pi i \alpha} \, E_3(x_1, x_2) \,, \; B_3(x_1 + 2\pi, x_2) = e^{2\pi i \alpha} \, B_3(x_1, x_2).$$

For $|x_2| \to \infty$ we impose the usual radiation condition. The physics of the problem imposes that the diffracted fields remain bounded and that they should be representable as superpositions of outgoing waves. Since the $\alpha$–quasiperiodic functions $E_3^\pm, B_3^\pm$ are analytic above $\Lambda_0$ resp. below $\Lambda_1$, they must take the form

$$\begin{aligned} &\left.\begin{aligned} E_3 &= p_3 \, e^{i(\alpha x_1 - \beta x_2)} + \sum_{n=-\infty}^{\infty} a_n^+ e^{i(\alpha_n x_1 + \beta_n^+ x_2)} \\ B_3 &= q_3 \, e^{i(\alpha x_1 - \beta x_2)} + \sum_{n=-\infty}^{\infty} c_n^+ e^{i(\alpha_n x_1 + \beta_n^+ x_2)} \end{aligned}\right\} \; x_2 > \max \Lambda_0, \\ &E_3 = \sum_{n=-\infty}^{\infty} a_n^- e^{i(\alpha_n x_1 - \beta_n^- x_2)}, \; B_3 = \sum_{n=-\infty}^{\infty} c_n^- e^{i(\alpha_n x_1 - \beta_n^- x_2)}, \; x_2 < \min \Lambda_1, \end{aligned} \tag{2.11}$$

where

$$\alpha_n = \alpha + n \,, \quad \beta_n^\pm = \beta_n^\pm(\alpha, \gamma) = \sqrt{k_\pm^2 - \gamma^2 - \alpha_n^2} \,,$$

and the square root is defined as in equation (2.6).

The complex scalars $a_n^\pm$, $b_n^\pm$ (the Rayleigh coefficients) are the main characteristics of diffraction gratings. They indicate the efficiency and the phase shift of the propagating modes, i.e. the outgoing plane waves corresponding to $\beta_n^\pm > 0$. The efficiencies represent the proportion of energy radiated in each mode. Defining the "energy" as the flux of the Pointing vector through a normed rectangle parallel to the $(x_1, x_3)$–plane, the ratio of the energies of a reflected or transmitted propagating mode and of the incident wave gives the efficiency of this mode. For gratings used in conical diffraction these efficiencies can be computed from the formulas

$$\begin{aligned} e_n^+ &= \frac{\beta_n^+}{\beta} \frac{|a_n^+|^2 + |c_n^+|^2}{p_3^2 + q_3^2} , \\ e_n^- &= \frac{(k_+^2 - \gamma^2)k_-^2}{(k_-^2 - \gamma^2)k_+^2} \frac{\beta_n^-}{\beta} \frac{|a_n^-|^2 + (k_+/k_-)^2 |c_n^-|^2}{p_3^2 + q_3^2} . \end{aligned}$$

## 3. Variational formulation

The transmission problem (2.9 – 2.10) together with the radiation conditions (2.11) can be transformed to a strongly elliptic variational formulation over a bounded domain. Introduce the functions $u = e^{-i\alpha x_1} E_3$, $v = e^{-i\alpha x_1} B_3$, which are $2\pi$–periodic in $x_1$, and the operators

$$\begin{aligned} &\nabla_\alpha = \nabla + i\,(\alpha, 0)\ ,\ \Delta_\alpha = \nabla_\alpha \cdot \nabla_\alpha = \Delta + 2i\alpha\partial_{x_1} - \alpha^2\ , \\ &\partial_{t,\alpha} = n_1\partial_2 - n_2\partial_1 - i\alpha n_2\ ,\ \partial_{n,\alpha} = n \cdot \nabla_\alpha\ . \end{aligned}$$

Then (2.9 – 2.10) lead to the differential equations

$$(\Delta_\alpha + k^2 - \gamma^2)\, u = (\Delta_\alpha + k^2 - \gamma^2)\, v = 0 \quad \text{in } \Omega \tag{3.1}$$

in regions where $k$ is constant, and the transmission conditions

$$\begin{aligned} &[u]_\Lambda = 0,\ \Big[\frac{k^2\, \partial_{n,\alpha}\, u + k_+\gamma\, \partial_{t,\alpha} v}{k^2 - \gamma^2}\Big]_\Lambda = 0\,, \\ &[v]_\Lambda = 0,\quad \Big[\frac{k_+\partial_{n,\alpha} v - \gamma\, \partial_{t,\alpha} u}{k^2 - \gamma^2}\Big]_\Lambda = 0\,. \end{aligned} \tag{3.2}$$

Due to (2.11) solutions $u, v \in H^1_{loc}$ have to satisfy the radiation conditions

$$\begin{aligned} &\left. \begin{aligned} u &= p_3\, e^{-i\beta x_2} + \sum_{n=-\infty}^{\infty} a_n^+ e^{i(nx_1 + \beta_n^+ x_2)} \\ v &= q_3\, e^{-i\beta x_2} + \sum_{n=-\infty}^{\infty} c_n^+ e^{i(nx_1 + \beta_n^+ x_2)} \end{aligned} \right\}\ x_2 > \max \Lambda_0, \\ &u = \sum_{n=-\infty}^{\infty} a_n^- e^{i(nx_1 - \beta_n^- x_2)},\ v = \sum_{n=-\infty}^{\infty} c_n^- e^{i(nx_1 - \beta_n^- x_2)},\ x_2 < \min \Lambda_1. \end{aligned} \tag{3.3}$$

We introduce two straight lines $\Gamma^\pm = \{(x_1, \pm b),\ x_1 \in [0, 2\pi]\}$, with $b > 0$ such that $b > \Lambda_0$ and $-b < \Lambda_1$, and the bounded periodic cell $\Omega = (0, 2\pi) \times (-b, b)$. Let us denote by $H_p^s(\Omega)$, $s \geq 0$, the restriction to $\Omega$ of all functions in the Sobolev space $H^s_{loc}(\mathbf{R}^2)$ which are $2\pi$–periodic in $x_1$.

