# Analysis and Numerics for the Optimal Design of Binary Diffractive Gratings

J. Elschner, G. Schmidt
WIAS Berlin

*1991 Mathematics Subject Classification.* 78-05, 78A45, 35J20, 65N30, 49J20.

*Keywords.* Diffraction problems, Helmholtz equation, transmission problem, strongly elliptic variational formulation, generalized FEM, gradient methods

**Abstract**

The aim of the paper is to provide the mathematical foundation of effective numerical algorithms for the optimal design of periodic binary gratings. Special attention is paid to fast and reliable methods for the computation of diffraction efficiencies and of the gradients of certain functionals with respect to the parameters of the non–smooth grating profile. The methods are based on a generalized finite element discretization of strongly elliptic variational formulations of quasi periodic transmission problems for the Helmholtz equation in a bounded domain coupled with boundary integral representations in the exterior. We prove uniqueness and existence results for quite general situations and analyse the convergence of the numerical solutions. Furthermore, explicit formulas for the partial derivatives of the reflection and transmission coefficients with respect to the parameters of a binary grating profile are derived. Finally, we briefly discuss the implementation of a gradient type algorithm for solving optimal design problems and present some numerical results.

# 1 Introduction

The practical application of diffractive optics technology has driven the need for mathematical models and numerical codes both to provide rigorous solutions of the full electromagnetic vector-field equations for complicated grating structures, thus predicting performance given the structure, and to carry out optimal design of new structures.

The aim of the present paper is to provide the mathematical foundation of effective numerical algorithms for the optimal design of periodic binary gratings. Special attention is paid to fast and reliable methods for the computation of diffraction efficiencies and of the gradients of certain functionals with respect to the parameters of the non–smooth grating profile.

The case of periodic gratings corresponds to quasi–periodic transmission problems for the Helmholtz equation in the whole plane. Special mathematical difficulties are associated with the numerical solution of these problems due to the highly oscillatory nature of waves and interfaces. Various methods based on Rayleigh expansion, ordinary differential and integral equations and on analytical continuation have been proposed (cf. the monograph [22] and the recent papers [23], [7]), which turned out to be efficient for solving the direct diffraction problem in the case of smooth interfaces between different materials. The situation is much worse for binary structures whose surface profile is given by a piecewise constant function. Here the mathematical complexities are amplified by singularities of the solutions caused by the non-smooth grating profile. Recently, a new variational approach was proposed by Bao and Dobson ([12], [4], [6], [15]) which appears to be well adapted to very general diffraction structures as well as complex materials. Furthermore, this approach may be generalized to the three-dimensional case and can be used in gradient methods for solving optimal design problems. However, the mathematical foundation of this approach seems to be incomplete; in particular, it does not cover all materials occuring in practice and excludes the so–called Rayleigh frequencies.

In the present paper a unified analysis is carried out both for the TE and the TM case leading to more general solvability results and to a rigorous convergence analysis for coupled finite element/boundary element solution methods. It turns out that the approach by Bao and Dobson results in fact from the coupling of the variational method for the Helmholtz equation in a bounded domain with the integral representation for solutions satisfying the radiation condition. This coupling leads here to nonsymmetric, but strongly elliptic variational formulations of both problems allowing the application of well established techniques to their study. Moreover, this technique can be applied to more general problems including the so-called conical diffraction on periodic gratings and various diffraction problems for biperiodic gratings. This will be the topic of a forthcoming paper.

The obtained results are used to derive explicit formulas for the partial derivatives of the reflection and transmission coefficients with respect to the parameters of a binary grating profile. This allows us to compute the gradients for a general class of functionals involving the Rayleigh coefficients of both TE and TM modes. It is proved that these functionals are $C^1$ so that gradient type methods can be applied to find local minima of functionals characterizing desired optical properties of binary gratings. There have been a number of papers from the engineering community that are concerned with optimal design of periodic gratings. In these papers descent methods based on simple difference quotients were used which, however, are very expensive for a large number of parameters. So far rigorous gradient formulas were obtained only for the TE case; see [12], [6], where interface mixture problems have been studied.

The outline of the paper is as follows. In Section 2 we formulate the diffraction problems and reduce them to strongly elliptic variational formulations in a bounded domain. This will be used in Section 3 to study existence and uniqueness questions for the continuous direct and adjoint problems. Further we investigate the regularity of the TM solution on the non–smooth grating profile. In Section 4 we consider problems connected with the optimization of grating efficiencies. We prove explicit formulas for the partial derivatives of the reflection and transmission coefficients with respect to the height and the transition points of the binary grating profile. These formulas are applied to evaluate the gradient of a typical functional occuring in the optimal design of binary gratings. Finally, in Section 5 we study the numerical solution of the diffraction problems. Using the strong ellipticity of the variational formulations a unified convergence analysis is performed for the Galerkin approximation of the equations with truncated hypersingular boundary operators. Furthermore, we study the generalized FEM with minimal pollution for the Helmholtz equation, leading to essentially better numerical results.

# 2 Preliminaries

## 2.1 The Helmholtz equation

Suppose that the whole space is filled with nonmagnetic material with a dielectric coefficient function $\epsilon$, which in Cartesian coordinates $(x_1, x_2, x_3)$ does not depend on $x_3$, is $2\pi$-periodic, $\epsilon(x_1 + 2\pi, x_2) = \epsilon(x_1, x_2)$, and homogeneous above and below certain interfaces. This paper is mainly concerned with the solution of optimal design problems by varying the form of the upper interface, denoted in the sequel by $\Lambda_0$ or $\Gamma$. The lower interface will be denoted by $\Lambda_1$. The surfaces $\Lambda_0$ and $\Lambda_1$ will be assumed to be given by $x_2 = f_j(x_1)$ for certain $2\pi$–periodic functions $f_j$, $j = 0, 1$. The material in the region $G^+$ above the grating surface $\Lambda_0$ has the constant dielectric coefficient $\epsilon = \epsilon^+$, whereas the medium in $G^-$ below $\Lambda_1$ is homogeneous with $\epsilon = \epsilon^-$. The medium in the region $G_0$

between $\Lambda_0$ and $\Lambda_1$ may be inhomogeneous with $\epsilon_0 = \epsilon$, where we assume for simplicity that the function $\epsilon_0$ is piecewise smooth with jumps at certain interfaces $\Lambda_j$, $j = 2, \ldots, \ell$. Assume the grating is illuminated by a monochromatic plane wave

$$\vec{E}^i = \vec{A} \exp(i\alpha x_1 - i\beta x_2) \exp(-i\omega t)\,, \ \vec{H}^i = \vec{B} \exp(i\alpha x_1 - i\beta x_2) \exp(-i\omega t)\,, \qquad (2.1)$$

with $\beta \neq 0$. Here, the complex amplitude vector $\vec{A}$ is perpendicular to the wave vector $\vec{k} = (\alpha, -\beta, 0)$, and $\vec{B} = (\omega\mu)^{-1} \vec{k} \times \vec{A}$ with the everywhere constant magnetic permeability $\mu$.

The incident wave $(\vec{E}^i, \vec{H}^i)$ will be diffracted by the grating, and the total fields will be given by

$$\vec{E}^{up} = \vec{E}^i + \vec{E}^{refl}\,, \ \vec{H}^{up} = \vec{H}^i + \vec{H}^{refl}$$

in the region $G^+$, by $\vec{E}^{int}$ and $\vec{H}^{int}$ in $G_0$ and by

$$\vec{E}^{down} = \vec{E}^{refr}\,, \ \vec{H}^{down} = \vec{H}^{refr}$$

in the region $G^-$. Dropping the factor $\exp(-i\omega t)$, the incident, diffracted and total fields satisfy the time–harmonic Maxwell equations

$$\nabla \times \vec{E} = i\omega\mu\vec{H}\,, \ \nabla \cdot \vec{E} = 0\,, \quad \nabla \times \vec{H} = -i\omega\epsilon\vec{E}\,, \ \nabla \cdot \vec{H} = 0\,.$$

Additionally the tangential components of the total fields are continuous when crossing an interface between two continous media

$$\nu \times (\vec{E}^1 - \vec{E}^2) = 0\,, \quad \nu \times (\vec{H}^1 - \vec{H}^2) = 0 \quad \text{on } \Lambda_j\,, \qquad (2.2)$$

where $\nu$ is the unit normal to the interface $\Lambda_j$. The periodicity of $\epsilon$, together with the form of the incident wave, imply that the physical solutions $\vec{E}$ and $\vec{H}$ are independent of $x_3$ and must be $\alpha$ quasi–periodic in $x_1$, i.e.

$$\vec{E}(x_1 + 2\pi, x_2) = \exp(2\pi i\alpha)\vec{E}(x_1, x_2) \qquad \vec{H}(x_1 + 2\pi, x_2) = \exp(2\pi i\alpha)\vec{H}(x_1, x_2)\,.$$

Further, $\vec{E}$ and $\vec{H}$ can be represented as the superposition of solutions corresponding to the TE case (Field Transverse Electric), where

$$\vec{E}^i = (0, 0, A_3) \exp(i\alpha x_1 - i\beta x_2)\,, \ \vec{H}^i = -(\omega\mu)^{-1}(\beta A_3, \alpha A_3, 0) \exp(i\alpha x_1 - i\beta x_2)$$

and to the TM case (Field Transverse Magnetic) with

$$\vec{E}^i = (A_1, A_2, 0) \exp(i\alpha x_1 - i\beta x_2)\,, \ \vec{H}^i = (\omega\mu)^{-1}(0, 0, \beta A_1 + \alpha A_2) \exp(i\alpha x_1 - i\beta x_2)\,.$$

Denote by $u^i$ the normed transverse component $\vec{E}^i \cdot \vec{x}_3$ for TE or $\vec{H}^i \cdot \vec{x}_3$ for TM. Obviously $u^i = \exp(i\alpha x_1 - i\beta x_2)$ with $k^+ = \omega(\mu\epsilon^+)^{1/2}$, $\alpha = k^+ \sin\theta$, $\beta = k^+ \cos\theta$, and the angle of incidence $\theta \in (-\pi/2, \pi/2)$. Then the diffraction problem for periodic gratings and incoming fields (2.1) splits into two scalar problems associated with the TE and TM mode:

The $\alpha$ quasi–periodic functions $u^\pm(x_1, x_2)$ and $u_0(x_1, x_2)$ equal to either the transverse component $\vec{E} \cdot \vec{x}_3$ for TE or $\vec{H} \cdot \vec{x}_3$ for TM in $G^\pm$ and $G_0$, resp., are easily seen to satisfy, in either case, the Helmholtz equations

$$\begin{aligned} \Delta u^\pm + (k^\pm)^2 u^\pm &= 0 \quad \text{in } G^\pm\,, \\ \Delta u_0 + (k_0)^2\, u_0 &= 0 \quad \text{in } G_0\,, \end{aligned} \qquad (2.3)$$

where $k^\pm = \omega(\mu\epsilon^\pm)^{1/2}$ are constants and $k_0 = \omega(\mu\epsilon_0)^{1/2}$.

The boundary conditions (2.2) are translated into transmission conditions for the unknowns $u^\pm$ and $u_0$ in the following way:

(i) TE mode:

$$\begin{array}{lll} u^+ + u^i = u_0\,, & \partial(u^+ + u^i)/\partial\nu = \partial u_0/\partial\nu & \text{on } \Lambda_0\,, \\ u^- = u_0\,, & \partial u^-/\partial\nu = \partial u_0/\partial\nu & \text{on } \Lambda_1\,, \\ u_0|^+_{\Lambda_j} = u_0|^-_{\Lambda_j}\,, & \partial u_0/\partial\nu|^+_{\Lambda_j} = \partial u_0/\partial\nu|^-_{\Lambda_j} & j = 2,..,\ell\,, \end{array} \tag{2.4}$$

with the incoming wave $u^i = \exp(i\alpha x_1 - i\beta x_2)$, where $u_0|^\pm_{\Lambda_j}$ denote the limits if $u_0$ approaches $\Lambda_j$, $j = 2,\dots,\ell$, from above or below, respectively.

(ii) TM mode:

$$\begin{array}{lll} u^+ + u^i = u_0\,, & (k^+)^{-2}\,\partial(u^+ + u^i)/\partial\nu = k^{-2}\,\partial u_0/\partial\nu & \text{on } \Lambda_0\,, \\ u^- = u_0\,, & (k^-)^{-2}\;\partial u^-/\partial\nu = k^{-2}\,\partial u_0/\partial\nu & \text{on } \Lambda_1\,, \\ u_0|^+_{\Lambda_j} = u_0|^-_{\Lambda_j}\,, & (k^{-2}\,\partial u_0/\partial\nu)|^+_{\Lambda_j} = (k^{-2}\,\partial u_0/\partial\nu)|^-_{\Lambda_j} & j = 2,..,\ell\,. \end{array} \tag{2.5}$$

Note that the components of the fields $\vec{E}$, $\vec{H}$ in the $x_1x_2$–plane can then be computed from the transverse components. We shall assume throughout that the grating material satisfies

$$\epsilon^+ > 0\,,\ \mathrm{Re}\,\epsilon^- > 0\,,\ \mathrm{Im}\,\epsilon^- \geq 0\,, \tag{2.6}$$

$$\mathrm{Re}\,\epsilon_0(x_1,x_2) > 0\,,\ \mathrm{Im}\,\epsilon_0(x_1,x_2) \geq 0\,. \tag{2.7}$$

Note that the case $\mathrm{Im}\,\epsilon > 0$ accounts for materials which absorb energy.

## 2.2 The radiation condition

Because the domain is unbounded in the $x_2$–direction, a radiation condition on the scattering problem must be imposed at infinity, namely that the diffracted fields $u^\pm$ remain bounded and that they should be representable as superpositions of outgoing waves.

Define the coefficients

$$\beta_n^\pm = \beta_n^\pm(\alpha) := \exp(i\gamma^\pm/2)|(k^\pm)^2 - (n-\alpha)^2|^{1/2}\,, \quad n \in \mathbb{Z}\,, \tag{2.8}$$

where

$$\gamma^\pm = \arg((k^\pm)^2 - (n+\alpha)^2)\,, \quad 0 \leq \gamma_j^\pm < 2\pi\,.$$

Note that $\beta_0^+ = \beta$ and that, for real $k^\pm$,

$$\beta_n^\pm = \begin{cases} ((k^\pm)^2 - (n+\alpha)^2)^{1/2}\,, & k^\pm > |n+\alpha|\,, \\ i((n+\alpha)^2 - (k^\pm)^2)^{1/2}\,, & k^\pm < |n+\alpha|\,. \end{cases}$$

Since the $\alpha$ quasi–periodic functions $u^\pm$ are analytic for $x_2 > \max f_0$ resp. $x_2 < \min f_1$ they can be expressed as a sum of plane waves (cf. [21], [9]):

$$u^+ = \sum_{n\in\mathbb{Z}} \left\{ A_n^+ \exp(i(n+\alpha)x_1 + i\beta_n^+ x_2) + B_n^+ \exp(i(n+\alpha)x_1 - i\beta_n^+ x_2) \right\} \tag{2.9}$$
for $x_2 > \max f_0$ ,

$$u^- = \sum_{n\in\mathbb{Z}} \left\{ A_n^- \exp(i(n+\alpha)x_1 - i\beta_n^- x_2) + B_n^- \exp(i(n+\alpha)x_1 + i\beta_n^- x_2) \right\} \quad \text{for } x_2 < \min f_1 \,, \tag{2.10}$$

where $A_n^\pm$, $B_n^\pm$ are complex numbers. The physics of the problem imposes the obvious condition that the diffracted field remains bounded as $|x_2| \to \infty$. Thus we will insist that $u^\pm$ satisfy the outgoing wave condition (OWC) $B_n^\pm = 0$, i.e. they are composed of bounded outgoing plane waves in $G^\pm$, plus the incident incoming wave $u^i$ in $G^+$:

$$\begin{aligned} u^+ - u^i &= \sum_{n\in\mathbb{Z}} A_n^+ \exp(i(n+\alpha)x_1 + i\beta_n^+ x_2)\,, \quad x_2 > \max f\,, \\ u^- &= \sum_{n\in\mathbb{Z}} A_n^- \exp(i(n+\alpha)x_1 - i\beta_n^- x_2)\,, \quad x_2 < \min f_1\,. \end{aligned} \tag{2.11}$$

Since $\beta_n^\pm$ is real for at most finitely many $n$, there are only a finite number of propagating plane waves in the sums of (2.11). Note that physically the case $\beta_n^\pm = 0$ corresponds to a plane wave propagating parallel to the grating. The remaining waves may be called surface waves for the grating since they propagate in the $x_1$–direction and are exponentially decayed as $|x_2| \to \infty$.