Let $\Omega_j$, $j = 1, \cdots, m$, be the subdomains of $\Omega$ in which the function $k$ is constant. Multiplying the equations (3.1) in each subdomain $\Omega_j$ by the constant factors $k^2/(k^2-\gamma^2)$ and $1/(k^2-\gamma^2)$, respectively, we get from Green's formula the equations

$$\begin{aligned}
&\sum_{j=1}^{m}\Big(\int_{\Omega_j}\Big(\frac{k^2}{k^2-\gamma^2}\nabla_\alpha u\,\overline{\nabla_\alpha\varphi}-k^2\,u\,\overline{\varphi}\Big)-\int_{\partial\Omega_j}\frac{k^2}{k^2-\gamma^2}\,\partial_{n,\alpha}u\,\overline{\varphi}\Big)=0\,,\\
&\sum_{j=1}^{m}\Big(\int_{\Omega_j}\Big(\frac{1}{k^2-\gamma^2}\,\nabla_\alpha v\,\overline{\nabla_\alpha\psi}-v\,\overline{\psi}\Big)-\int_{\partial\Omega_j}\frac{1}{k^2-\gamma^2}\,\partial_{n,\alpha}v\,\overline{\psi}\Big)=0\,,
\end{aligned}\tag{3.4}$$

for all functions $\varphi, \psi \in H^1_p(\Omega)$. Here the normal derivative on $\partial\Omega_j$ corresponds to the outer normal with respect to $\Omega_j$.

Using the periodicity in $x_1$ and the transmission conditions (3.2) at the interfaces $\Lambda = \{\partial\Omega_j \cap \partial\Omega_i\}$ the equations (3.4) can be transformed to

$$\begin{aligned}
&\int_{\Omega}\Big(\frac{k^2}{k^2-\gamma^2}\nabla_\alpha u\,\overline{\nabla_\alpha\varphi}-k^2\,u\,\overline{\varphi}\Big)-k_+\gamma\int_{\Lambda}\Big[\frac{1}{k^2-\gamma^2}\Big]_\Lambda\,\partial_{t,\alpha}v\,\overline{\varphi}\\
&\qquad\qquad-\frac{k_+^2}{k_+^2-\gamma^2}\int_{\Gamma^+}\partial_n u\,\overline{\varphi}-\frac{k_-^2}{k_-^2-\gamma^2}\int_{\Gamma^-}\partial_n u\,\overline{\varphi}=0\,,\\
&\int_{\Omega}\Big(\frac{1}{k^2-\gamma^2}\nabla_\alpha\,v\,\overline{\nabla_\alpha\psi}-v\,\overline{\psi}\Big)+\frac{\gamma}{k_+}\int_{\Lambda}\Big[\frac{1}{k^2-\gamma^2}\Big]_\Lambda\,\partial_{t,\alpha}u\,\overline{\psi}\\
&\qquad\qquad-\frac{1}{k_+^2-\gamma^2}\int_{\Gamma^+}\partial_n v\,\overline{\psi}-\frac{1}{k_-^2-\gamma^2}\int_{\Gamma^-}\partial_n v\,\overline{\psi}=0\,.
\end{aligned}\tag{3.5}$$

Note that integrals over the upper and lower boundaries $\Gamma^\pm$ of $\Omega$ contain the usual normal derivatives since the artificial boundaries are parallel to the $x_1$–axis. Let us explain the integrals over the interface $\Lambda$. In view of the identity

$$\int_{\Omega_j}\nabla_\alpha f\,\overline{\nabla_\alpha^\perp g}=-\int_{\partial\Omega_j}\partial_{t,\alpha}f\,\overline{g}\quad\text{with }\nabla_\alpha^\perp:=(\partial_2 f,-\partial_1 f)-i\,(0,\alpha)\tag{3.6}$$

for all $f, g \in H^1_p(\Omega)$, the tangential derivatives $\partial_{t,\alpha}u$ and $\partial_{t,\alpha}v$ are uniquely defined on $\Lambda$. Furthermore, fixing the tangential direction on $\Lambda$, the jump $\big[1/(k^2-\gamma^2)\big]_\Lambda$ stands for the difference "value on the left minus value on the right".

Following [3], [2] (see also section 4) one can show that the normal derivatives on $\Gamma^\pm$ satisfy the nonlocal boundary conditions

$$\begin{aligned}
\partial_n u|_{\Gamma^+}&=-T^+_{\alpha\gamma}(u|_{\Gamma^+})-2i\beta p_3 e^{-i\beta b}\,,\quad \partial_n u|_{\Gamma^-}=-T^-_{\alpha\gamma}(u|_{\Gamma^-})\,,\\
\partial_n v|_{\Gamma^+}&=-T^+_{\alpha\gamma}(v|_{\Gamma^+})-2i\beta q_3 e^{-i\beta b}\,,\quad \partial_n v|_{\Gamma^-}=-T^-_{\alpha\gamma}(v|_{\Gamma^-})\,,
\end{aligned}\tag{3.7}$$

where $T^\pm_{\alpha\gamma}$ are periodic pseudodifferential operators of order 1 acting on $2\pi$-periodic functions on $\mathbf{R}$ by the formula

$$(T^\pm_{\alpha\gamma}f)(x)=-\sum_{n\in\mathbf{Z}}i\beta_n^\pm\hat{f}_n e^{inx}\,,\quad \hat{f}_n=(2\pi)^{-1}\int_0^{2\pi}f(x)e^{-inx}\,dx\,.\tag{3.8}$$

Thus the conical diffraction problem admits the following weak formulation

$$\begin{aligned}
&\int\limits_{\Omega} \Big(\frac{k^2}{k^2-\gamma^2}\nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - k^2\, u\,\overline{\varphi}\Big) + \frac{k_+^2}{k_+^2-\gamma^2}\int\limits_{\Gamma^+} (T_{\alpha\gamma}^+ u)\,\overline{\varphi} + \frac{k_-^2}{k_-^2-\gamma^2}\int\limits_{\Gamma^-} (T_{\alpha\gamma}^- u)\,\overline{\varphi} \\
&\qquad + k_+\gamma \int\limits_{\Lambda} \Big[\frac{1}{k^2-\gamma^2}\Big]_\Lambda v\,\overline{\partial_{\tau,\alpha}\varphi} = -\frac{2ip_3\beta k_+^2}{k_+^2-\gamma^2}\, e^{-i\beta b} \int\limits_{\Gamma^+} \overline{\varphi} \\
&\int\limits_{\Omega} \Big(\frac{1}{k^2-\gamma^2}\nabla_\alpha v \cdot \overline{\nabla_\alpha \psi} - v\,\overline{\psi}\Big) + \frac{1}{k_+^2-\gamma^2}\int\limits_{\Gamma^+} (T_{\alpha\gamma}^+ v)\,\overline{\psi} + \frac{1}{k_-^2-\gamma^2}\int\limits_{\Gamma^-} (T_{\alpha\gamma}^- v)\,\overline{\psi} \\
&\qquad - \frac{\gamma}{k_+}\int\limits_{\Lambda} \Big[\frac{1}{k^2-\gamma^2}\Big]_\Lambda u\,\overline{\partial_{\tau,\alpha}\psi} = -\frac{2iq_3\beta}{k_+^2-\gamma^2}\, e^{-i\beta b} \int\limits_{\Gamma^+} \overline{\psi}
\end{aligned} \tag{3.9}$$