In the following we will use the integral representation for $\alpha$ quasi–periodic solutions of the Helmholtz equation satisfying the OWC. These representations are the basis for the treatment of diffraction problems with integral equation methods (cf. [23], [21], [9]). We assume that

$$(k^\pm)^2 \neq (n+\alpha)^2 \quad \text{for all } n \in \mathbb{Z}\,. \tag{2.12}$$

and introduce the $\alpha$ quasi–periodic fundamental solutions

$$\begin{aligned} \Psi^\pm(x) &= \frac{i}{2} \sum_{n\in\mathbb{Z}} H_0^{(1)}\left(k^\pm \sqrt{(x_1 - 2\pi n)^2 + x_2^2}\right) \exp(2\pi i n\alpha) \\ &= \frac{i}{2\pi} \sum_{n\in\mathbb{Z}} \frac{\exp(i(n+\alpha)x_1 + i\beta_n^\pm |x_2|)}{\beta_n^\pm}\,, \end{aligned} \tag{2.13}$$

where $H_0^{(1)}$ is the first Hankel function of order zero. Note that for fixed $\epsilon^\pm$ and incidence angle $\theta$ condition (2.12) is violated for a discrete set of frequencies $\omega_j$, $\omega_j \to \infty$, referred to as Rayleigh frequencies and corresponding to physically anomalous behaviour first observed by Wood.

It is well known that under condition (2.12) the series in (2.13) converge uniformly in compact subsets of $\mathbb{R}^2 \setminus \{0\}$ and that the difference $\Psi^\pm(x) - \log|x|/\pi$ is smooth ([7], [9]). Let us introduce two simple curves $\Gamma^\pm$ which are the restriction to the strip $\{0 \le x_1 \le 2\pi\}$ of the graph of smooth periodic functions lying in $G^\pm$, resp. The single and double layer potentials are defined by

$$\begin{aligned} V^\pm\varphi(x) &:= \int_{\Gamma^\pm} \Psi^\pm(x-y)\,\varphi(y)\,ds\,, \\ K^\pm\varphi(x) &:= \int_{\Gamma^\pm} \frac{\partial}{\partial\nu_y}\Psi^\pm(x-y)\,\varphi(y)\,ds\,, \end{aligned}$$

where the normals $\nu$ to $\Gamma^\pm$ are directed away from the grating profile. Then the $\alpha$ quasi–periodic functions $u^\pm$ solve the Helmholtz equation

$$\Delta u^\pm + (k^\pm)^2 u^\pm = 0$$

and satisfy the outgoing wave condition iff the representations

$$u^\pm = \frac{1}{2}\Big(K^\pm u^\pm - V^\pm \frac{\partial u^\pm}{\partial \nu}\Big) \tag{2.14}$$

are valid in the corresponding exterior domains. Using the jump relations for the potentials and their normal derivatives one obtains the well-known relations between the values of $u^\pm$ and their normal derivatives for $x \in \Gamma^\pm$:

$$\begin{aligned} u^\pm(x) - K^\pm u^\pm(x) + V^\pm \frac{\partial u^\pm}{\partial \nu}(x) &= 0\ , \\ D^\pm u^\pm(x) + \frac{\partial u^\pm}{\partial \nu}(x) + (K^\pm)' \frac{\partial u^\pm}{\partial \nu}(x) &= 0\ , \end{aligned} \tag{2.15}$$

with $(K^\pm)'$ the transpose to the double layer potential operators and $D^\pm$ the hypersingular integral operators

$$\begin{aligned} (K^\pm)'\varphi(x) &= \frac{\partial}{\partial \nu} \int_{\Gamma^\pm} \Psi^\pm(x-y)\,\varphi(y)\,ds\ , \\ D^\pm\varphi(x) &= -\frac{\partial}{\partial \nu} \int_{\Gamma^\pm} \frac{\partial}{\partial \nu_y} \Psi^\pm(x-y)\,\varphi(y)\,ds\ . \end{aligned}$$

## 2.3 Variational formulation

We are interested in $\alpha$ quasi–periodic solutions $u^\pm$, $u_0$ to the TE diffraction problem (2.3), (2.4), and the TM diffraction problem (2.3), (2.5) fulfilling the radiation condition (2.11). For the variational approach to these problems we follow a procedure which couples the variational method for the transmission problem near the inhomogeneities with the integral equation method in the exterior domain. This procedure was introduced in [10] as a symmetric method for coupling finite elements and boundary elements which, in case of self-adjoint boundary value problems, yields symmetric matrices and allows a simple error analysis. In our case the method results in strongly elliptic variational formulations, obtained recently by Bao and Dobson ([12], [5], [6]) using a different approach.

Fix numbers $b > \max f_0$ and $a < \min f_1$, and let $\Omega = (0, 2\pi) \times (a, b)$, $\Omega^\pm = \Omega \cap G^\pm$, $\Omega_0 = \Omega \cap G_0$, $\Gamma^+ = \{x_2 = b\} \cap \overline{\Omega}$, $\Gamma^- = \{x_2 = a\} \cap \overline{\Omega}$ ; see Fig. 1. With a solution of (2.3) we associate the function

$$u = \begin{cases} \exp(-i\alpha x_1)\,(u^+ + u^i) & \text{in } \Omega^+\ , \\ \exp(-i\alpha x_1)\,u_0 & \text{in } \Omega_0\ , \\ \exp(-i\alpha x_1)\,u^- & \text{in } \Omega^-\ . \end{cases} \tag{2.16}$$

defined in $\Omega$, which is $2\pi$–periodic in $x_1$. To formulate the differential problem for $u$ we define $\nabla_\alpha = \nabla + i(\alpha, 0)$, $\Delta_\alpha = \nabla_\alpha \cdot \nabla_\alpha = \Delta + 2i\alpha\partial_{x_1} - \alpha^2$, and let

$$k = \begin{cases} k^+ = \omega(\mu\epsilon^+)^{1/2} & \text{in } \Omega^+\ , \\ k_0 = \omega\,(\mu\epsilon_0)^{1/2} & \text{in } \Omega_0\ , \\ k^- = \omega(\mu\epsilon^-)^{1/2} & \text{in } \Omega^-\ . \end{cases} \tag{2.17}$$

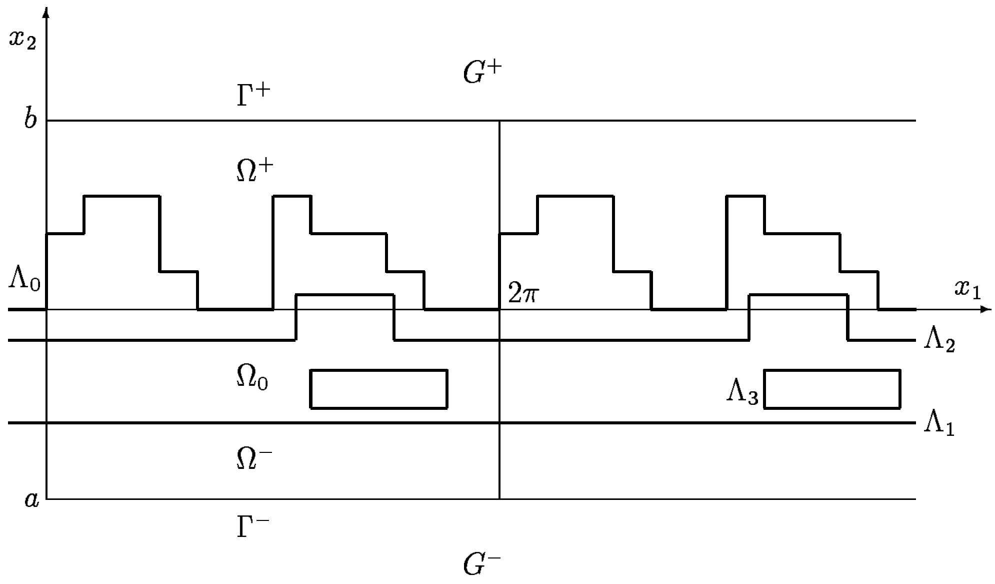


Figure 1: Problem geometry

Let further $H_p^s(\Omega)$, $s \geq 0$, denote the restriction to $\Omega$ of all functions in the Sobolev space $H^s(\mathbb{R}^2)$ which are $2\pi$–periodic in $x_1$. Obviously, $H_p^s(\Omega)$ can be identified with the Sobolev space $H^s(\mathbf{T} \times (a, b))$, where $\mathbf{T}$ stands for the unit circle. Then $(H_p^s(\Omega))'$, the dual space with respect to the scalar product in $L^2(\Omega)$, is isomorphic to the space $H_{\overline{\Omega}}^{-s}(\mathbf{T} \times \mathbb{R}) = \{f \in H^{-s}(\mathbf{T} \times \mathbb{R}) : \text{ supp } f \in \overline{\Omega}\}$ .

The TE diffraction problem can now be formulated as follows. Due to (2.3) the function $u \in H_p^1(\Omega)$ has to satisfy the differential equation

$$(\Delta_\alpha + k^2)u = 0 \quad \text{in } \Omega \tag{2.18}$$

and the transmission conditions (2.4). Integration by parts results in the variational relation

$$\int\limits_{\Omega} \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - \int\limits_{\Omega} k^2 u \, \bar{\varphi} - \int\limits_{\Gamma^+} \frac{\partial u}{\partial \nu} \, \bar{\varphi} - \int\limits_{\Gamma^-} \frac{\partial u}{\partial \nu} \, \bar{\varphi} = 0 \, , \tag{2.19}$$

for all $\varphi \in H_p^1(\Omega)$. Having in mind that $u^\pm$ satisfy the OWC and are smooth on $\Gamma^\pm$ we use relation (2.15), characterizing the exterior fields. Since $\Gamma^\pm$ are straight lines, the integral operators are very simple. So $(K^\pm)' = 0$ and

$$D^\pm \varphi(x) = \frac{1}{2\pi i} \int\limits_0^{2\pi} \sum_{n \in \mathbb{Z}} \beta_n^\pm \exp\left(i(n+\alpha)(x_1 - y_1)\right) \varphi(y_1) \, dy_1 \, .$$

Note that even for Rayleigh frequencies, i.e. condition (2.12) does not hold, we obtain that the functions $u^\pm$ described in (2.9) and (2.10) satisfy the equality

$$D^\pm u^\pm(x) + \frac{\partial u^\pm}{\partial \nu}(x) = 0 \, , \quad x \in \Gamma^\pm \, .$$

Thus one gets in the general case

$$\frac{\partial u}{\partial \nu}\Big|_{\Gamma+} = \exp(-i\alpha x_1)\Big(\frac{\partial u^+}{\partial \nu} + \frac{\partial u^i}{\partial \nu}\Big) = -\exp(-i\alpha x_1)\, D^+u^+ - i\beta\exp(-i\beta b)$$
$$\frac{\partial u}{\partial \nu}\Big|_{\Gamma-} = -\exp(-i\alpha x_1)\, D^-u^- \ ,$$

relating the normal derivatives of $u$ with its boundary values.

Let us define the operators $T_\alpha^\pm$ acting on $2\pi$-periodic functions on $\mathbb{R}$ by

$$(T_\alpha^\pm v)(x) := -\sum_{n\in\mathbb{Z}} i\beta_n^\pm \hat{v}_n \exp(inx)\ , \quad \hat{v}_n = (2\pi)^{-1}\int_0^{2\pi} v(x)\exp(-inx)\,dx\ , \tag{2.20}$$

which are periodic pseudodifferential operators of order 1; see e.g. [14]. In the following the action of these operators on boundary values $u|_{\Gamma\pm} \in H_p^{s-1/2}(\Gamma^\pm)$ of functions $u \in H_p^s(\Omega)$ is denoted by $T_\alpha^\pm u$. Obviously $T_\alpha^\pm$ maps $H_p^s(\Gamma^\pm)$ isomorphically onto $H_p^{s-1}(\Gamma^\pm)$ for any $s \in \mathbb{R}$ if condition (2.12) holds.

With this notations one gets evidently

$$\begin{aligned}\exp(-i\alpha x_1)\, D^+u^+ &= \exp(-i\alpha x_1)D^+(\exp(i\alpha x_1)u) - \exp(-i\alpha x_1)\, D^+u^i \\ &= T_\alpha^+ u + i\beta\exp(-i\beta b)\ ,\\ \exp(-i\alpha x_1)\, D^-u^- &= T_\alpha^- u\ ,\end{aligned}$$

leading to the nonlocal boundary conditions

$$\frac{\partial u}{\partial \nu}\Big|_{\Gamma+} = -T_\alpha^+ u - 2i\beta\exp(-i\beta b)\ , \quad \frac{\partial u}{\partial \nu}\Big|_{\Gamma-} = -T_\alpha^- u \tag{2.21}$$

Thus the coupling of the transmission problem in $\Omega$ and the integral representation for the exterior domain results in the variational formulation for the TE diffraction problem (2.3), (2.4), (2.11)

$$\begin{aligned}B_{TE}(u,\varphi) &:= \int_\Omega \nabla_\alpha u\cdot\overline{\nabla_\alpha\varphi} - \int_\Omega k^2\, u\,\bar\varphi + \int_{\Gamma+}(T_\alpha^+ u)\,\bar\varphi + \int_{\Gamma-}(T_\alpha^- u)\,\bar\varphi \\ &= -\int_{\Gamma+} 2i\beta\exp(-i\beta b)\,\bar\varphi\ , \quad \forall\varphi\in H_p^1(\Omega)\,.\end{aligned} \tag{2.22}$$

Similarly, the TM diffraction problem (2.3), (2.5), (2.11) can be formulated as follows (cf. [6], [5]). Find $u \in H_p^1(\Omega)$ that satisfies

$$\begin{aligned}B_{TM}(u,\varphi) &:= \int_\Omega \frac{1}{k^2}\nabla_\alpha u\cdot\overline{\nabla_\alpha\varphi} - \int_\Omega u\,\bar\varphi + \frac{1}{(k^+)^2}\int_{\Gamma+}(T_\alpha^+ u)\,\bar\varphi + \frac{1}{(k^-)^2}\int_{\Gamma-}(T_\alpha^- u)\,\bar\varphi \\ &= -\frac{1}{(k^+)^2}\int_{\Gamma+} 2i\beta\,\exp(-i\beta b)\,\bar\varphi\ , \quad \forall\,\varphi\in H_p^1(\Omega)\,.\end{aligned} \tag{2.23}$$

This formulation will be written also as the equation

$$\nabla_\alpha\cdot[(1/k^2)\nabla_\alpha u] + u = 0 \quad \text{in } \Omega \tag{2.24}$$

with the boundary conditions (2.21).

Obviously, by

$$B_{TE}(u,\varphi) = (\mathcal{B}_{TE}\,u,\varphi)_{L^2(\Omega)}\ , \quad B_{TM}(u,\varphi) = (\mathcal{B}_{TM}\,u,\varphi)_{L^2(\Omega)}\ , \tag{2.25}$$

the forms $B_{TE}$ and $B_{TM}$ generate bounded linear operators $\mathcal{B}_{TE}$ resp. $\mathcal{B}_{TM}$ acting on $H^1_p(\Omega)$. Since

$$\Big|\int_{\Gamma^\pm} (T^\pm_\alpha u)\,\bar\varphi\Big| \le c\,\|u\|_{H^1(\Omega)}\|\varphi\|_{H^1(\Omega)}$$

it is clear that these operators map $H^1_p(\Omega)$ boundedly into its dual:

$$\mathcal{B}_{TE}, \mathcal{B}_{TM} : H^1_p(\Omega) \longrightarrow (H^1_p(\Omega))'\ .$$

For the calculation of gradients of reflection and transmission coefficients in Sec. 4. solutions of the corresponding adjoint problems are needed. The adjoint TE problem seeks $v \in H^1_p(\Omega)$ such that

$$B_{TE}(\varphi,v) = (\varphi,f^+)_{L^2(\Gamma^+)} + (\varphi,f^-)_{L^2(\Gamma^-)}\ , \quad \text{for all } \varphi\in H^1_p(\Omega)\ , \tag{2.26}$$

where $f^\pm \in H_p^{-1/2}(\Gamma^\pm)$. Note that this problem is equivalent to

$$(\Delta_\alpha + \bar k^2)v = 0 \quad \text{in } \Omega\,, \quad ((T^\pm_\alpha)^* + \partial/\partial\nu)v = f^\pm \quad \text{on } \Gamma^\pm\,,$$

where the adjoint of the boundary operator is given by

$$(T^\pm_\alpha)^* v = \sum_{n\in\mathbb{Z}} i\,\overline{\beta^\pm_n}\,\hat v_n\,\exp(inx_1)\,. \tag{2.27}$$

Moreover, if $v$ is a solution of the adjoint problem (2.26), then the function $w=\bar v$ solves the boundary value problem

$$(\Delta_{-\alpha} + k^2)\,w = 0 \quad \text{in } \Omega\,, \quad (T^\pm_{-\alpha} + \partial/\partial\nu)\,w = \overline{f^\pm} \quad \text{on } \Gamma^\pm\,.$$

The analogue of (2.26) in the TM case reads as follows: Find $v\in H^1_p(\Omega)$ such that

$$B_{TM}(\varphi,v) = (\varphi,f^+)_{L^2(\Gamma^+)} + (\varphi,f^-)_{L^2(\Gamma^-)}\ , \quad \text{for all } \varphi\in H^1_p(\Omega)\ , \tag{2.28}$$

which is equivalent to

$$\nabla_\alpha\cdot((1/\bar k^2)\nabla_\alpha v) + v = 0 \quad \text{in } \Omega\,, \quad ((T^\pm_\alpha)^* + \partial/\partial\nu)v = (k^\pm)^2\,f^\pm \quad \text{on } \Gamma^\pm\,,$$

and $w=\bar v$ solves

$$\nabla_{-\alpha}\cdot((1/k^2)\nabla_{-\alpha} w) + w = 0 \quad \text{in } \Omega\,, \quad (T^\pm_{-\alpha} + \partial/\partial\nu)w = (\overline{k^\pm})^2\,\overline{f^\pm} \quad \text{on } \Gamma^\pm\,.$$

# 3 Solvability and regularity of the diffraction problem

## 3.1 Strong ellipticity of the variational forms

Here we consider an arbitrary grating characterized by a piecewise smooth function $\epsilon$ in $\Omega$ which is constant near the upper and lower boundaries $\Gamma^{\pm}$. We are interested in the existence and uniqueness of solutions for ranges of $\omega$ and incident angles $\theta$. Recall that $k^2 = \omega^2\mu\epsilon$ with $\epsilon = \epsilon^{\pm}$ in $\Omega^{\pm}$, $\epsilon = \epsilon_0$ in $\Omega_0$, and that $\alpha = k^+ \sin\theta$. The results are essentially based on the strong ellipticity of the variational forms $B_{TE}$ and $B_{TM}$. We call a bounded sesquilinear form $a(\cdot,\cdot)$ given on some Hilbert space $X$ strongly elliptic if there exist a complex number $\phi$, $|\phi| = 1$, a constant $c > 0$ and a compact form $q(\cdot,\cdot)$ such that

$$\operatorname{Re} a(\phi u, u) \geq c\|u\|_X^2 - q(u,u)\ , \quad \forall\, u \in X\ .$$

**Theorem 3.1** *Suppose that $k$ satisfies condition* (2.6)*. Then the sesquilinear form $B_{TE}$ is strongly elliptic over $H_p^1(\Omega)$. If additionally condition* (2.7) *holds then the form $B_{TM}$ is strongly elliptic, too.*

*Proof.*

TE mode: Split the sesquilinear form $B_{TE} = B_1 - B_2$ with the compact form

$$B_2(u,\varphi) = \omega^2\mu \int_\Omega \epsilon\, u\, \bar\varphi\ , \quad \text{such that} \quad |B_2(u,\varphi)| \leq c_1\omega^2 \|u\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)}\,. \tag{3.1}$$

The form $\operatorname{Re} B_1(\exp(i\pi/4)\,u,u)$ is coercive over $H_p^1(\Omega)$. Indeed, from condition (2.6) one gets $0 \leq \arg\,((k^{\pm})^2 - (n+\alpha)^2) \leq \pi$ with sharp inequalities for nonreal $k^-$. Therefore $-\pi/4 \leq \arg\,(\exp(-i\pi/4)\,\beta_n^{\pm}) \leq \pi/4$ and for any $u \in H_p^1(\Omega)$

$$\operatorname{Re}\Big(\exp(i\pi/4)\,(T_\alpha^{\pm}u,u)_{L^2(\Gamma^\pm)}\Big) \geq 0\ .$$

Suppose now that there exists a sequence $\{u_j\}$ with $\|u_j\|_{H^1(\Omega)} = 1$ and weakly converging in $H_p^1(\Omega)$ such that

$$\operatorname{Re} B_1(\exp(i\pi/4)\,u_j,u_j) \longrightarrow 0 \quad \text{as } j \to \infty\ .$$

Since

$$\operatorname{Re}\, \exp(i\pi/4) \int_\Omega |\nabla_\alpha u|^2 = (\sqrt{2}/2) \int_\Omega (|\nabla u|^2 + \alpha^2 |u|^2)$$

one gets $\|\nabla u_j\|_{L^2(\Omega)} \to 0$. Hence $u_j$ is a Cauchy sequence in $H_p^1(\Omega)$, and consequently this sequence converges strongly in $H_p^1(\Omega)$ to a function $u_0 = \mathit{const}$. But for constant functions there holds

$$\operatorname{Re}\Big(\exp(i\pi/4)(T_\alpha^{+}u_0,u_0)\Big) = (\sqrt{2}/2)\ k^+\ \cos\theta\,|u_0|^2$$

implying $u_0 = 0$, which contradicts the assumption $\|u_j\|_{H^1(\Omega)} = 1$. Thus we obtain

$$\operatorname{Re} B_1(\exp(i\pi/4)\,u,u) \geq c\,\|u\|_{H^1(\Omega)}^2$$

for any $u \in H^1_p(\Omega)$ and $\omega > 0$.

Remark that under the condition (2.12) there exist constants such that

$$c_1\,\omega \le |\beta_n^\pm| \le c_2\,(|n| + \omega)$$

which implies even

$$\mathrm{Re}\,\Big(\exp(i\pi/4)\,(T_\alpha^\pm u, u)_{L^2(\Gamma^\pm)}\Big) \ge c\,\omega \|u\|^2_{H^{1/2}(\Gamma^\pm)}\,.$$

Hence if $\theta_0 \in (0, \pi/2)$ is some maximum incidence angle and $\omega_0 > 0$ is chosen small enough then the estimate

$$\mathrm{Re}\,B_1(\exp(i\pi/4)\,u, u) \ge c\,\omega \|u\|^2_{H^1(\Omega)} \tag{3.2}$$

is valid, where the constant $c$ does not depend on the frequencies $\omega$ with $0 < \omega \le \omega_0$ and on the incidence angles $\theta$ with $|\theta| \le \theta_0$.

TM mode: Decompose the sesquilinear form $B_{TM} = B_1 - B_2$ with

$$B_1(u,\varphi) = \int_\Omega \frac{1}{k^2}\,\nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} + \frac{1}{(k^+)^2}\int_{\Gamma^+} (T_\alpha^+ u)\,\bar\varphi + \frac{1}{(k^-)^2}\int_{\Gamma^-} (T_\alpha^- u)\,\bar\varphi\,,$$

and the compact form $B_2$ satisfies

$$|B_2(u,\varphi)| \le \|u\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)}\,. \tag{3.3}$$

Similar to the previous considerations one can prove the existence of an angle $\phi$ such that $\mathrm{Re}\,B_1(\exp(i\phi)\,u, u)$ is coercive over $H^1_p(\Omega)$. First we consider the arguments of $-i\beta_n^\pm/(k^\pm)^2$. For $k^\pm$ real we have

$$\arg\,\big(-i\beta_n^\pm/(k^\pm)^2\big) \in \{-\pi/2, 0\}\,. \tag{3.4}$$

If $\arg\,(k^-)^2 = \chi = \pi - \tau$, $\tau \in (0,\pi)$, then

$$\arg\,\big(-i\beta_n^-/(k^-)^2\big) \in (\tau/2 - \pi, \tau - \pi)\,.$$

In view of (2.7) there exists $\phi$ such that $\max(\arg k^2) \le \phi < \pi$ and $\phi > \pi - \tau/2$. Then

$$\arg\,\frac{\exp(i\phi/2)\,\beta_n^-}{i\,(k^-)^2} \in \Big(-\frac{\pi}{2} + \frac{\tau}{4}, \frac{\pi}{2} - \chi\Big)\,.$$

Since

$$\arg\,\frac{\exp(i\phi/2)\,\beta_n^+}{i\,(k^+)^2} \in \Big(-\frac{\tau}{4}, \frac{\phi}{2}\Big)$$

it is clear that

$$\mathrm{Re}\,(k^\pm)^{-2}\big(\exp(i\phi/2)\,T_\alpha^\pm u, u\big) \ge 0\,.$$

On the other hand we have

$$\mathrm{Re}\int_\Omega \frac{1}{k^2}\,\exp(i\phi/2)\,|\nabla_\alpha u|^2 \ge c_1\,\omega^{-2}\big(\|\nabla u\|^2_{L^2(\Omega)} + \alpha^2\,\|u\|^2_{L^2(\Omega)}\big)\,,$$

such that the same arguments as before imply the coerciveness of $\mathrm{Re}\,B_1(\exp(i\phi/2)\,u, u)$.

Again, under the condition (2.12) one obtains

$$\mathrm{Re}\,(k^{\pm})^{-2}(\exp(i\phi/2)\,T_{\alpha}^{\pm}u,u) \geq c\,\omega^{-1}\|u\|^2_{H^{1/2}(\Gamma^{\pm})}\,,$$

such that for some maximum incidence angle $\theta_0 \in (0,\pi/2)$ and $\omega_0 > 0$ sufficiently small

$$\mathrm{Re}\,B_1(\exp(i\phi/2)\,u,u) \geq c\,\omega^{-1}\|u\|^2_{H^1(\Omega)} \tag{3.5}$$

for all frequencies $\omega$ with $0 < \omega \leq \omega_0$ and for all incidence angles $\theta$ with $|\theta| \leq \theta_0$. ∎

**Remark 3.1** Obviously the results of Theorem 3.1 remain true if $k$ is replaced by the complex conjugate $\overline{k}$. Furthermore, the function $\epsilon_0$ (cf. (2.17)) can be chosen quite arbitrarily. In the TE case it suffices that $\epsilon_0 \in L^\infty(\Omega)$, whereas in the TM case one has to suppose that $\epsilon_0^{-1} \in L^\infty(\Omega)$ and $\arg\epsilon_0 \in [0,\pi-\delta]$.

## 3.2 Existence and uniqueness of the variational solution

Several existence and uniqueness results for the problem of diffraction by periodic gratings are known ([7], [9], [1]). Here we give some results for the variational formulations in both TE and TM modes in the general case (2.6), (2.7). Note that existence and uniqueness results were proved for TE polarization in [12] and for the TM case in [5], the latter under the restriction $\mathrm{Re}\,k^2 > 0$.

From the estimates (3.1), (3.2) and (3.3), (3.5) one obtains a first uniqueness result for all sufficiently small frequencies $\omega$.

**Theorem 3.2** *Choose some maximum incidence angle $\theta_0 \in (0,\pi/2)$. Then under the assumptions of Theorem 3.1 there exists a frequency $\omega_0 > 0$ such that the variational problem (2.22) resp. (2.23) admits a unique solution $u \in H^1_p(\Omega)$ for all incidence angles $\theta$ with $|\theta| \leq \theta_0$ and all frequencies $\omega$ with $0 < \omega \leq \omega_0$. Moreover, let $\mathcal{S}$ be an arbitrary set of interfaces $\Lambda_j$, $j = 1,\ldots,\ell$, lying in the strip $a < \tilde{a} < x_2 < \tilde{b} < b$, and for fixed $\Lambda = \{\Lambda_j, j = 1,\ldots,\ell\} \in \mathcal{S}$ let $u_\Lambda$ denote the corresponding variational solution of the TE or TM diffraction problem. Then $\|u_\Lambda\|_{H^1_p(\Omega)} \leq c$, where $c$ is independent of $\theta \in [-\theta_0,\theta_0]$, $\omega \in (0,\omega_0]$ and $\Lambda \in \mathcal{S}$.*

Let us assume that the piecewise smooth interfaces $\Lambda_j$, $j = 1,\ldots,\ell$, may only intersect with angles different from 0 and $2\pi$. Using (2.4) and the elliptic regularity of the Laplacian it can be shown by standard methods that the inverse of the operator $\mathcal{B}_{TE}$, if it exists, maps boundedly

$$\mathcal{B}_{TE}^{-1} : L^2(\Omega) \times H^{1/2}(\Gamma^+) \times H^{1/2}(\Gamma^-) \to H^2_p(\Omega)\,. \tag{3.6}$$

**Corollary 3.1** *Under the assumptions of Theorem 3.2 the solution of the TE diffraction problem (2.22) satisfies $\|u_\Lambda\|_{H^2_p(\Omega)} \leq c$ uniformly in $\theta \in [-\theta_0,\theta_0]$, $\omega \in (0,\omega_0]$ and $\Lambda$.*

**Remark 3.2** The results of Theorem 3.2 and Corollary 3.1 extend to the variational solutions of the adjoint problems (2.26) resp. (2.28).