for all $\varphi, \psi \in H_p^1(\Omega)$. The left–hand side of the system (3.9) generates a sesquilinear form $\mathcal{B}((u,v),(\varphi,\psi))$, which is bounded on $(H_p^1(\Omega))^2 \times (H_p^1(\Omega))^2$. In the following we describe the main analytic results obtained in [4] for equations of the form

$$\mathcal{B}((u,v),(\varphi,\psi)) = (f,\varphi)_{L_2} + (g,\psi)_{L_2} \tag{3.10}$$

with $(f,g) \in ((H_p^1(\Omega))^2)'$.

**Theorem 1.** Suppose that $k$ satisfies (2.8).

1. If $\operatorname{Im} k > 0$ in some subdomain $\Omega_j \subset \Omega$, then for any $\omega > 0$ equation (3.10) has at most one solution $(u,v) \in (H_p^1(\Omega))^2$.
2. If $k \notin [0,\gamma]$, then $\mathcal{B}$ is strongly elliptic, i.e. there exists a complex number $\sigma$ and a compact form $q$ such that
$$\operatorname{Re} \sigma\mathcal{B}((u,v),(u,v)) \geq c\|(u,v)\|^2_{(H_p^1(\Omega))^2} - q((u,v),(u,v)) .$$
3. Let $k \notin [0,\gamma]$, $k_-^2 \notin [0,\alpha^2+\gamma^2]$ and fix $\Phi \in (0,\pi/2)$. Then there exists $\omega_0 > 0$ such that, for any incidence angles $\theta, \phi$ with $|\theta|, |\phi| \leq \Phi$ and any frequency $\omega$ with $0 < \omega \leq \omega_0$, (3.10) is uniquely solvable.
4. If $k^2 > \gamma^2$ and $k_-^2 > \alpha^2+\gamma^2$, then for all but a countable set of frequencies $\omega_j$, $\omega_j \to \infty$, there exists a unique solution in $(H_p^1(\Omega))^2$.
5. Denote by $\mathcal{R}$ the set of Rayleigh frequencies
$$\mathcal{R} = \Big\{(\omega,\theta,\phi) : \exists\, n \in \mathbf{Z} \text{ s. th. } k_\pm^2 - \gamma^2 = (n+\alpha)^2\Big\} .$$
If for $(\omega^0,\theta^0,\phi^0) \notin \mathcal{R}$ the equation (3.10) is uniquely solvable, then in a neighborhood of this point the unique solution depends analytically on $\omega, \theta, \phi$.

*Remark.* By Snell's law the condition $k^2 > \alpha^2+\gamma^2$ is necessary that the incident wave will be transmitted into the grating material. Hence, the assumptions have a reasonable physical interpretation. They are satisfied for any relevant application.

## 4. Finite element solution

Since the sesquilinear form $\mathcal{B}$ is strongly elliptic, it is natural to use a Galerkin method for solving the direct diffraction problem. By standard theory one can easily show that FE discretizations of the weak formulations (3.9) are stable and provide quasioptimal convergence orders.

In the following we describe some aspects connected with the nonlocal boundary terms and its discretization. In modern diffractive optics the grating structure is very often incorporated into a stack of thin–film layers in order to combine and enhance the properties of these two types of optical devices. In this case one can introduce new nonlocal boundary operators which model the multilayer stack and the radiation conditions. Therefore one can reduce the integration domain $\Omega$ used for the FE discretization to the inhomogeneous grating structure.

Before describing the construction of these boundary operators we first show that the new variational form remains strongly elliptic. To do so we use integral representations for $2\pi$–periodic solutions of the modified Helmholtz equations with constant $k$

$$(\Delta_\alpha + k^2 - \gamma^2)\,\varphi = 0 \quad \text{in some domain } G \tag{4.1}$$

with the property that $(x_1, x_2) \in G$ implies $(x_1 + 2\pi, x_2) \in G$ and $|x_2| < b$. First we briefly recall some basic properties. The fundamental solution is given by the infinite series

$$\Psi(x) = \frac{i}{2\pi} \sum_{n\in\mathbf{Z}} \frac{e^{inx_1 + i\beta_n |x_2|}}{\beta_n} \tag{4.2}$$

with $\beta_n = \sqrt{k^2 - \gamma^2 - \alpha_n^2}$. If one of the denominators $\beta_n$ in (4.2) is zero, then the corresponding term in the series must be replaced by $ie^{inx_1}(C + |x_2|)$, where $C$ is an arbitrary constant. By standard potential theory the periodic function $\varphi$ solves (4.1) if and only if it admits the representation

$$\varphi(x) = \frac{1}{2}\Big(V\partial_\nu \varphi(x) - K\varphi(x)\Big)\,, \quad x \in G\,. \tag{4.3}$$

Here $V$ and $K$ are the single and double layer potentials

$$V\varphi(x) = \int\limits_{\partial G} \Psi(x-y)\,\varphi(y)\,ds\,,\ K\varphi(x) = \int\limits_{\partial G} \partial_{\nu_y}\Psi(x-y)\,\varphi(y)\,ds\,,$$

and $\nu$ denotes the exterior normal of $\partial G$. Taking in (4.3) the normal derivatives at $\partial G$ one derives the equation

$$2\partial_\nu\varphi|_{\partial G} = K'(\partial_\nu\varphi|_{\partial G}) + \partial_\nu\varphi|_{\partial G} + D(\varphi|_{\partial G}) \tag{4.4}$$

with the integral operators

$$K'\varphi(x) = \partial_\nu \int\limits_{\partial G} \Psi(x-y)\,\varphi(y)\,ds\,,\ D\varphi(x) = -\partial_\nu \int\limits_{\partial G} \partial_{\nu_y}\Psi(x-y)\,\varphi(y)\,ds\,.$$