We now study the uniqueness of the diffraction problems in the case that the frequency $\omega$ is arbitrary, but the grating geometry is fixed. Introduce the set of exceptional values (the Rayleigh frequencies), where condition (2.12) is violated:

$$\mathcal{R}(\epsilon) = \{(\omega,\theta) : \exists\, n \in \mathbb{Z} \ \text{ such that } \ (n + \omega(\mu\epsilon^+)^{1/2}\sin\theta)^2 = \omega^2\mu\epsilon^\pm\}\,. \tag{3.7}$$

**Theorem 3.3** *(i) For all but a sequence of countable frequencies $\omega_j$, $\omega_j \to \infty$, the diffraction problem* (2.22) *resp.* (2.23) *has a unique solution $u \in H^1_p(\Omega)$.*
*(ii) If for $(\omega_0, \theta_0) \notin \mathcal{R}(\epsilon)$ the diffraction problem* (2.22) *resp.* (2.23) *is uniquely solvable then the solution $u$ depends analytically on $\omega$ and $\theta$ in a neighbourhood of this point.*

*Proof.* In view of the proof of Theorem 3.1 the operators $\mathcal{B}_{TE}$ and $\omega^2\mathcal{B}_{TM}$ can be represented in the form

$$\mathcal{B}_{TE} = \mathcal{A}_{TE} - \omega^2\mu\epsilon\ , \quad \omega^2\mathcal{B}_{TM} = \mathcal{A}_{TM} - \omega^2\ , \tag{3.8}$$

where the operators $\mathcal{A}_{TE}$ and $\mathcal{A}_{TM}$ generated by the forms

$$(\mathcal{A}_{TE}u, \varphi) = \int_\Omega \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} + \int_{\Gamma^+} (T^+_\alpha u)\,\bar\varphi + \int_{\Gamma^-} (T^-_\alpha u)\,\bar\varphi$$

$$(\mathcal{A}_{TM}u, \varphi) = \int_\Omega (\mu\epsilon)^{-1} \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} + (\mu\epsilon^+)^{-1}\int_{\Gamma^+} (T^+_\alpha u)\,\bar\varphi + (\mu\epsilon^-)^{-1}\int_{\Gamma^-} (T^-_\alpha u)\,\bar\varphi$$

are invertible mappings from $H^1_p(\Omega)$ onto $(H^1_p(\Omega))'$. Recall that $\alpha = \omega(\mu\epsilon^+)^{1/2}\sin\theta$. Hence $\mathcal{B}_{TE}$ and $\omega^2\mathcal{B}_{TM}$ are compact perturbations of invertible operator functions depending on $\omega > 0$. Moreover, for any fixed $\theta$, $|\theta| < \pi/2$, from the definition (2.20) of $T^\pm_\alpha$ follows that these functions depends analytically on $\omega \notin \mathcal{R}(\epsilon)$. Thus by [16], Theorem I.5.1, the number of linearly independent solutions of the equation $B_{TE}(u, \varphi) = 0$ resp. $B_{TM}(u, \varphi) = 0$, $\varphi \in H^1_p(\Omega)$, is constant for all $\omega \in \mathbb{R}^+ \backslash \mathcal{R}(\epsilon)$ with the possible exception of certain isolated points in that domain. Consequently, due to (3.2) and (3.5) the operators $\mathcal{B}_{TE}$ and $\mathcal{B}_{TM}$ are invertible with the possible exception of a discrete set in $\mathbb{R}^+ \backslash \mathcal{R}(\epsilon)$. Thus assertion (i) is proved if we show that $\omega_0 \in \mathcal{R}(\epsilon)$ cannot be an accumulation point of this set. Since $\theta$ is fixed it follows from the definition of $\beta^\pm_n$ that in some neighbourhood of $\omega_0 \in \mathcal{R}(\epsilon)$ the operator $\mathcal{B}_{TE}$ resp. $\mathcal{B}_{TM}$ can be expanded into a Puiseux series of the form

$$\sum_{j=1}^\infty (\omega - \omega_0)^{j/2} A_j\ ,$$

where the branch of the root is chosen as in (2.8). Replacing $(\omega - \omega_0)^{1/2}$ by $\lambda$ one obtains an analytic operator function in a neighbourhood of $\lambda = 0$, and applying Theorem I.5.1 of [16] to that operator function gives the result. Assertion (ii) follows immediately from the fact that the inverse of an analytic operator function is also analytic. ∎

**Remark 3.3** A less precise version of Theorem 3.3 (i) was stated for TE polarization in [12] and for the TM case in [5]. The above arguments also fill a gap in the proofs of those results. We are grateful to Professor I.C. Gohberg for discussing this topic and for pointing out that Theorem I.5.1 in [16] can be generalized to analytic operator functions with algebroid branching points.
The analytic dependence of solutions was known only for the special case of TE polarization and perfectly conducting gratings (see [18]). Note that the non–smooth behaviour of efficiencies at $(\omega_0, \theta_0) \in \mathcal{R}(\epsilon)$, known as Wood anomalies, is caused by the non–analytic dependence of the inverse operators.

Finally we give a simple proof of an uniqueness result if the imaginary part of the dielectric constant of one of the grating materials is positive. For some special cases this was shown in [1], [7].

**Lemma 3.1** *Suppose that $k$ is piecewise constant with nonnegative imaginary part and that $\operatorname{Im} k(x) > 0$ for all $x$ from some subdomain $\Omega_1 \subset \Omega$ with piecewise smooth boundary. Then the operator $\mathcal{B}_{TE}$ is invertible for all $\omega > 0$. If $\Omega_1$ contains a curve connecting the boundary points $(0, c)$ and $(2\pi, c)$ then the operator $\mathcal{B}_{TM}$ is invertible, too.*

*Proof.* Suppose that $\mathcal{B}_{TE}u = 0$. Then

$$\begin{aligned}\operatorname{Im} B_{TE}(u,u) &= \operatorname{Im}\Big(-\int_\Omega k^2|u|^2 + \int_{\Gamma^+}(T_\alpha^+ u)\,\bar u + \int_{\Gamma^+}(T_\alpha^- u)\,\bar u\Big)\\ &= -\int_\Omega \operatorname{Im} k^2|u|^2 - \operatorname{Re}\sum_{n\in\mathbb{Z}}\beta_n^+|\hat u_n^+|^2 - \operatorname{Re}\sum_{n\in\mathbb{Z}}\beta_n^-|\hat u_n^-|^2 = 0\,,\end{aligned}$$

where $\hat u_n^\pm$ are the Fourier coefficients of $u|_{\Gamma^\pm}$. It follows from (2.8) that all terms of this expression vanish since they are nonnegative. Thus if $k^\pm > 0$ then $\hat u_n^\pm = 0$ for all $n$ with $|n+\alpha| < k^\pm$, and if $\operatorname{Im} k^- > 0$ then $\hat u_n^- = 0$ for all n. Additionally we obtain $u(x) = 0$ for $x \in \Omega_1$, such that in any subdomain of $\Omega$, where $k$ is constant, $u$ solves a Helmholtz equation with the conditions $u = \partial u/\partial\nu = 0$ on some part of the boundary in view of the transmission conditions (2.4). Therefore $u$ must vanish everywhere.

The case $\mathcal{B}_{TM}u = 0$ can be considered analogously. Using (3.4) the conclusions concerning $\hat u_n^\pm$ follow immediately. Further, from

$$\operatorname{Im}\int_\Omega \frac{1}{k^2}\,|\nabla_\alpha u|^2 = \operatorname{Im}\int_\Omega \frac{1}{k^2}\,|\nabla(u\,\exp(i\alpha x_1)|^2 = 0$$

one gets $u\,\exp(i\alpha x_1) = \mathit{const}$ in $\Omega_1$. Since $u$ is $2\pi$-periodic in $x_1$ and $\Omega_1$ ranges from the left to the right boundary of $\Omega$ we derive $u = 0$. ■

**Remark 3.4** Any solution of the homogeneous equations $\mathcal{B}_{TE}u = 0$ and $\mathcal{B}_{TM}u = 0$ has vanishing Rayleigh coefficients $A_n^\pm = 0$ for all $n$ with $\beta_n^\pm = \sqrt{(k^\pm)^2 - (n+\alpha)^2} > 0$.

## 3.3 Additional regularity for the TM diffraction problem

For the calculation of gradients of reflection and transmission coefficients for TM polarisation with respect to variations of the non–smooth grating surface $\Lambda_0$ in Sec. 4 we need auxiliary results about the regularity of the solution near this surface. We will restrict here to the case that $\epsilon$ is constant is some neighbourhood below the grating surface and that the other interfaces $\Lambda_j$, $j = 1, \ldots, \ell$, do not intersect and are smooth. Then the solutions of the equations (2.24) and (2.21) are sufficiently regular everywhere with the exception of a neighbourhood of $\Lambda_0$ which will be denoted in the sequel by $\Gamma$. Since the regularity of the solution is a local problem we may simplify the notations further by assuming that $G_0 = \emptyset$.

Consider the transmission problem (2.24), (2.21), or equivalently (cp. (2.3), (2.5), (2.16)),

$$\Delta u^\pm + (k^\pm)^2 u^\pm = 0 \quad \text{in } \Omega^\pm\,,$$

$$u^+ - u^- = -u^i\,,\ \partial u^+/\partial\nu - C\partial u^-/\partial\nu = -\partial u^i/\partial\nu \quad \text{on } \Gamma\,, \tag{3.9}$$

where $C = (k^+/k^-)^2 \neq 1$, and $u^\pm$ are $\alpha$ quasiperiodic in $x_1$ and satisfy the radiation condition (2.11). Note that the right–hand sides of (3.9) are (infinitely) smooth on $\Gamma$. If the grating profile $\Gamma$ is smooth, then standard regularity theory shows that any solution $(u^+, u^-)$ of (3.9) is contained in $H^s(\Omega^+) \times H^s(\Omega^-)$ for arbitrary $s \geq 1$. For non–smooth $\Gamma$, this is not true, even for $s = 2$, due to the singularities at the corner points.

In this paragraph, we consider the case when $\Gamma$ is a curved polygon, i.e. $\Gamma$ is smooth, with the exception of a finite number of corner points $P_j$ with angles $\delta_j$, $j = 1, ..., J$. In the practically important case of a binary grating, $\Gamma$ consists of straight lines only and $\delta_j \in \{\pi/2, 3\pi/2\}$ for any $j$; see Sec. 4.

For the solution of the transmission problem (3.9), the corner singularities at $P_j$ can be determined with Kondratiev's method of local Mellin transformation [19] (see, in particular, [11], [20] in the case of transmission problems), which implies the following decomposition:

Define the sets

$$\mathcal{A}_j := \left\{ \lambda \in \mathbb{C} : \left( \frac{\sin(\pi - \delta_j)\lambda}{\sin \pi\lambda} \right)^2 = \left( \frac{C+1}{C-1} \right)^2 \right\} \cup \mathbb{N} \,, \tag{3.10}$$

and let $s > 1$ be given such that

$$s - 1 \neq \operatorname{Re} \lambda \quad \text{for all } \lambda \in \cup_{j=1}^J \mathcal{A}_j \,. \tag{3.11}$$

Let $(\varrho_j, \theta_j)$, $\varrho_j(x) = \text{dist}\,(x, P_j)$, be polar coordinates centered at $P_j$. Then

$$u^\pm = \sum_{j=1}^J \sum_{l=1}^{L_j} c_{jl} u_{jl}^\pm + w^\pm \,, \quad \text{with } w^\pm \in H^s(\Omega^\pm) \,, \tag{3.12}$$

when the $u_{jl}^\pm$ are of the form

$$\chi_j(x) d_{jlr}^\pm(\theta_j) \varrho_j(x)^{\lambda_j} \log^r(\varrho_j(x)) \,, \quad r \in \{0, 1, 2\} \,. \tag{3.13}$$

Here $\chi_j \in C_0^\infty(\mathbb{R}^2)$ are cut–off functions near the corner point $P_j$, $d_{jlr}^\pm$ are smooth functions in $\theta_j$, $c_{jl}$ complex constants, and $\lambda_j \in \mathcal{A}_j$ with $0 < \operatorname{Re} \lambda_j < s - 1$.

The $d_{jlr}^\pm$ depend only on the geometry of $\Gamma$ near $P_j$, whereas the constants $c_{jl}$ depend also on $u^i$. We write $X^s$ for the subspace of $H^1(\Omega^+) \times H^1(\Omega^-)$ of all $(u^+, u^-)$ possessing a decomposition (3.12). Then $X^s$ is a Hilbert space with the norm

$$\|(u^+, u^-)\|_{X^s}^2 = \|w^+\|_{H^s(\Omega^+)}^2 + \|w^-\|_{H^s(\Omega^-)}^2 + \sum_{j=1}^J \sum_{l=1}^{L_j} |c_{jl}|^2 \,. \tag{3.14}$$

It follows from (3.12) and (3.13) that near $P_j$ one has the estimates

$$|\nabla^k u^\pm(x)| = O\left( \varrho_j(x)^{\operatorname{Re} \lambda_j^0 - k - \varepsilon} \right) \,, \quad \forall \varepsilon > 0 \,,\ 0 \leq k \leq s - 1 \,, \tag{3.15}$$

where $\nabla^k$ denotes the vector of all partial derivatives of order $k$ and $\lambda_j^0$ is the solution $\lambda_j$ of the transcendental equation

$$\left( \frac{\sin(\pi - \delta_j)\lambda}{\sin \pi\lambda} \right)^2 = \left( \frac{C+1}{C-1} \right)^2 \tag{3.16}$$

with minimal $\operatorname{Re}\lambda_j \in (0,1)$, and with the convention that $\operatorname{Re}\lambda_j^0 := 1$ if there is no root of (3.16) with real part $< 1$. Define

$$\mu^0 = \min\{\operatorname{Re}\lambda_j^0 : j = 1, \ldots, J\}\,. \tag{3.17}$$

**Remark 3.5** It was shown in [11, Lemma 6.2] that if $k^- > 0$, i.e. $C > 0$, then $\mu^0 \in (1/2, 1)$. In particular, for a binary grating, (3.16) takes the form

$$2\cos(\pi\lambda/2) = \sigma\frac{C-1}{C+1}\,, \quad \sigma = \pm 1 \tag{3.18}$$

for any $j$, which easily implies $\mu^0 \in (2/3, 1)$ if $k^- > 0$. However, if $\operatorname{Im} k^- > 0$ then, even for a right angle at $P_j$, one may have $\operatorname{Re}\lambda_j^0 \leq 1/2$ so that the solution $u^\pm$ to (3.9) does not belong to $H^{3/2}(\Omega^\pm)$, in general. More precisely, for a binary grating we show that $\mu^0 > 1/2$ holds if and only if the condition

$$P^2 - Q^2 < 2 \quad \text{with } P = \operatorname{Re}\frac{C-1}{C+1} = \frac{|C|^2-1}{|C+1|^2}\,,\ Q = \operatorname{Im}\frac{C-1}{C+1} = \frac{2\operatorname{Im} C}{|C+1|^2} \tag{3.19}$$

is satisfied. Note that (3.19) is always valid if $\operatorname{Re} k^- \geq \operatorname{Im} k^-$.

Taking real and imaginary parts of (3.18), we obtain for $\lambda = \mu + i\kappa$, $\mu, \kappa \in \mathbb{R}$,

$$2\cos(\pi\mu/2)\cosh(\pi\kappa/2) = \sigma P\,,\ 2\sin(\pi\mu/2)\sinh(\pi\kappa/2) = \sigma Q\,,\ \sigma = \pm 1\,.$$

or equivalently

$$R(\mu) = \frac{P^2}{\cos^2(\pi\mu/2)} - \frac{Q^2}{\sin^2(\pi\mu/2)} = 4\,,\ \sinh^2(\pi\kappa/2) = \frac{Q^2}{4\sin^2(\pi\mu/2)} \tag{3.20}$$

Assume (3.20) has a solution $\mu \in (0, 1/2]$. Then $\cos^2(\pi\mu/2) \geq 1/2$ , $\sin^2(\pi\mu/2) \leq 1/2$, and the first equation of (3.20) implies $P^2 - Q^2 \geq 2$. Conversely, if $P^2 - Q^2 \geq 2$ then there exists $\mu \in (0, 1/2]$ such that $R(\mu) = 4$, since $R(1/2) = 2(P^2 - Q^2) \geq 4$ and $R(\mu) \to -\infty$ as $\mu \to 0$. (Note that $Q \neq 0$ if $\operatorname{Im} k^- > 0$ and $\operatorname{Re} k^- > 0$.)

Finally, we observe that $\mu^0 > 0$ may be arbitrarily small if we choose $\operatorname{Re} C$ sufficiently close to $-3$ and $|\operatorname{Im} C|$ sufficiently small.

To obtain a regularity result in weighted spaces of differentiable functions, which will be applied to binary gratings in Sec. 4, set $\varrho(x) = \min\{\varrho_j(x) : j = 1, \ldots, J\}$ and introduce the spaces

$$Y^\mu = \left\{(u^+, u^-) \in C(\bar\Omega^+) \times C(\bar\Omega^-) : \varrho^{1-\mu}\nabla u^\pm \in C(\bar\Omega^\pm)\right\}\,, \quad 0 \leq \mu \leq 1\,,$$

equipped with the canonical norm $\|u^+\|_\mu + \|u^-\|_\mu$, with

$$\|u^\pm\|_\mu = \max_{x \in \bar\Omega^\pm}\left\{|u^\pm(x)| + \varrho(x)^{1-\mu}\nabla u^\pm(x)|\right\}\,.$$

Let $\mu^0 \in (0, 1]$ be the number defined in (3.17). Then it follows from the definition of the space $X^s$ (cf. (3.12), (3.14)) and (3.15) that the continuous embeddings

$$X^2 \hookrightarrow H^s(\Omega^+) \times H^s(\Omega^-)\,, \quad \text{for any } s \in [1, 1+\mu^0)\,, \tag{3.21}$$

$$X^s \hookrightarrow Y^\mu\,, \quad \text{for any } s > 2 \quad \text{and } \mu \in [0, \mu^0)\,, \tag{3.22}$$

hold. Summarizing (in particular, (3.12), (3.21), (3.22) and Remark 3.5), we then have:

**Theorem 3.4** *Let $(u^+, u^-) \in H^1(\Omega^+) \times H^1(\Omega^-)$ be a solution of the transmission problem* (3.9). *Then $(u^+, u^-) \in X^s$ for any $s > 1$ satisfying condition* (3.11). *Moreover, $(u^+, u^-) \in H^{1+\mu}(\Omega^+) \times H^{1+\mu}(\Omega^-)$ and $(u^+, u^-) \in Y^\mu$ for any $\mu \in [0, \mu^0)$, where $\mu^0$ is given by* (3.17). *Notice that $\mu^0 \in (1/2, 1)$ if condition* (3.19) *is satisfied.*

Returning to the variational formulation (2.23) of problem (3.9), we obtain from (2.16) and the above theorem:

**Corollary 3.2** *Let $u \in H^1_p(\Omega)$ be a solution of the TM diffraction problem* (2.23). *Then $u \in H^{1+s}_p(\Omega)$ for any $s$ with $0 < s < \min(1/2, \mu^0)$ and*

$$\max_{x \in \bar{\Omega}} |u(x)| + \sup_{x \in \Omega \setminus \Gamma} |\varrho(x)^{1-\mu} \nabla u(x)| < \infty \tag{3.23}$$

*for any $\mu \in [0, \mu^0)$. Note that $\partial(\exp(i\alpha x_1)u)/\partial\nu$ suffers a jump on $\Gamma$, hence $u \notin H^s(\Omega)$ for $s \geq 3/2$, in general.*

**Remark 3.6** The inverse of the operator $\mathcal{B}_{TM}$, if it exists, maps boundedly

$$\mathcal{B}_{TM}^{-1} : (H_p^{1-s}(\Omega))' \times H^{s-1/2}(\Gamma^+) \times H^{s-1/2}(\Gamma^-) \to H_p^{1+s}(\Omega) \,, \ |s| < \min(1/2, \mu^0) \,, \tag{3.24}$$

and

$$\mathcal{B}_{TM}^{-1} : (H_p^{1-\mu}(\Omega))' \times H^{\mu-1/2}(\Gamma^+) \times H^{\mu-1/2}(\Gamma^-) \to Z^\mu \,, \quad \mu \in [0, \mu^0) \,, \tag{3.25}$$

where

$$Z^\mu = \{u \in H^1_p(\Omega) : u|_{\Omega^\pm} \in H^{1+\mu}(\Omega^\pm)\} \,.$$

Finally, for the calculation of gradients in Sec. 4.3, we need a uniform version of Corollary 3.1. Let $\Gamma_0 \subset \Omega$ be a piecewise smooth grating profile with $J$ corner points, and assume that, in some (fixed) neighbourhood $U_j$ of the $j$th corner, $\Gamma_0$ consists of two straight lines intersecting with the angle $\delta_j (j = 1, ..., J)$. Let $\mathcal{S}$ be a set of grating profiles $\Gamma$ sufficiently close to $\Gamma_0$ and such that, for each $\Gamma \in \mathcal{S}$, $\Gamma \cap U_j$ is a translate of $\Gamma_0 \cap U_j$ $(j = 1, ..., J)$, and $\Gamma$ is smooth outside these neighbourhoods. Defining the space $X^s_\Gamma$, $\Gamma \in \mathcal{S}$, as in (3.12)–(3.14), we observe that the singular functions $u^\pm_{jl}$ occuring in the corresponding decomposition (3.12) are simply translates of the functions (3.13) (for $\Gamma_0$).

Suppose further that the assumptions of Theorem 3.2 are satisfied, and for $\Gamma \in \mathcal{S}$ let $(u^+_\Gamma, u^-_\Gamma) \in H^1(\Omega^+) \times H^1(\Omega^-)$ resp. $u_\Gamma \in H^1_p(\Omega)$ denote the corresponding (unique) solution to (3.9) resp. (2.23). Then the Mellin transformation techniques of [19] and [20] imply, for any $s > 1$ satisfying (3.11),

$$\|(u^+_\Gamma, u^-_\Gamma)\|_{X^s_\Gamma} \leq c \,, \quad \text{uniformly in } \theta, \omega \text{ and } \Gamma \,. \tag{3.26}$$

Denoting by $\varrho_\Gamma(x)$ the distance of $x$ to the set of corner points of $\Gamma$, in analogy to Corollary 3.1 one obtains from (3.26):

**Corollary 3.3** *There exists $c > 0$ independent of $\theta \in [-\theta_0, \theta_0]$, $\omega \in (0, \omega_0]$ and $\Gamma \in \mathcal{S}$ such that*

$$\|u_\Gamma\|_{H^{1+s}(\Omega)} \leq c \,, \quad \textit{for any } s \in (0, \min(1/2, \mu^0)) \,, \tag{3.27}$$

$$\max_{x \in \bar{\Omega}} |u_\Gamma(x)| + \sum_{x \in \Omega \setminus \Gamma} |\varrho_\Gamma(x)^{1-\mu} \nabla u_\Gamma(x)| \leq c \,, \quad \textit{for any } \mu \in [0, \mu^0) \,. \tag{3.28}$$

# 4 Minimization problems for binary gratings

## 4.1 Optimization of grating efficiencies

Define the finite sets of indices

$$P^{\pm} = \{n \in \mathbb{Z} : \beta_n^{\pm} \in \mathbb{R}\}\,,$$

where $\beta_n^{\pm}$ is given by (2.8). Then the Rayleigh amplitudes $A_n^+$ $(n \in P^+)$ resp. $A_n^-$ $(n \in P^-)$, which are called the reflection resp. transmission coefficients, correspond to the propagating modes in (2.11). Note that $P^- = \emptyset$ if $\operatorname{Im} k^- \neq 0$.

Let $u$ be the solution of the TE or TM variational problem (2.22) or (2.23). The reflection and transmission coefficients are determined by the traces of $u$ on the artifical boundaries $\Gamma^{\pm}$ (cp. (2.11), (2.16)):

$$\begin{aligned}
A_n^+ &= (2\pi)^{-1} \exp(-i\beta_n^+ b) \int\limits_{\Gamma^+} u \exp(-inx_1)\,, \quad n \in P^+ \backslash \{0\}\,, \\
A_0^+ &= -\exp(-2i\beta b) + (2\pi)^{-1} \exp(-i\beta b) \int\limits_{\Gamma^+} u\,, \qquad (4.1)\\
A_n^- &= (2\pi)^{-1} \exp(i\beta_n^- a) \int\limits_{\Gamma^-} u \exp(-inx_1)\,, \quad n \in P^-\,.
\end{aligned}$$

Then the reflected and transmitted efficiencies in the TE case are defined by

$$e_n^{TE,\pm} = (\beta_n^{\pm}/\beta)|A_n^{\pm}|^2\,, \quad n \in P^{\pm}\,, \tag{4.2}$$

and in the TM case by

$$e_n^{TM,+} = (\beta_n^+/\beta)|A_n^+|^2\,, \quad n \in P^+\,, \quad e_n^{TM,-} = (k^+/k^-)^2(\beta_n^-/\beta)|A_n^-|^2\,, \quad n \in P^-\,. \tag{4.3}$$

For lossless gratings, i.e. all optical indices $k$ are real, the principle of conservation of energy then, in either case, yields the relation

$$\sum_{n\in P^+} e_n^+ + \sum_{n \in P^-} e_n^- = 1\,. \tag{4.4}$$

Consider a binary grating profile $\Gamma$ which is composed of a finite number of horizontal and vertical segments and is determined by the height $H$ and by, say $m+1$, transition points $0 = t_0 < t_1 < ... < t_{m-1} < t_m = 2\pi$; see Fig. 2. Since $t_0$ and $t_m$ are assumed to be fixed, we write $\Gamma = \Gamma(t_1, ..., t_{m-1}, H)$.

We retain the notation of Sec. 2.3 (cf. Fig. 1) and denote the vertical segment of $\Gamma = \Lambda_0$ at $t_j$ $(j = 1, ..., m-1)$ by $\Sigma_j$. The union of all upper horizontal segments lying in $\Omega = \Omega^+ \cup \Omega_0$ is denoted by $\Sigma_m$.

A typical minimization problem occurring in the optimal design of binary gratings is the following. Assume that the number of transition points is fixed and, for given numbers $c_n^{TE,\pm}, c_n^{TM,\pm} \in \{-1, 0, 1\}$, define the functional

$$\begin{aligned}
J(\Gamma) &= J(t_1, ..., t_{m-1}, H) \\
&:= \sum_{n \in P^+} \left(c_n^{TE,+} e_n^{TE,+} + c_n^{TM,+} e_n^{TM,+}\right) + \sum_{n\in P^-} \left(c_n^{TE,-} e_n^{TE,-} + c_n^{TM,-} e_n^{TM,-}\right)\,. \qquad (4.5)
\end{aligned}$$

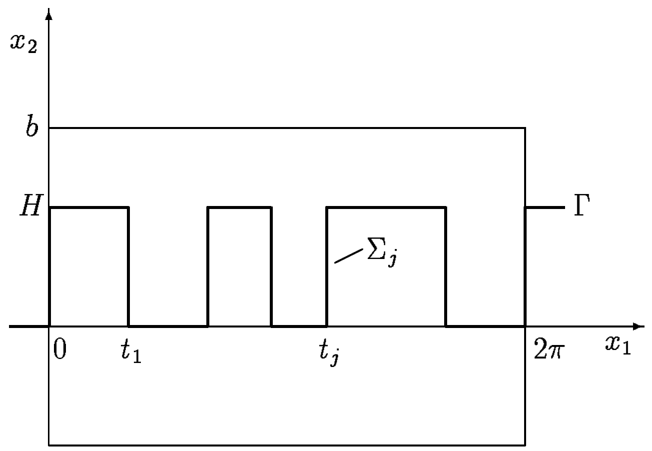


Figure 2: Binary grating, $m = 6$

Note that the efficiencies $e_n^\pm$ are functions of the grating profile $\Gamma$, and thus they are functions of $t_1, ..., t_{m-1}, H$. Now the minimization problem reads as follows:

Find a binary grating profile $\Gamma^0 = \Gamma(t_1^0, ..., t_{m-1}^0, H^0)$ such that

$$\min_{(t_1,...,t_{m-1},H)\in K} J(\Gamma) = J(\Gamma^0)\,, \tag{4.6}$$

where $K$ is some compact set in the parameter space $\mathbb{R}^m$ reflecting e.g. natural constraints on the design of the profile. Note that the choice $c_n^\pm = -1$ resp. $c_n^\pm = 1$ in (4.5) amounts to maximizing resp. minimizing the efficiency of the corresponding reflected or transmitted propagating mode of order $n$.

To find local minima of problem (4.6), the method of gradient descent can be applied. Thus we must calculate the gradient of $J$, $\nabla J(\Gamma) = (D_j J(\Gamma))_1^m$, where e.g. for $j = 1$

$$\begin{aligned} D_1 J(\Gamma) &= \lim_{h\to 0} h^{-1}(J(\Gamma_h) - J(\Gamma)) \\ &= \lim_{h\to 0} h^{-1}(J(t_1 + h, t_2, ..., H) - J(t_1, t_2, ..., H)) \end{aligned} \tag{4.7}$$

Here $\Gamma_h$ denotes the binary profile with the height $H$ and the transition points $t_1 + h$, $t_2, ..., t_{m-1}$. Analogously, $D_j J(\Gamma)$ $(j = 2, ..., m-1)$ denote the partial derivatives with respect to the other transition points, and $D_m J(\Gamma)$ will denote the derivative with respect to the height $H$.

From (4.1), (4.2) and (4.5) we obviously have, for $j = 1, ..., m$,

$$\begin{aligned} D_j J(\Gamma) = &\sum_{n\in P^+} 2(\beta_n^+/\beta)\Big\{ c_N^{TE,+}\mathrm{Re}\,(A_n^{TE,+}(\Gamma) D_j A_n^{TE,+}(\Gamma)) \\ &+ c_n^{TM,+}\mathrm{Re}\,(A_n^{TM,+}(\Gamma) D_j A_n^{TM,+}(\Gamma))\Big\} \\ + &\sum_{n\in P^-} 2(\beta_n^-/\beta)\Big\{ c_n^{TE,-}\mathrm{Re}\,(A_n^{TE,-}(\Gamma) D_j A_n^{TE,-}(\Gamma)) \\ &+ (k^+/k^-)^2 c_n^{TM,-}\mathrm{Re}\,(A_n^{TM,-}(\Gamma) D_j A_n^{TM,-}(\Gamma))\Big\}\,. \end{aligned} \tag{4.8}$$

Therefore, we have to calculate the partial derivatives $D_j A_n^\pm(\Gamma)$ of the reflection and transmission coefficients in both the TE and TM case. This will be done in the following

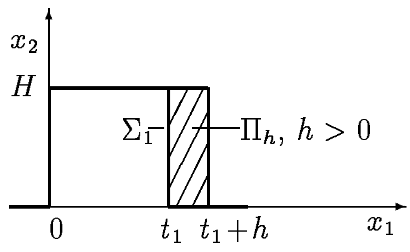


Figure 3: Geometry for the calculation of gradients

two paragraphs. Furthermore, it will be shown that those partial derivatives depend continuously on the grating parameters $t_1, ..., t_{m-1}, H$, so that $J$ turns out to be a $C^1$ functional, at least in the case when the frequency $\omega$ is sufficiently small. So we assume for the rest of this section that the conditions of Theorem 3.2 are satisfied.