Let us consider the special case that $G$ represents a layer parallel to the $(x_1, x_3)$–plane. Then $\partial G$ consists of two lines, say $x_2 = t_1$ and $x_2 = t_2$, $t_1 < t_2$, and therefore the boundary integral operators have a very simple form. In the following we set $K'_{ij}\varphi = K'(\varphi|_{t_j})(x, t_i)$, $D_{ij}\varphi = D(\varphi|_{t_j})(x, t_i)$, $i, j = 1, 2$. These operators are diagonal in the Fourier basis with $K'_{ii} = 0$ and

$$\begin{aligned}
D_{ii}\varphi(x_1, t_i) &= -i \sum_{n\in\mathbf{Z}} \beta_n \hat\varphi_n(t_i)\, e^{inx_1}\,,\\
D_{ij}\varphi(x_1, t_i) &= i \sum_{n\in\mathbf{Z}} \beta_n \hat\varphi_n(t_j)\, e^{i\beta_n d}\, e^{inx_1}\,,\\
K'_{ij}\varphi(x_1, t_i) &= -\sum_{n\in\mathbf{Z}} \hat\varphi_n(t_j)\, e^{i\beta_n d}\, e^{inx_1}\,,
\end{aligned}$$

where $d = t_2 - t_1$ denotes the layer thickness. Therefore the boundary integral equations (4.4) for solutions $u, v$ can be transformed to

$$\begin{aligned} (I - K'_{12}K'_{21})\partial_\nu u|_{t_1} &= (I + K'_{21}K'_{12})D_{11}u|_{t_1} + 2D_{12}u|_{t_2} \\ (I - K'_{21}K'_{12})\partial_\nu u|_{t_2} &= 2D_{21}u|_{t_1} + (I + K'_{21}K'_{12})D_{22}u|_{t_2} \end{aligned} \tag{4.5}$$

If we suppose that at the line $\{x_2 = t_2\}$ the normal derivatives are given by the equation

$$\begin{pmatrix} \partial_n u|_{t_2} \\ \partial_n v|_{t_2} \end{pmatrix} = \begin{pmatrix} a_{uu} \, a_{uv} \\ a_{vu} \, a_{vv} \end{pmatrix} \begin{pmatrix} u|_{t_2} \\ v|_{t_2} \end{pmatrix} + \begin{pmatrix} f_2 \\ g_2 \end{pmatrix} = A \begin{pmatrix} u|_{t_2} \\ v|_{t_2} \end{pmatrix} + \begin{pmatrix} f_2 \\ g_2 \end{pmatrix} \tag{4.6}$$

where the elements of $A$ are certain pseudodifferential operators of finite order, then we obtain from (4.5) a linear system for $u|_{t_2}$ and $v|_{t_2}$ with the operator matrix

$$\begin{pmatrix} a_{uu} - D_{22} - (a_{uu} + D_{22})K'_{21}K'_{12} & (I - K'_{21}K'_{12})a_{uv} \\ (I - K'_{21}K'_{12})a_{vu} & a_{vv} - D_{22} - (a_{vv} + D_{22})K'_{21}K'_{12} \end{pmatrix} \tag{4.7}$$

In the following we assume that this matrix is invertible. Putting the solution of this system into the first equations of (4.5) we therefore obtain

$$\begin{aligned} \partial_\nu u|_{t_1} &= D_{11}u|_{t_1} + (K'_{12}K'_{21}(I + K'_{21}K'_{12})D_{11} + 2D_{12}b_{uu})u|_{t_1} + 2D_{12}b_{uv}v|_{t_1} + f_1 \\ \partial_\nu v|_{t_1} &= D_{11}v|_{t_1} + 2D_{12}b_{vu}u|_{t_1} + (K'_{12}K'_{21}(I + K'_{21}K'_{12})D_{11} + 2D_{12}b_{vv})v|_{t_1} + g_1 \end{aligned}$$

with certain functions $f_1$ and $g_1$. Thus we have proved

**Lemma 2.** If for $x_2 = t_2$ the normal derivatives and boundary values of solutions $u$ and $v$ to (4.1) are connected by (4.6) and the operator (4.7) is invertible, then

$$\begin{pmatrix} \partial_n u|_{t_1} \\ \partial_n v|_{t_1} \end{pmatrix} = B \begin{pmatrix} u|_{t_1} \\ v|_{t_1} \end{pmatrix} + \begin{pmatrix} f_1 \\ g_1 \end{pmatrix} \tag{4.8}$$

where the operator

$$B - \begin{pmatrix} D_{11} & 0 \\ 0 & D_{11} \end{pmatrix}$$

maps $H_p^{1/2} \times H_p^{1/2}$ into the set of functions with exponentially decaying Fourier coefficients.

Now we apply this result to a stack of $N$ layers $L_j = \{x_2 \in (t_{j-1}, t_j)\}$, $1 \le j \le N$, with Helmholtz coefficients $k_j$, located above the grating structure. For $x_2 = b > t_N$ we have the boundary conditions

$$\begin{aligned} \partial_n u|_{\Gamma^+} &= -T^+_{\alpha\gamma}(u|_{\Gamma^+}) - 2i\beta p_3 e^{-i\beta b} \, , \\ \partial_n v|_{\Gamma^+} &= -T^+_{\alpha\gamma}(v|_{\Gamma^+}) - 2i\beta q_3 e^{-i\beta b} \, , \end{aligned} \tag{4.9}$$

Note that $T^+_{\alpha\gamma}\varphi = D(\varphi|_{\Gamma^+})$ for $k = k_+$. The boundary conditions (3.7) obviously follow from (4.4) applied in the outer domains $\{|x_2| > b\}$ to the sum of the incoming and diffracted fields. The change of signs is caused by the fact that $\nu$ and the outer normal to $\Gamma^+$ with respect to $\Omega$ have opposite directions.