**Remark 4.1** Once one has derived explicit formulas for the partial derivatives of the reflection and transmission coefficients with respect to the parameters of the grating profile it is possible to compute the gradients for a much more general class of functionals involving the Rayleigh coefficients for a given range of incidence angles or wavelengths.

## 4.2 Calculation of gradients. TE case

We fix $n \in P^+$ and derive a formula for the partial derivative $D_1 A_n^+(\Gamma)$ of the Rayleigh coefficient of the $n$th reflected mode in the TE case. At the end of this paragraph we present formulas for the other derivatives and for the transmission coefficients.

Let $u$ be the solution of the TE transmission problem (2.22) and let $u_h$ denote the solution of the corresponding problem for the profile $\Gamma_h = \Gamma(t_1 + h, t_2, ..., H)$:

$$\begin{aligned} B_{TE}^h(u_h, \varphi) := & \int_\Omega \nabla_\alpha u_h \cdot \overline{\nabla_\alpha \varphi} + \int_\Omega k_h^2 \, u_h \, \bar\varphi + \int_{\Gamma^+} (T_\alpha^+ u_h) \, \bar\varphi + \int_{\Gamma^+} (T_\alpha^- u_h) \, \bar\varphi \\ = & - \int_{\Gamma^+} 2i\beta \exp(-i\beta b) \, \bar\varphi \, , \quad \forall \varphi \in H_p^1(\Omega) \, , \end{aligned} \tag{4.9}$$

where (see Fig. 3)

$$k_h = \begin{cases} k_0 \, , & h > 0 \\ k^+ \, , & h < 0 \end{cases} \quad \text{in } \Pi_h \, , \; k_h = k \quad \text{in } \Omega \backslash \Pi_h \, . \tag{4.10}$$

By (4.1) and the definition of $D_1 A_n^+(\Gamma)$ (cf. (4.7)), we have

$$D_1 A_n^+(\Gamma) = \lim_{h \to 0} \frac{\exp(-i\beta_n^+)}{h} \int_{\Gamma^+} (u_h - u) \exp(-inx_1) \, dx_1 \, . \tag{4.11}$$

Let $w$ be the solution of the adjoint transmission problem (cf. (2.26))

$$B_{TE}(\varphi, w) = \int_{\Gamma^+} \varphi \, \exp(-inx_1) \, dx_1 \, , \quad \forall \varphi \in H_p^1(\Omega) \, . \tag{4.12}$$

Then obviously

$$h^{-1}\int\limits_{\Gamma^+}(u_h-u)\exp(-inx_1)\,dx_1 = h^{-1}B_{TE}(u_h-u,w) = h^{-1}(B_{TE}(u_h,w)-B^h_{TE}(u_h,w))$$
$$= h^{-1}\int\limits_{\Omega}(k_h^2-k^2)u_h\bar{w} = ((k_0)^2-(k^+)^2)|h|^{-1}\int\limits_{\Pi_h}u_h\bar{w}\,.$$

Together with (4.11), this implies the formula

$$D_1A_n^+(\Gamma) = (2\pi)^{-1}\exp(-i\beta_n^+b)((k_0)^2-(k^+)^2)\int\limits_{\Sigma_1}u\,\bar{w}\,dx_2\,, \tag{4.13}$$

provided we have shown that

$$\lim_{h\to 0}|h|^{-1}\int\limits_{\Pi_h}u_h\bar{w}\,dx = \int\limits_{\Sigma_1}u\,\bar{w}\,dx_2\,. \tag{4.14}$$

*Proof of* (4.14):
Corollary 3.1 and Remark 3.2 imply that $w\in H^2_p(\Omega)$ and, for $h_0>0$ sufficiently small,

$$\|u_h\|_{H^2_p(\Omega)}\le c\quad\text{for all } h \text{ with } |h|\le h_0\,. \tag{4.15}$$

Moreover, one has the compact embeddings

$$H^2_p(\Omega)\hookrightarrow H^1_p(\Omega)\,,\ H^2_p(\Omega)\hookrightarrow C_p(\bar{\Omega})\,, \tag{4.16}$$

where $C_p$ denotes the space of continuous functions which are $2\pi$–periodic in $x_1$. Therefore, given any sequence $u_n=u_{h_n}$, $h_n\to 0$, we can select a subsequence, again denoted by $u_n$, such that $u_n\to\tilde{u}$ in $H^1_p(\Omega)$ for some $\tilde{u}$. It is easily seen that $\tilde{u}$ satisfies the variational problem (2.22). Indeed,

$$\begin{aligned} B_{TE}(\tilde{u},\varphi) &= \lim_{n\to\infty}B_{TE}(u_n,\varphi) = \lim_{n\to\infty}(B_{TE}(u_n,\varphi)-B^{h_n}_{TE}(u_n,\varphi))+B^{h_n}_{TE}(u_n,\varphi)\\ &= \lim_{n\to\infty}\int\limits_{\Pi_{h_n}}(k_h^2-k^2)u_n\bar{\varphi}-\int\limits_{\Gamma^+}2i\beta\exp(-i\beta b)\,\bar{\varphi}\,. \end{aligned}$$

Hence $u_h$ converges in $H^1_p(\Omega)$ to the unique solution $u$ as $h\to 0$, and the same is true for the space $C_p(\Omega)$.

Consequently, given any $\varepsilon>0$, we observe that

$$|h|^{-1}\int\limits_{\Pi_h}|(u-u_h)\bar{w}|\le\varepsilon$$

and, by considering the Riemann sums for the continuous integrands,

$$\left||h|^{-1}\int\limits_{\Pi_h}u\bar{w}-\int\limits_{\Sigma_1}u\bar{w}\right|\le\varepsilon$$

for all sufficiently small $|h|$, which proves (4.14). ∎

**Remark 4.2** Let $h$ be the vector $(h_j)_1^m$, and let $\Gamma_h$ denote the profile with transition points $t_j + h_j$ and height $H + h_m$. Then, by applying Corollary 3.1 to the set of profiles $\mathcal{S} = \{\Gamma_h : \ |h| \leq h_0\}$ and using similar arguments as above, we obtain $D_1 A_n^+(\Gamma_h) \to D_1 A_n^+(\Gamma)$ as $|h| \to 0$, i.e. the partial derivative depends continuously on the parameters of the profile.

We finally collect the formulas for all components of the gradient of $A_n^\pm$; the proof is completely analogous to that of (4.13):

$$\begin{aligned} D_j A_n^\pm(\Gamma) &= \frac{(-1)^{j-1}}{2\pi} \exp(-i\beta_n^\pm b)((k_0)^2 - (k^+)^2) \int_{\Sigma_j} u\bar{w}_\pm dx_2 \,, \quad j = 1, ..., m-1 \,, \\ D_m A_n^\pm(\Gamma) &= \frac{1}{2\pi} \exp(-i\beta_n^\pm b)((k_0)^2 - (k^+)^2) \int_{\Sigma_m} u\bar{w}_\pm dx_1 \,, \end{aligned} \tag{4.17}$$

where $u$ is the solution to the TE diffraction problem (2.22), $w_+$ solves the adjoint problem (4.12) and $w_-$ the adjoint problem

$$B_{TE}(\varphi, w_-) = \int_{\Gamma^-} \varphi \, \exp(-inx_1) \, dx_1 \,, \quad \forall \varphi \in H_p^1(\Omega) \,. \tag{4.18}$$

Recall that $\Sigma_m$ is the union of all upper horizontal segments of $\Gamma$, whereas $\Sigma_j$ ($j = 1, ..., m-1$) denotes the vertical segment at the transition point $t_j$.

## 4.3 Calculation of gradients. TM case

Retaining the notation of the preceding paragraph, we wish to compute the partial derivative $D_1 A_n^+(\Gamma)$ of the $n$th reflected TM mode from the relation (4.11), where $u$ is the solution to the problem (2.24), and $u_h$ solves

$$\begin{aligned} B_{TM}^h(u_h, \varphi) &:= \int_\Omega \frac{1}{k_h^2} \nabla_\alpha u_h \cdot \overline{\nabla_\alpha \varphi} - \int_\Omega u_h \, \bar{\varphi} + \frac{1}{(k^+)^2} \int_{\Gamma^+} (T^+ u_h) \, \bar{\varphi} + \frac{1}{(k^-)^2} \int_{\Gamma^-} (T_\alpha^- u_h) \, \bar{\varphi} \\ &= -\frac{1}{(k^+)^2} \int_{\Gamma^+} 2i\beta \, \exp(-i\beta b) \, \bar{\varphi} \,, \quad \forall \, \varphi \in H_p^1(\Omega) \,. \end{aligned} \tag{4.19}$$

with $k_h$ defined in (4.10). If $w$ is the solution to the adjoint problem (compare (2.28))

$$B_{TM}(\varphi, w) = \int_{\Gamma^+} \varphi \, \exp(-inx_1) \, dx_1 \,, \quad \forall \varphi \in H_p^1(\Omega) \,. \tag{4.20}$$

then one obtains

$$\begin{aligned} & h^{-1} \int_{\Gamma^+} (u_h - u) \exp(-inx_1) \, dx_1 = h^{-1} B_{TM}(u_h - u, w) \\ &= h^{-1} \int_\Omega \left( \frac{1}{k^2} - \frac{1}{k_h^2} \right) \nabla_\alpha u_h \cdot \overline{\nabla_\alpha w} = h^{-1} \int_{\Pi_h} (k_h^2 - k^2) \frac{1}{k_h^2} \nabla_\alpha u_h \cdot \frac{1}{k^2} \overline{\nabla_\alpha w} \qquad (4.21) \\ &= ((k_0)^2 - (k^+)^2) \, |h|^{-1} \int_{\Pi_h} \left\{ \frac{1}{(k^+ k_0)^2} \, \partial_{x_2} u_h \partial_{x_2} \bar{w} + \frac{1}{k_h^2} \, \partial_{x_1,\alpha} u_h \frac{1}{k^2} \, \overline{\partial_{x_1,\alpha} w} \right\} , \end{aligned}$$

where $\partial_{x_1,\alpha} = \partial_{x_1} + i\alpha$, cp. (2.23).

To determine the limit on the right–hand side of (4.11), we thus have to compute

$$\lim_{h\to 0} |h|^{-1} \int\limits_{\Pi_h} gr(u_h) \cdot \overline{gr(w)} \,, \tag{4.22}$$

where

$$gr(u_h) = \left( \frac{1}{k_h^2}\, \partial_{x_1,\alpha} u_h \,,\ \frac{1}{k^+ k_0} \partial_{x_2} u_h \right) \,,\ gr(w) = \left( \frac{1}{\overline{k}^2}\, \partial_{x_1,\alpha} w, \frac{1}{\overline{k^+ k_0}}\, \partial_{x_2} w \right) .$$

Let $\Sigma_{1,h}$ be the vertical segment of the profile $\Gamma_h$ at the transition point $t_1 + h$, with the convention that $\Sigma_{1,0} = \Sigma_1$ and $\Gamma_0 = \Gamma$. Fix $h_0 > 0$ sufficiently small, and for any sufficiently small $\varepsilon \geq 0$, consider the rectangle $R_\varepsilon = (t_1 - h_0, t_1 + h_0) \times (\varepsilon, H - \varepsilon)$; see Fig. 4.

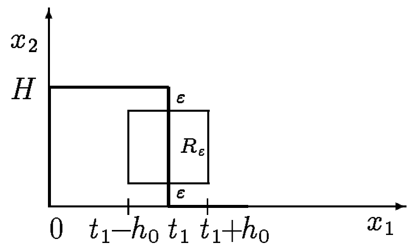


Figure 4: Geometry for gradients in TM case

Corollary 3.3 applied to the set of profiles $\Gamma_h$, $|h| \leq h_0$, implies

$$\|u_h\|_{H_p^{1+s}(\Omega)} \leq c \,,\ |h| \leq h_0 \,, \quad \text{for some } s > 1 \,, \tag{4.23}$$

$$\max_{x\in \bar{R}_0} \left| (x_2(H - x_2))^{1-\mu} gr(u_h) \right| \leq c \,,\ |h| \leq h_0 \,, \quad \text{for any } \mu \in [0, \mu^0) \,. \tag{4.24}$$

These estimates also include $u_0 = u$ and hold for the solution $w$ to (4.20). Note that the expressions on the left side of (4.24) are well defined since $\partial_{x_2} u_h$ and $k_h^{-2} \partial_{x_1,\alpha} u_h$ are continuous across $\Sigma_{1,h} \cap R_\varepsilon$ for any $\varepsilon > 0$, because of (3.28) and the transmission conditions

$$u_h|_+ = u_h|_- \,,\ \frac{1}{(k_h^+)^2} \frac{\partial}{\partial\nu} \left(e^{i\alpha x_1} u_h\right)\Big|_+ = \frac{1}{(k_h^-)^2} \frac{\partial}{\partial\nu} \left(e^{i\alpha x_1} u_h\right)\Big|_- \quad \text{on } \Gamma_h \,,$$

where the plus resp. minus sign denotes the limit as the interface $\Gamma_h$ is approached from the region above resp. below.

Thus the expressions $gr(u_h)$ and $gr(w)$ are well defined on $\Sigma_{1,h}$ and $\Sigma_1$, respectively. In particular, we have on $\Sigma_1$

$$gr(u) = \frac{1}{k^+ k_0} \left( \frac{k_0}{k^+}\, \partial_{x_1,\alpha} u|_+ \,,\ \partial_{x_2} u|_+ \right) = \frac{1}{k^+ k_0} \left( \frac{k^+}{k_0}\, \partial_{x_1,\alpha} u|_- \,,\ \partial_{x_2} u|_- \right) \,, \tag{4.25}$$

and the same holds for $w$, with $k$ replaced by $\bar{k}$.

Now we prove that the limit (4.22) exists and that

$$\lim_{h\to 0} |h|^{-1} \int\limits_{\Pi_h} gr(u_h) \cdot \overline{gr(w)} = \int\limits_{\Sigma_1} gr(u) \cdot \overline{gr(w)}\, dx_2 \,, \tag{4.26}$$

which, together with (4.11) and (4.21), then implies

$$D_1 A_n^+(\Gamma) = \frac{\exp(-i\beta_n^+ b)}{2\pi} ((k_0)^2 - (k^+)^2) \int\limits_{\Sigma_1} gr(u) \cdot \overline{gr(w)}\, dx_2 \,. \tag{4.27}$$

*Proof of* (4.26):
Here we use the regularity results given in Subsec. 3.3. Therefore we assume that $k_0$ is constant in a neighbourhood below the grating profile $\Gamma$ and that condition (3.19) with $C = (k^+/k_0)^2$ is fulfilled. Then estimate (4.24) applied to $u_h$ and $w$ gives, with some $\mu > 1/2$,

$$\begin{aligned} &\frac{1}{|h|} \int\limits_{\Pi_h \setminus R_\varepsilon} \left| gr(u_h) \cdot \overline{gr(w)} \right| dx + \int\limits_{\Sigma_1 \setminus R_\varepsilon} \left| gr(u) \cdot \overline{gr(w)} \right| dx_2 \\ &\le c \left\{ \int\limits_0^\varepsilon x_2^{2\mu-2} dx_2 + \int\limits_{H-\varepsilon}^H (H - x_2)^{2\mu-2} dx_2 \right\} \le c\varepsilon^\delta \,, \end{aligned}$$

where $\delta = 2\mu - 1 > 0$ and $c$ is independent of $\varepsilon$ and $h$.