The transmission conditions (3.2) for $u$ and $v$ at the flat interfaces $x_2 = t_j$, $1 \le j \le N$, generate an equation of the form (4.6). Thus if we choose a new artificial boundary $\widetilde{\Gamma}^+ = \{(x_1, \tilde{b}),\ x_1 \in [0, 2\pi],\ \tilde{b} \in (t_0, t_1)\}$ within the layer $L_1$ we obtain nonlocal boundary conditions

$$\begin{pmatrix} \partial_n u|_{\widetilde{\Gamma}^+} \\ \partial_n v|_{\widetilde{\Gamma}^+} \end{pmatrix} = M^+_{\alpha\gamma} \begin{pmatrix} u|_{\widetilde{\Gamma}^+} \\ v|_{\widetilde{\Gamma}^+} \end{pmatrix} + \delta_{0n} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \tag{4.10}$$

which have to be satisfied by the solutions $u, v$ of the conical diffraction problem (3.1 – 3.3). Here $\delta_{0n}$ is the Kronecker symbol. Repeating an analogous procedure with a layer-stack below the grating structure, we obtain a new artificial boundary $\widetilde{\Gamma}^-$ within the first layer below the grating with Helmholtz coefficient $k_{-1}$ and an analogous nonlocal boundary condition with an operator $M^-_{\alpha\gamma}$. From Lemma 2 we obtain additionally

**Lemma 3.** The matrix operators $M^\pm_{\alpha\gamma}$ are compact perturbations of the diagonal matrix operators

$$\widetilde{T}^\pm_{\alpha\gamma} := \begin{pmatrix} T^{(\pm 1)}_{\alpha\gamma} & 0 \\ 0 & T^{(\pm 1)}_{\alpha\gamma} \end{pmatrix} : \begin{array}{c} H^{1/2}_p(\widetilde{\Gamma}^\pm) \\ \times \\ H^{1/2}_p(\widetilde{\Gamma}^\pm) \end{array} \longrightarrow \begin{array}{c} H^{-1/2}_p(\widetilde{\Gamma}^\pm) \\ \times \\ H^{-1/2}_p(\widetilde{\Gamma}^\pm) \end{array},$$

where the pseudodifferential operators $T^{(\pm 1)}_{\alpha\gamma}$ are defined as in (3.8) with $k_\pm$ replaced by $k_{\pm 1}$.

Thus we have derived a reduced periodic cell $\widetilde{\Omega}$ with upper resp. lower boundaries $\widetilde{\Gamma}^\pm$ which becomes the new integration domain. Let $\widetilde{\Lambda}$ denote the set of all interfaces between different materials inside $\widetilde{\Omega}$. Taking into account that $\nu$ has the opposite direction of the outer normal to $\widetilde{\Gamma}^+$ (w.r.t. $\widetilde{\Omega}$) we see that the problem (3.1 – 3.3) is transformed to the variational system

$$\begin{aligned}
&\int\limits_{\widetilde{\Omega}} \Big( \frac{k^2}{k^2-\gamma^2} \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - k^2\, u \overline{\varphi} \Big) + k_+ \gamma \int\limits_{\widetilde{\Lambda}} \Big[ \frac{1}{k^2-\gamma^2} \Big]_{\widetilde{\Lambda}} v\, \overline{\partial_{\tau,\alpha}\varphi} \\
&+ \frac{k_1^2}{k_1^2-\gamma^2} \int\limits_{\widetilde{\Gamma}^+} \big( (M^+_{\alpha\gamma})_{11} u + (M^+_{\alpha\gamma})_{12} v \big) \overline{\varphi} + \frac{k_{-1}^2}{k_{-1}^2-\gamma^2} \int\limits_{\widetilde{\Gamma}^-} \big( (M^-_{\alpha\gamma})_{11} u + (M^-_{\alpha\gamma})_{12} v \big) \overline{\varphi} \\
&\qquad\qquad = \frac{f_1 k_1^2}{k_1^2-\gamma^2} \int\limits_{\widetilde{\Gamma}^+} \overline{\varphi} \\
&\int\limits_{\widetilde{\Omega}} \Big( \frac{1}{k^2-\gamma^2} \nabla_\alpha v \cdot \overline{\nabla_\alpha \psi} - v\, \overline{\psi} \Big) - \frac{\gamma}{k_+} \int\limits_{\widetilde{\Lambda}} \Big[ \frac{1}{k^2-\gamma^2} \Big]_{\widetilde{\Lambda}} u\, \overline{\partial_{\tau,\alpha}\psi} \\
&+ \frac{1}{k_1^2-\gamma^2} \int\limits_{\widetilde{\Gamma}^+} \big( (M^+_{\alpha\gamma})_{21} u + (M^+_{\alpha\gamma})_{22} v \big) \overline{\psi} + \frac{1}{k_{-1}^2-\gamma^2} \int\limits_{\widetilde{\Gamma}^-} \big( (M^-_{\alpha\gamma})_{21} u + (M^-_{\alpha\gamma})_{22} v \big) \overline{\psi} \\
&\qquad\qquad = -\frac{f_2}{k_1^2-\gamma^2} \int\limits_{\widetilde{\Gamma}^+} \overline{\psi}
\end{aligned} \tag{4.11}$$

for all $\varphi, \psi \in H^1_p(\widetilde{\Omega})$. Here $k_{\pm 1}$ are the Helmholtz coefficients of the layers directly above resp. below the inhomogeneous structure and $(M^\pm_{\alpha\gamma})_{ij}$ denote the entries of the matrix operators $M^\pm_{\alpha\gamma}$. We denote this sesquilinear form by $\mathcal{B}\big((u,v),(\varphi,\psi)\big)$. For the numerical solution we choose some finite–dimensional subspace $S_h$ of $H_p^1(\Omega)$ and seek a solution $(u_h, v_h) \in (S_h)^2$ of the discrete equations

$$\widetilde{\mathcal{B}}((u_h, v_h), (\varphi_h, \psi_h)) = \frac{f_1 k_1^2}{k_1^2 - \gamma^2} \int\limits_{\widetilde{\Gamma}^+} \overline{\varphi_h} + \frac{f_2}{k_1^2 - \gamma^2} \int\limits_{\widetilde{\Gamma}^+} \overline{\psi_h} \tag{4.12}$$

for all $\varphi_h, \psi_h \in S_h$. Since the domain integrals and the line integrals over the interfaces in the variational form $\widetilde{\mathcal{B}}$ can be exactly computed for piecewise polynomial functions, a consistency error may occur only if the nonlocal boundary operators

$$\int\limits_{\widetilde{\Gamma}^\pm} (M_{\alpha\gamma}^\pm)_{ij} u \overline{\varphi} \, dx = \frac{1}{2\pi} \sum_{n \in \mathbf{Z}} (m_{ij}^\pm)_n \hat{u}_n \overline{\hat{\varphi}}_n \tag{4.13}$$

are discretized. But for finite elements whose traces on $\widetilde{\Gamma}^\pm$ are periodic splines with uniformly distributed break points this can be done very efficiently and accurately, so that this discretization error can be neglected. We will show this in the next section.

**Theorem 4.** Suppose that $k$ satisfies (2.8) and $k \notin [0, \sqrt{\alpha^2 + \gamma^2}]$. Then for all but a sequence of countable frequencies $\omega_j$, $|\omega_j| \to \infty$, and all sufficiently small $h$, the Galerkin equations (4.12) are uniquely solvable. If the exact solution satisfies $(u, v) \in (H_p^s(\Omega))^2$, $1 < s \le 2$, then the difference between the finite element solutions and the exact solution can be estimated by

$$\|u - u_h\|_{H^1(\widetilde{\Omega})} + \|v - v_h\|_{H^1(\widetilde{\Omega})}) \le C h^{s-1} (\|u\|_{H^s(\widetilde{\Omega})} + \|v\|_{H^s(\widetilde{\Omega})}) \,,$$
$$(\|u - u_h\|_{L^2(\widetilde{\Omega})} + \|v - v_h\|_{L^2(\widetilde{\Omega})}) \le C h^{2s-2} (\|u\|_{H^s(\widetilde{\Omega})} + \|v\|_{H^s(\widetilde{\Omega})}) \,,$$

where the constants depend on $k$ but are independent of $h$ and $u$.

*Proof.* Consider the same inhomogeneous grating structure but without multilayer stacks and with material parameters $k_1$ above the grating and $k_{-1}$ for the substrate. Then Theorem 1 can be applied. Note that the variational form corresponding to this problems differs by Lemma 3 from the sesquilinear form $\widetilde{\mathcal{B}}$ by a compact perturbation. Hence all conclusions of Theorem 1 are also valid for $\widetilde{\mathcal{B}}$, and therefore by standard theory the assertion follows. $\square$

## 5. Implementation

Here we discuss the efficient and accurate computation of the nonlocal boundary terms which model a system of say $N$ horizontal layers $\{x_2 \in (t_{j-1}, t_j)\}$, $1 \le j \le N$, with Helmholtz coefficients $k_j$. The operators $M_{\alpha\gamma}^\pm$ are diagonal in the Fourier basis. Therefore one should use the Fourier series of the traces of the finite elements on $\widetilde{\Gamma}^\pm$. Moreover, the solutions of the equations (3.1) are analytic and $2\pi$–periodic in $x_1$, so that the application of Fourier series techniques in the layer system is justified. For $t_0 < x_2 < t_N$ the solution can be written in the form

$$u(x_1, x_2) = \sum_{n \in \mathbf{Z}} \phi_n(x_2) \, e^{inx_1} \,, \quad v(x_1, x_2) = \sum_{n \in \mathbf{Z}} \psi_n(x_2) \, e^{inx_1} \,, \tag{5.1}$$

where the Fourier coefficients $\phi_n$ and $\psi_n$ solve the differential equation

$$\Big( \frac{d^2}{dx_2^2} + (k_j^2 - \gamma^2 - (n+\alpha)^2) \Big) \phi_n = 0 \,, \quad t_{j-1} < x_2 < t_j \,, \; 1 \le j \le N. \tag{5.2}$$

Therefore

$$\phi_n(x_2) = a_1\rho_j^+(x_2) + a_2\rho_j^-(x_2)\,, \quad \psi_n(x_2) = a_3\rho_j^+(x_2) + a_4\rho_j^-(x_2)$$

with $\rho_j^\pm = \exp(\pm i\sqrt{k_j^2 - \gamma^2 - (n+\alpha)^2}\,x_2)$ or constant and linear functions in the case $k_j^2 = \gamma^2 + (n+\alpha)^2$.

The transmission conditions (3.2) for $u$ and $v$ at the flat interfaces $x_2 = t_j$, $1 \le j \le N-1$, transform to conditions for the solutions of (5.2) at the inner points of the form

$$\begin{gathered} \phi_n(t_j^-) = \phi_n(t_j^+)\,,\ \psi_n(t_j^-) = \psi_n(t_j^+)\,, \\ \frac{k_j^2\phi_n' - i\gamma k^+(n+\alpha)\psi_n}{k_j^2-\gamma^2}\Big|_{t_j^-} = \frac{k_{j+1}^2\phi_n' - i\gamma k^+(n+\alpha)\psi_n}{k_{j+1}^2-\gamma^2}\Big|_{t_j^+}\,, \\ \frac{i\gamma\,(n+\alpha)\phi_n + k^+\psi_n'}{k_j^2-\gamma^2}\Big|_{t_j^-} = \frac{i\gamma\,(n+\alpha)\phi_n + k^+\psi_n'}{k_{j+1}^2-\gamma^2}\Big|_{t_j^+}\,, \end{gathered} \tag{5.3}$$

where $t_j^\pm$ means that the interface $x_2 = t_j$ is approached from above resp. below. It can be easily checked that (5.3) generates an one–to–one mapping between the coefficients $a_i$, $i = 1,\dots,4$, of the solutions corresponding to adjacent subintervals.

Let the multi–layer stack be located above the grating structure. Then for $x_2 = t_N$ we have (5.3) with $k_{N+1}$ replaced by $k_+$. Moreover, since for $x_2 > t_N$ the radiation condition has to be satisfied we obtain from (3.7)

$$\begin{aligned} \phi_n'(b) &= i\beta_n^+\,\phi_n(b) - 2i\beta p_3\delta_{0n}e^{-i\beta b}\,, \\ \psi_n'(b) &= i\beta_n^+\,\psi_n^+(b) - 2i\beta q_3\delta_{0n}e^{-i\beta b}\,. \end{aligned}$$

Hence the system of differential equations (5.2) with the interior jump conditions (5.3) has a solution depending on two parameters, which for $x_2 \in (t_N, b)$ is given by

$$\begin{aligned} \phi_n(x_2) &= c_1e^{i\beta_n^+x_2} + p_3\delta_{0n}e^{-i\beta x_2}\,, \\ \psi_n(x_2) &= c_2e^{i\beta_n^+x_2} + q_3\delta_{0n}e^{-i\beta x_2}\,. \end{aligned}$$