Thus it is sufficient to verify, for any fixed $\varepsilon > 0$ sufficiently small, that

$$\lim_{h\to 0} |h|^{-1} \int\limits_{\Pi_h \cap R_\varepsilon} gr(u_h) \cdot \overline{gr(w)}\, dx = \int\limits_{\Sigma_1 \cap R_\varepsilon} gr(u) \cdot \overline{gr(w)}\, dx_2 \,. \tag{4.28}$$

By (4.23) and the compact embedding $H_p^{1+s}(\Omega) \hookrightarrow H_p^1(\Omega)$, we obtain as in the proof of (4.14) that $u_h \to u$ in $H_p^1(\Omega)$. Indeed, for any $\varphi$ from the dense subset $H_p^1(\Omega) \cap C^\infty(\bar\Omega)$ of $H_p^1(\Omega)$ and for any sequence $u_h$ converging to some $\tilde u$ in $H_p^1(\Omega)$ we have

$$B_{TM}^h(u_h, \varphi) - B_{TM}(\tilde u, \varphi) = \int\limits_{\Pi_h} \left( \frac{1}{k_h^2} - \frac{1}{k^2} \right) \nabla_\alpha u_h \cdot \overline{\nabla_\alpha \varphi} + B_{TM}(u_h - \tilde u, \varphi) \to 0 \,,\ h \to 0$$

since by (4.23) there holds $\nabla_\alpha u_h \cdot \nabla_\alpha \varphi \in L^1(\Omega)$ uniformly. Hence $B_{TM}(u - \tilde u, \varphi) = 0$ for any $\varphi \in H_p^1(\Omega)$.

Moreover, (3.26) (or standard regularity theory for transmission problems) gives

$$\sup_{x \in \bar R_\varepsilon \setminus \Sigma_{1,h}} \left\{ |\nabla u_h(x)| + |\nabla^2 u_h(x)| \right\} \le c \,,\ |h| \le h_0 \,;$$

note that $\bar R_\varepsilon$ stays away from the corner points of $\Gamma_h$. Together with the continuity of $gr(u_h)$ across $\Sigma_{1,h} \cap R_\varepsilon$ and a compactness argument, we then have $gr(u_h) \to gr(u)$ $(h \to 0)$ in the norm of $C(\bar R_\varepsilon)$ and, since $gr(w)$ is also continuous on $\bar R_\varepsilon$,

$$\lim_{h\to 0} |h|^{-1} \int\limits_{\Pi_h \cap R_\varepsilon} (gr(u_h) - gr(u)) \cdot \overline{gr(w)} = 0 \,.$$

Finally, the continuity of the integrands implies

$$\lim_{h\to 0} \frac{1}{|h|} \int\limits_{\Pi_h \cap R_\varepsilon} gr(u) \cdot \overline{gr(w)}\, dx = \int\limits_{\Sigma_1 \cap R_\varepsilon} gr(u) \cdot \overline{gr(w)}\, dx_2 \,,$$

which completes the proof of (4.28) and hence that of (4.26). ■

**Remark 4.3** So far we have not been able to prove (4.27) in the case when condition (3.19) is violated. Since then $\mu^0 \leq 1/2$, the function $gr(u) \cdot \overline{gr(w)}$ might be non–integrable on $\Sigma_1$; see (3.15) and Remark 3.5.
However, for materials occuring in practice the condition (3.19) on the optical index is violated only in some exceptional cases, e.g. for silver and aluminium in a certain range of small wavelengths (less than 450 nm). Additionally, simple numerical examples indicate that even in the case of small $\mu^0 > 0$ the Rayleigh coefficients $A_n^+$ depend smoothly on the variation of transition points or the height of a binary grating. So we believe that the restriction $\mu^0 > 1/2$ is only of technical nature.

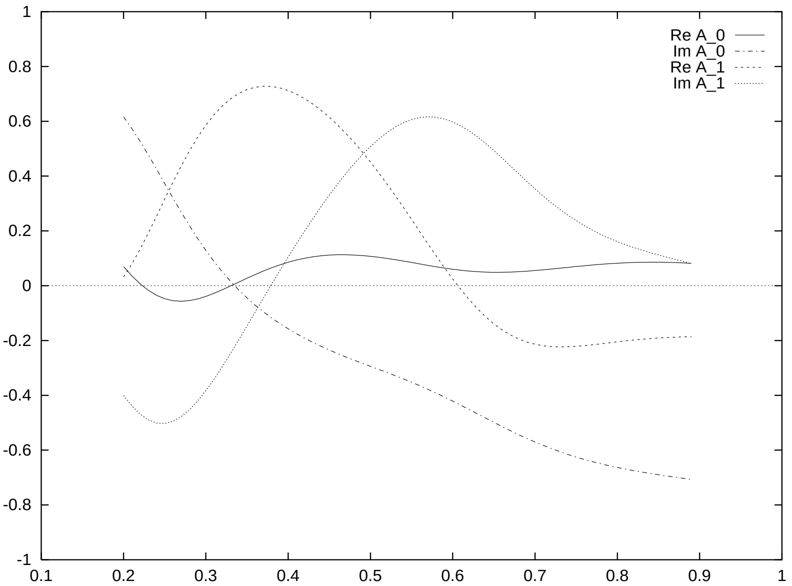


Figure 5: Real and imaginary part of Rayleigh coefficients for a simple binary grating with $C = -3 +$ i 0.01 versus variation of the transition point.

**Remark 4.4** Applying Corollary 3.3 to the profiles $\Gamma_h$, $h = (h_1, ..., h_m)$, considered in Remark 4.2 and using the above arguments, one can show that $D_1 A_n^{\pm}(\Gamma_h) \to D_1 A_n^{\pm}(\Gamma)$ as $|h| \to 0$.

Proceeding as in the proof of (4.27), we get the following analogues of the formulas (4.17) in the TM case:

$$\begin{aligned} D_j A_n^{\pm}(\Gamma) &= \frac{(-1)^{j-1}}{2\pi} \exp(-i\beta_n^{\pm} b)((k_0)^2 - (k^+)^2) \int\limits_{\Sigma_j} gr(u) \cdot \overline{gr(w_\pm)}\, dx_2 \,, \\ &\quad j = 1, ..., m-1 \,, \\ D_m A_n^{\pm}(\Gamma) &= \frac{\exp(-i\beta_n^{\pm} b)}{2\pi}((k_0)^2 - (k^+)^2) \int\limits_{\Sigma_m} gr_H(u) \cdot \overline{gr_H(w_\pm)}\, dx_1 \,. \end{aligned} \tag{4.29}$$

Here, $u$ is the solution of the direct TM problem (2.23), $w_+$ solves the adjoint problem (4.20), and $w_-$ the adjoint problem

$$B_{TM}(\varphi, w) = \int_{\Gamma^-} \varphi \exp(-inx_1)\, dx_1\,, \quad \forall \varphi \in H_p^1(\Omega)\,. \tag{4.30}$$

Furthermore, $gr(u)$ is defined by (4.25),

$$gr_H(u) = \frac{1}{k^+k_0}\left(\partial_{x_1,\alpha} u\Big|_+, \frac{k_0}{k^+}\,\partial_{x_2} u\Big|_+\right) = \frac{1}{k^+k_0}\left(\partial_{x_1,\alpha} u\Big|_-, \frac{k^+}{k_0}\,\partial_{x_2} u\Big|_-\right), \tag{4.31}$$

and the corresponding expressions for $w_\pm$ are obtained by replacing $k$ with $\bar{k}$.

# 5 Generalized finite element solution

Since the sesquilinear forms $B_{TE}$ and $B_{TM}$ corresponding to the TE and TM diffraction problems are strongly elliptic it is natural to use a Galerkin method for solving the corresponding direct and adjoint variational problems. Here we describe some aspects of the discretization of these problems with bilinear finite elements given on a piecewise uniform rectangular partitioning of $\Omega$. The traces of these functions on $\Gamma^\pm$ are the linear boundary elements so that the presented approach is in fact a coupled FE/BE method for treating the diffraction on periodic gratings. Our choice of bilinear test and trial functions on a uniform mesh is motivated by the singular behavior of the solutions at the non-smooth grating profile, by some special features of the FE solution of Helmholtz-type equations and by the simple implementation of the methods.

## 5.1 Stability and convergence

The error analysis is very simple due to the strong ellipticity of the coupled variational formulations (2.22) and (2.23). Since for nonsmooth $\Gamma$ the solutions are not better than $H^2$, in general, the trial functions are restricted to piecewise bilinear ones, although the convergence analysis can be applied to any spaces of FE functions. We remark that the Galerkin method for solving the direct problems was investigated by using a different technique in [4] and [5], for the TM case under the restriction $\operatorname{Re} k^2 > 0$.

Let $\Omega_h$ be a partitioning of $\Omega$ into simple rectangles of the size $h_1 \times h_2$, by $S_h$ we denote the subspace of $H_p^1(\Omega)$, formed by the bilinear functions on $\Omega_h$. In the following we consider a family of these spaces assuming that the quotient $h_1/h_2$ is bounded from below and above, and set $h = \sqrt{h_1\, h_2}$.

The finite element solutions $u_h \in S_h$ of the direct and adjoint problems can be determined from the linear systems

$$B_{TE}(u_h, \varphi_h) = \int_{\Gamma^+} f^+ \varphi_h + \int_{\Gamma^-} f^- \varphi_h\,, \quad \text{for all } \varphi_h \in S_h\,, \tag{5.1}$$

or correspondingly

$$B_{TM}(u_h, \varphi_h) = \frac{1}{(k^+)^2}\int_{\Gamma^+} f^+ \varphi_h + \frac{1}{(k^-)^2}\int_{\Gamma^-} f^- \varphi_h\,, \quad \text{for all } \varphi_h \in S_h\,, \tag{5.2}$$

with some smooth periodic functions $f^{\pm}$ given on $\Gamma^{\pm}$. Using Theorems 3.1 and 3.3 as well as (3.6) and (3.24) one gets by standard Galerkin techniques (see [24], Chapter 12)

**Theorem 5.1** *Suppose that $k \in L^{\infty}(\Omega)$ takes constant values $k^{\pm}$ in a neighbourhood of $\Gamma^{\pm}$, respectively, and satisfies condition (2.6). For TM problems suppose additionally that $k^{-1} \in L^{\infty}(\Omega)$ and condition (2.7) is valid. Then for all but a sequence of countable frequencies $\omega_j$, $|\omega_j| \to \infty$, and all sufficiently small $h$ the Galerkin equations (5.1) resp. (5.2) are uniquely solvable. If the exact solution $u \in H_p^s(\Omega)$, $1 < s \leq 2$, then the difference between the finite element solutions and the exact solution can be estimated by*

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{s-1} \|u\|_{H^s(\Omega)} \, , \quad \|u - u_h\|_{L^2(\Omega)} \leq C h^{2s-2} \|u\|_{H^s(\Omega)}$$

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

In practical computations the operators $T_{\alpha}^{\pm}$ cannot be computed from the infinite series expansion (2.20). Denote their truncation by

$$T_{\alpha,N}^{\pm} v = - \sum_{|n| \leq N} i \beta_n^{\pm} \hat{v}_n \exp(i n x_1)$$

and let $P_N^{\pm}$ be the bounded mapping $H_p^1(\Omega) \to H_p^{1/2}(\Gamma^{\pm})$ defined by

$$P_N^{\pm} u = \sum_{|n| \leq N} \hat{u}_n^{\pm} \exp(i n x_1) \, .$$

Here $\hat{u}_n^{\pm}$ denote the Fourier coefficients of $u|_{\Gamma^{\pm}}$. Furthermore, set $Q_N^{\pm} u = u|_{\Gamma^{\pm}} - P_N^{\pm} u$. By standard arguments one easily verifies

**Lemma 5.1** *There exists a constant $c > 0$ such that for any $N > 0$ and any $u \in H_p^1(\Omega)$*

$$\|\nabla u\|_{L^2(\Omega)} \geq c \, \|Q_N^{\pm} u\|_{H^{1/2}(\Gamma^{\pm})} \, .$$

Denote by $B_{TE}^N$ and $B_{TM}^N$ the sesquilinear forms corresponding to the truncated operators $T_{\alpha,N}^{\pm}$ In the practical computations the approximate solution of the direct and adjoint problems in TE or TM mode is obtained as the FE solution of the equations

$$B_{TE}^N(u, \varphi) = \int_{\Gamma^+} f^+ \varphi + \int_{\Gamma^-} f^- \varphi \, , \quad \text{for all } \varphi \in H_p^1(\Omega) \, , \tag{5.3}$$

or correspondingly

$$B_{TM}^N(u, \varphi) = \frac{1}{(k^+)^2} \int_{\Gamma^+} f^+ \varphi + \frac{1}{(k^-)^2} \int_{\Gamma^-} f^- \varphi \, , \quad \text{for all } \varphi \in H_p^1(\Omega) \, , \tag{5.4}$$

for some $N$, which has to be specified. The error analysis can be performed similarly to the untruncated case and relies on the following

**Lemma 5.2** *For any $N > 0$ the forms $B_{TE}^N$ and $B_{TM}^N$ are strongly elliptic over $H_p^1(\Omega)$. If the operator $\mathcal{B}_{TE}$ respectively $\mathcal{B}_{TM}$ is invertible then there exists $N_0$ and a constant $c > 0$ such that for all $N \geq N_0$ and $u \in H_p^1(\Omega)$ the inequalities*

$$\sup_{\|\varphi\|_{H_p^1(\Omega)}=1} |B_{TE}^N(u, \varphi)| \geq c \, \|u\|_{H^1(\Omega)} \quad \textit{resp.} \quad \sup_{\|\varphi\|_{H_p^1(\Omega)}=1} |B_{TM}^N(u, \varphi)| \geq c \, \|u\|_{H^1(\Omega)}$$

*are valid. Moreover, there exists $h_0 > 0$ such for any $N > N_0$ and $h < h_0$ the Galerkin approximations $B^N_{TE}(u_h, \varphi_h)$ or $B^N_{TM}(u_h, \varphi_h)$, $u_h, \varphi_h \in S_h$, are stable, i.e.*

$$\sup_{\|\varphi_h\|_{H^1_p(\Omega)}=1} |B^N_{TE}(u_h, \varphi_h)| \geq c\,\|u_h\|_{H^1(\Omega)} \quad \textit{resp.} \quad \sup_{\|\varphi\|_{H^1_p(\Omega)}=1} |B^N_{TM}(u_h, \varphi_h)| \geq c\,\|u_h\|_{H^1(\Omega)}$$

*with constants not depending on $N$ and $h$.*

*Proof.* To show the strong ellipticity make use of

$$\begin{aligned}
&(T^{\pm}_{\alpha,N} u, u) = (T^{\pm}_{\alpha,N} P^{\pm}_N u, P^{\pm}_N u)\ ,\\
&\int_\Omega |\nabla u|^2 \geq \frac{1}{2}\int_\Omega |\nabla u|^2 + c\,\|Q^{\pm}_N u\|^2_{H^{1/2}(\Gamma^\pm)}\ ,\\
&\|P^{\pm}_N u\|^2_{H^{1/2}(\Gamma^\pm)} + \|Q^{\pm}_N u\|^2_{H^{1/2}(\Gamma^\pm)} = \|u\|^2_{H^{1/2}(\Gamma^\pm)}\ ,
\end{aligned}$$

and proceed as in the proof of Theorem 3.1.