Now we choose the left boundary $\tilde b \in (t_0, t_1)$ and boundary conditions

$$\begin{aligned} &\phi_n'(b) = i\beta_n^+\,\phi_n(b)\,,\ \psi_n'(b) = i\beta_n^+\,\psi_n(b)\,, \\ &\phi_n(\tilde b) = \psi_n(\tilde b) = 0\,. \end{aligned} \tag{5.4}$$

Suppose that the problem (5.2 – 5.4) has only the trivial solution for any $n \in \mathbf{Z}$. Then there exists a matrix $M_n^+$ such that at this left boundary

$$\begin{pmatrix} \phi'(\tilde b) \\ \psi'(\tilde b) \end{pmatrix} = M_n^+ \begin{pmatrix} \phi(\tilde b) \\ \psi(\tilde b) \end{pmatrix} + \delta_{0n}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

It is clear that the matrices $\{M_n^+,\, n \in \mathbf{Z}\}$ form the operator $M_{\alpha\gamma}^+$. Note that the unique solvability of the boundary value problem (5.2 – 5.4) corresponds to the invertibility of the operator (4.7) mentioned in Lemma 2.

Hence, the scalars $(m_{ij}^\pm)_n$ in (4.13) can be found from solving the differential equations (5.2). But due to Lemma 2 the numbers $(m_{ij}^\pm)_n$ and the differences $(m_{ii}^\pm)_n + i\beta_n^{\pm 1}$ decay exponentially as $|n| \to \infty$. Therefore these equations have to be solved only for a relatively small number of different $n$. This can be done efficiently by a recursive algorithm described in [7], which is numerically stable for any number of layers, and there is no limit in layer thicknesses. Algorithms of this type are widely used in other numerical methods for analyzing layered structures (see [7] and the references therein).

Thus for the practical computation of the integrals (4.13) we solve the differential equations (5.2) by the above mentioned method as long as $(m_{ii}^{\pm})_n + i\beta_n^{\pm 1}$ and $(m_{ij}^{\pm})_n$ are greater than a certain tolerance. Hence for $i \neq j$ one has to sum only a few terms, whereas the sums for $i = j$ involve a factor $(m_{ii}^{\pm})_n$ of the order $O(|n|^{-1})$. However, these summations can also be performed very efficiently with an accuracy comparable with the computer precision. As mentioned above, we discretize (4.11) with finite elements with traces on $\widetilde{\Gamma}^{\pm}$ which are periodic splines with uniformly distributed break points. Then we can use recurrence relations for the Fourier coefficients of spline functions and convergence acceleration methods.

For example, choose a piecewise linear or bilinear finite element discretization of (4.11). Then its traces on $\widetilde{\Gamma}^{\pm}$ are spanned by the shifts of the hat functions. If $\widetilde{\Gamma}^{\pm}$ is divided into $p$ subintervals of equal length then in this basis the integral (4.13) corresponds to an $p \times p$ circulant matrix with the eigenvalues

$$\tau_0 = 2\pi (m_{ii}^{\pm})_0 \ , \ \tau_q = 2\pi \Big(\frac{\sin(\pi q/p)}{\pi}\Big)^4 \sum_{r=-\infty}^{\infty} \frac{(m_{ii}^{\pm})_{rp+q}}{p(r+q/p)^4} \ , \quad q = 1, \ldots, p-1 \ .$$

Since $(m_{ii}^{\pm})_n + i\beta_n^{\pm 1}$ decay exponentially as $|n| \to \infty$, it is advantageous to replace $(m_{ii}^{\pm})_{rp+q}$ by $-i\beta_{rp+q}^{\pm 1}$ and to expand

$$\frac{\beta_{rp+q}^{\pm 1}}{p} = \sqrt{\Big(\frac{k^{\pm 1}}{p}\Big)^2 - \Big(\frac{\gamma}{p}\Big)^2 - \Big(\frac{\alpha}{p} + \Big(r + \frac{q}{p}\Big)\Big)^2}$$

with respect to powers of $|r + q/p|$. Then the corresponding sums can be obtained immediately by using approximation formulas of the generalized Zeta function

$$\zeta(x, s) = \sum_{r=0}^{\infty} (r + x)^{-s} \ .$$

To compute the eigenvalues $\tau_q$, one has to correct the result with the few terms where the sums $|(m_{ii}^{\pm})_{rp+q} + i\beta_{rp+q}^{\pm 1}|$ are larger than the fixed tolerance. Therefore the discretization of the nonlocal boundary operators does not affect the error of the FE discretization of (4.11).

Due to Theorem 4 the FE discretization converges with quasioptimal order, but with a constant depending on $k$. It is well known that for usual FE approximations of the Helmholtz equation these constants can become very large. This pollution error is caused by the fact that the discretization of the Helmholtz equation with wave number $k$ results in an approximate solution possessing a different wave number $k_h$. In [6] we have extended a generalized FEM of minimal pollution, proposed in [1], to the case of piecewise uniform rectangular partitions and bilinear finite elements. Such partitions are well suited to treat binary and multilevel gratings, which are fabricated using modern semiconductor technologies and whose cross sections have a rectangular structure. A simple example is shown in Fig. 1.

For the case of arbitrary polygonal interfaces and therefore nonuniform triangulations, a method with reduced pollution effect is presently not available.

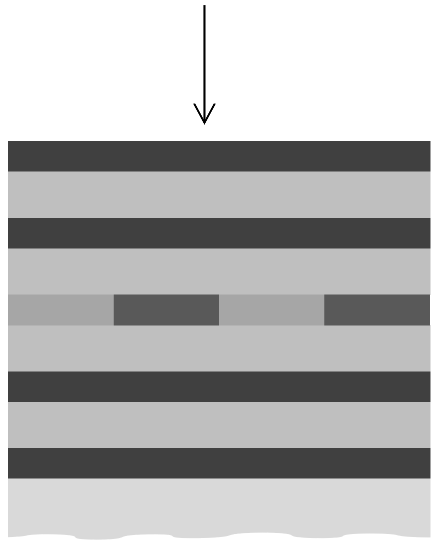

Figure 1. Grating within a thin–film multilayer stack

## 6. Numerical example

The method was used to analyze the diffraction properties of binary and multilayer gratings of different geometries and materials. As an example we report on a thin–film multilayer stack incorporating a single grating in the center layer shown in Fig. 1. This optical transmission filter was introduced in [10]. It was shown that in the case of normal incidence this element is a highly efficient (almost 100%) narrow–line transmission filter close to the wavelength $\lambda = 500$ nm with low sidebands over a range of approximatively 60 nm. This is illustrated in Fig. 2, where the efficiency of the transmitted mode of order zero is shown for two different resolutions.

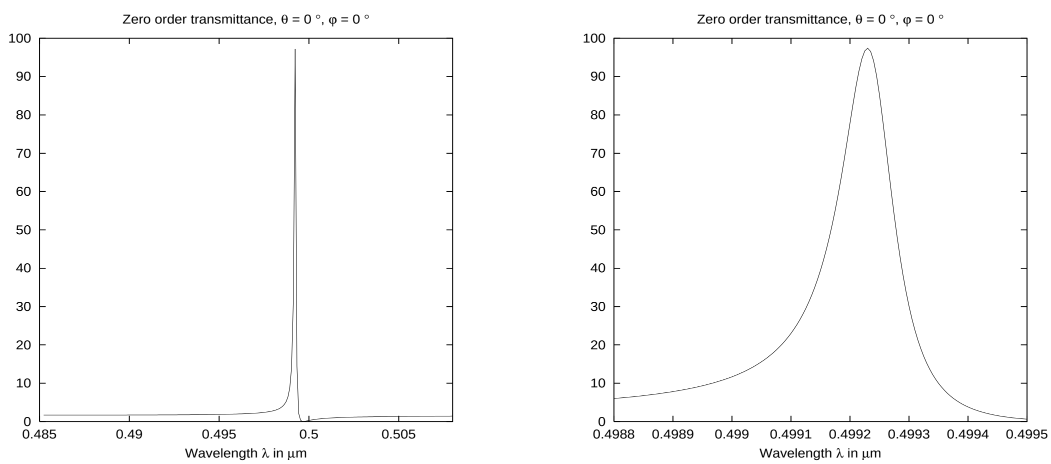


Figure 2. Efficiency of the zero transmitted mode under normal incidence

With the data taken from [10] we analyzed the dependence of the transmittance on small perturbations of the direction of the incident wave. The period of the grating is $d = 300$ nm, the odd homogeneous layers have a thickness of 53.2 nm with dielectric constant $\varepsilon = 2.35^2$. The even homogeneous layers parameters are $\varepsilon = 1.38^2$ and a thickness of 90.6 nm. The grating in the fifth layer consists of two material with $\varepsilon = 2.5^2$ and $\varepsilon = 2.2^2$, respectively, the height is

53.1 nm. The dielectric constant for the cover region is 1.0 and that of the substrate is $\varepsilon = 1.52^2$.

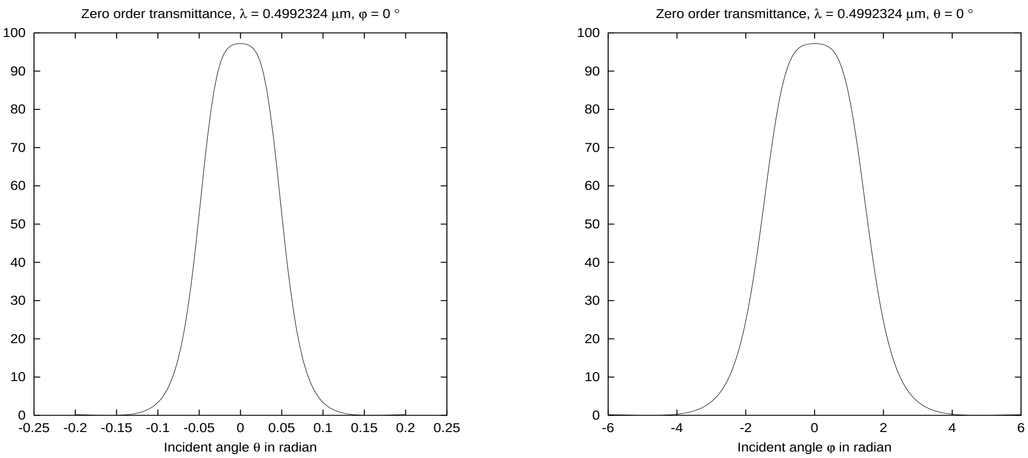


Figure 3. Efficiency of the zero transmitted mode with for incidence fields with wave vectors perpendicular to $x_3$ (left) and to $x_1$ (right)

Fig. 3 depicts the transmission efficiency obtained for the peak wavelength $\lambda = 499.2324$ nm and the wave vectors

$$(\sin\theta, -\cos\theta, 0) \quad \text{and} \quad (0, -\cos\phi, \sin\phi)$$

for small $\theta$ and $\phi$, respectively. We see that small perturbations of the incidence in the $(x_1, x_2)$–plane lead to rapid changes of the transmitted energy, whereas the energy is not so sensitive with respect to perturbations in the $(x_2, x_3)$–plane.

## References

[1] I. Babuška and S. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers, SIAM J. Num. Anal. **34** (1997) 2392–2423.

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

[3] A.-S. Bonnet-Bendhia and F. Starling, Guided waves by electromagnetic gratings and non–uniqueness examples for the diffraction problem, Math. Meth. Appl. Sci. **17** (1994) 305–338.

[4] J. Elschner, R. Hinder, F. Penzel and G. Schmidt, Existence, uniqueness and regularity for solutions of the conical diffraction problem, Math. Models and Methods in Appl. Sci. **10** (2000) 317 – 341.

[5] J. Elschner and G. Schmidt, Diffraction in Periodic Structures and Optimal Design of Binary Gratings I. Direct Problems and Gradient Formulas, Math. Meth. Appl. Sci. **21** (1998) 1297–1342.

[6] J. Elschner and G. Schmidt, Numerical Solution of Optimal Design Problems for Binary Gratings, J. Comput. Physics **146** (1998) 603–626.

[7] L. Li, Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings, J. Opt. Soc. Am. A, **13** (1996) 1024–1035.

[8] R. Petit, (ed.), *Electromagnetic theory of gratings*, Topics in Current Physics, Vol. **22** (Springer, 1980).

[9] R. Petit and F. Zolla, The method of fictitious source as applied to the electromagnetic diffraction of a plane wave by a grating in conical diffraction mounts, SPIE Proc. Vol. 2532 (1997) 374–385.

[10] S. Tibuleac and R. Magnusson, Diffractive narrow–band transmission filters based on guided–mode resonance effects in thin–film multilayers, IEEE Photonics Tech. Letters, **9** (1997) 464–466.