Now the proof of the second assertion follows from standard arguments. Suppose that there exists a weakly converging sequence $\{u^N\}$ with $\|u^N\|_{H^1(\Omega)} = 1$ such that

$$\sup_{\|\varphi\|_{H^1_p(\Omega)}=1} |B^N_{TE}(u^N, \varphi)| \to 0\ . \tag{5.5}$$

If the form $B_{TE}$ has a trivial kernel then $u^N$ converge weakly to 0. Indeed, since the operators $(T^{\pm}_{\alpha,N})^*$ converge strongly to $(T^{\pm}_{\alpha})^*$ (cf. (2.27)) and

$$B_{TE}(u^N, \varphi) = B^N_{TE}(u^N, \varphi) + ((T^+_\alpha - T^+_{\alpha,N})u^N, \varphi) + ((T^-_\alpha - T^-_{\alpha,N})u^N, \varphi)\ ,$$

one concludes that $B_{TE}(u^N, \varphi) \to 0$ for any fixed $\varphi$. Hence $u^N \to 0$ strongly in $L^2(\Omega)$. But the strong ellipticity of $B^N_{TE}$ was shown by the estimate

$$\mathrm{Re}\, B^N_{TE}(\exp(-i\pi/4)u^N, u^N) \geq c_1\,\omega\|u^N\|_{H^1(\Omega)} - c_2 \int_\Omega |k|^2\,|u^N|^2$$

implying

$$|B^N_{TE}(u^N, u^N)| \geq c\omega\ .$$

The stability of the Galerkin method for $B^N_{TE}$ uniformly in $h$ follows analogously if one chooses a weakly converging sequence $u_{h_N} \in S_{h_N}$ and any strongly converging sequence $\varphi_{h_N} \in S_{h_N}$ fulfilling (5.5).

Obviously, the same arguments can be used to prove the assertions in the case of the form $B^N_{TM}$, too. ∎

Lemma 5.2 implies that the Galerkin equations of (5.3) or (5.4) are uniquely solvable if $N > N_0$ and $h < h_0$. Moreover, the solutions $u^N_h \in S_h$ converge to the exact solution $u^N$ of (5.3) or (5.4) with the rates given in Theorem 5.1. To estimate $u - u^N_h$ it suffices therefore to consider the difference $u - u^N$. In the TE case one has

$$\begin{aligned}
B^N_{TE}(u - u^N, \varphi) &= B_{TE}(u, \varphi) - B^N_{TE}(u - u^N, \varphi) - B^N_{TE}(u^N, \varphi)\\
&= ((T^+_{\alpha,N} - T^+_\alpha)u, \varphi) - ((T^-_{\alpha,N} - T^-_\alpha)u, \varphi)\ .
\end{aligned}$$

Since by Lemma 5.2

$$\|u - u^N\|_{H^1_p(\Omega)} \le c \sup_{\|\varphi\|_{H^1_p(\Omega)}=1} |B^N_{TE}(u - u^N, \varphi)| ,$$

it suffices therefore to estimate $|(T^\pm_\alpha - T^\pm_{\alpha,N})u, \varphi)|$ for the solution $u$ of the original problem. By (2.11) it has the form

$$u = \sum_{n\in\mathbb{Z}} A^+_n \exp(inx_1 + i\beta^+_n x_2) , \quad x_2 > \max f_0 ,$$
$$u = \sum_{n\in\mathbb{Z}} A^-_n \exp(inx_1 - i\beta^-_n x_2) , \quad x_2 < \min f_1 .$$

Recall that $\Gamma^+ = \{x_2 = b\}$, take $\tilde b \in (\max f_0, b)$ and denote $\widetilde\Gamma^+ = \{x_2 = \tilde b\}$. Then

$$u|_{\Gamma^+} = \sum_{n\in\mathbb{Z}} \hat u_n(b) \exp(inx_1) = \sum_{n\in\mathbb{Z}} \hat u_n(\tilde b) \exp(i\beta^+_n(b - \tilde b)) \exp(inx_1) ,$$

where $\hat u_n(b)$ denote the Fourier coefficients of the function $u(x_1, b)$. Hence

$$\begin{aligned}|((T^+_\alpha - T^+_{\alpha,N})u, \varphi)| &\le \Big\{ \sum_{|n|>N} |\hat u_n(\tilde b)|^2 |\exp(2i\beta^+_n(b - \tilde b))| \frac{|\beta^+_n|^2}{|n|} \Big\}^{1/2} \Big\{ \sum_{|n|>N} |n|\, |\hat\varphi_n|^2 \Big\}^{1/2} \\ &\le N^{1-s} \exp\Big( -(b-\tilde b)\sqrt{(N-|\alpha|)^2 - (k^+)^2} \Big) \|u\|_{H^{s-1/2}(\widetilde\Gamma^+)} \|\varphi\|_{H^{1/2}(\Gamma^+)} ,\end{aligned}$$

if $N > |\alpha| + k^+$ and $s > 1$. A similar estimate is valid for $|((T^-_\alpha - T^-_{\alpha,N})u, \varphi)|$ if $N$ is sufficiently large. Therefore

$$\|u - u^N\|_{H^1_p(\Omega)} \le c\, N^{1-s}\, \gamma(N)\, \|u\|_{H^s_p(\Omega)} ,$$

where $\gamma(N)$ is defined by

$$\gamma(N) = \exp\Big( -(b-\tilde b)\sqrt{(N-|\alpha|)^2 - (k^+)^2} \Big) + \exp\Big( -(\tilde a - a)\sqrt{|(N-|\alpha|)^2 - (k^-)^2|} \Big)$$

and $\tilde a, \tilde b$ are chosen such that $a < \tilde a < \min f_1$ and $\max f_0 < \tilde b < b$.

Summarizing we obtain the final convergence result in the TE case.

**Theorem 5.2** *Suppose that the variational problem* (2.22) *is uniquely solvable. Then there exists* $N_0 > |k^\pm| + |\alpha|$ *such that for all* $N > N_0$ *and sufficiently small* $h$ *the FE discretization of* (5.3) *has a unique solution* $u^N_h \in S_h$ *approximating the exact solution* $u \in H^2_p(\Omega)$ *with*

$$\|u - u^N_h\|_{H^1(\Omega)} \le (Ch + c\,N^{-1}\,\gamma(N))\, \|u\|_{H^2(\Omega)} ,$$
$$\|u - u^N_h\|_{L^2(\Omega)} \le (Ch^2 + c\,N^{-2}\,\gamma(N))\, \|u\|_{H^2(\Omega)} ,$$

*where the constants depend on* $k$ *but are independent on* $h$, $N$ *and* $u$. *Moreover, the Fourier coefficients* $A^\pm_{n,h}$ *of the discrete solution* $u^N_h$ *converge to the Rayleigh coefficients* (4.1) *with the rate*

$$|A^\pm_n - A^\pm_{n,h}| \le (Ch^2 + c\,N^{-2}\,\gamma(N))\, \|u\|_{H^2(\Omega)} .$$

To formulate the convergence result in the TM case we remark that the solutions of the direct and adjoint problems do not belong to $H_p^{3/2}(\Omega)$ due to the jump of the normal derivatives (cf. Corollary 3.2). But Theorem 3.4 states that the restriction of u to some subdomain $\Omega_1$, where $k$ is continuous, satisfies $u \in H^{1+\mu}(\Omega_1)$ with $\mu \in [0, \mu_0)$ (cp. also (3.25)). The improved smoothness for $\mu_0 > 1/2$ allows us to derive higher convergence rates compared with Theorem 5.1. We assume that the grating and the partitioning $\Omega_h$ are such that the discontinuities of $k$ lie on mesh lines. This is possible for example for binary gratings situated on some multilayer system. Then it is well known that for bilinear FE functions there holds the estimate

$$\inf_{\varphi_h \in S_h} \|u - \varphi_h\|_{H^s(\Omega)} \leq c_u\, h^{1+\mu-t} \ , \quad 0 \leq t \leq 1 \ ,$$

where $c_u$ is the sum of $H^{1+\mu}$-norms of $u$ on subdomains of $\Omega$. Furthermore, the estimate of $u - u^N$ can be considered analogously to the TE case, the only difference are the factors in front of $((T_\alpha^\pm - T_{\alpha,N}^\pm)u, \varphi)$. Thus under the assumptions concerning the jumps of $k$ and the partitioning $\Omega_h$ we obtain

**Theorem 5.3** *Suppose that the variational problem* (2.23) *is uniquely solvable. Then there exists* $N_0 > |k^\pm| + |\alpha|$ *such that for all* $N > N_0$ *and sufficiently small* $h$ *the FE discretization of* (5.4) *has a unique solution* $u_h^N \in S_h$ *approximating the exact solution* $u$ *with*

$$\|u - u_h^N\|_{H^1(\Omega)} \leq c_u\, (h^\mu + N^{-\mu}\, \gamma(N)) \ ,$$
$$\|u - u_h^N\|_{L^2(\Omega)} \leq c_u\, (h^{2\mu} + N^{-2\mu}\, \gamma(N)) \ ,$$

*for any* $\mu \in [0, \mu_0)$, *where* $\mu_0$ *is defined in Subsec.* 3.3. *The constants* $c_u$ *depend on* $k$ *and* $u$ *but are independent on* $h$ *and* $N$. *Moreover, the Fourier coefficients* $A_{n,h}^\pm$ *of the discrete solution* $u_h^N$ *converge to the Rayleigh coefficients* (4.1) *with the rate*

$$|A_n^\pm - A_{n,h}^\pm| \leq c_u\, (h^{2\mu} + N^{-2\mu}\, \gamma(N)) \ .$$

Finally we consider the approximation of the gradients. Recall that in the TE case

$$D_j A_n^+(\Gamma) = (2\pi)^{-1} \exp(-i\beta_n^+ b)((k_0)^2 - (k^+)^2) \int\limits_{\Sigma_j} u\, \bar{w} \ ,$$

where $u$ is the solution of the direct problem (2.22) and $w$ solves the associated adjoint problem (4.12) resp. (4.18). $\Sigma_j$ is a segment of the grating profile $\Gamma = \Lambda_0$. The approximation of $D_j A_n^+(\Gamma)$ is of course

$$D_j A_n^+(\Gamma)_h := (2\pi)^{-1} \exp(-i\beta_n^+ b)((k_0)^2 - (k^+)^2) \int\limits_{\Sigma_j} u_h \bar{w}_h \ , \tag{5.6}$$

with the corresponding FE solutions $u_h$ and $w_h$ of a truncated form $B_{TE}^N$.

From Theorem 5.2 and the inequality

$$\|u - u_h\|_{H^{-1/2+\delta}(\Sigma_j)} \leq c\, \|u - u_h\|_{H^\delta(\Omega)} \ , \quad 0 < \delta \leq 1 \ ,$$

which holds for FE solutions to elliptic second order equations (see [17]), one gets immediately the convergence rate for the gradient

$$|D_j A_n^+(\Gamma) - D_j A_n^+(\Gamma)_h| \leq (C h^{2-\delta} + c\, N^{-2+\delta}\, \gamma(N))\, \|u\|_{H^2(\Omega)} \tag{5.7}$$

for any $\delta > 0$.

In the TM case one has to estimate (compare (4.25), (4.31) and condition (3.19))

$$\int_{\Sigma_j} \left( gr(u) \cdot \overline{gr(w)} - gr(u_h) \cdot \overline{gr(w_h)} \right) .$$

Using the inequalities

$$\begin{aligned} \|u' - u_h'\|_{H^{-\mu}(\Sigma_j)} &\leq c\, \|u - u_h\|_{H^{1-\mu}(\Sigma_j)} \leq c\, \|u - u_h\|_{H^{3/2-\mu}(\Omega^+)} , \\ \|\partial(u - u_h)/\partial\nu\|_{H^{-\mu}(\Sigma_j)} &\leq c\, \|u - u_h\|_{H^{3/2-\mu}(\Omega^+)} , \end{aligned}$$

for $0 < \mu < \mu_0 - 1/2$, one obtains easily the estimate

$$\left| \int_{\Sigma_j} gr(u - u_h) \cdot \overline{gr(w)} \right| \leq c\, \|u - u_h\|_{H^{3/2-\mu}(\Omega^+)} \|w\|_{H^{3/2+\mu}(\Omega^+)} .$$

Then Theorem 5.3 and the inverse property of $S_h$ lead to the following approximation rate for the gradients in TM mode:

$$|D_j A_n^+(\Gamma) - D_j A_n^+(\Gamma)_h| \leq c_u\, (h^{2\mu} + N^{-2\mu}\, \gamma(N)) . \tag{5.8}$$

## 5.2 Generalized FEM with minimal pollution

It is well know that the accuracy of Galerkin FEM for boundary value problems governed by the Helmholtz equation deteriorates with increasing wave number and enlarging domains. This effect was observed in our problems, too. For example, the reflection and transmission efficiencies (4.2) and (4.3) do not depend on the choice of the artificial boundaries $\Gamma^\pm$. But it turned out in numerical tests that for relatively large wave numbers the computed efficiencies strongly depend on the $x_2$-dimensioning of the domain $\Omega$.

In recent years many attempts have been made in the mathematical and engineering literature to overcome the so–called pollution effect, i.e. the non-robust behavior of finite element and other domain based methods with respect to the wave number. Roughly speaking, this effect is originated by the discretization of the Helmholtz operator in the interior of the domain. Note that this effect does not arise by using boundary element methods.

In one-dimensional problems the usual piecewise linear FE solution of the equation $u" + k^2 u = 0$ on a uniform grid has the discrete wave number

$$k_h = \frac{1}{h} \arccos \frac{2(3 - (kh)^2)}{6 + (kh)^2} = k - \frac{k^3 h^2}{24} + O(k^5 h^4) .$$

It is clear that this "phase lag" leads to large discretization errors. The simplest way to improve the phase accuracy consists in solving Galerkin FE systems where $k$ is replaced by

$$k' = \frac{6(1 - \cos kh)}{h^2 (2 + \cos kh)}$$

which ensures that this modified FEM has no pollution. This approach is also effective if $k$ is piecewise constant as shown in Fig 5. Here the FE and GFE solutions for $h = 1/32$ are plotted of the equation

$$u'' + k^2\,u = 0\ ,\ u'(a) + iku(a) = 0\ ,\ u'(b) - iku(b) = 2ik\exp(-ikb) \tag{5.9}$$

with $a = -5.5$, $b = 1$, $k(x) = 5$ for $-5 < x < 0$ and otherwise $k = 1$, corresponding to the diffraction by some layer. We see that due to the phase lag the FE solution cannot approximate the exact one, whereas the GFEM solution is very accurate. Note that a rigorous error analysis for one–dimensional problems is contained in [2].

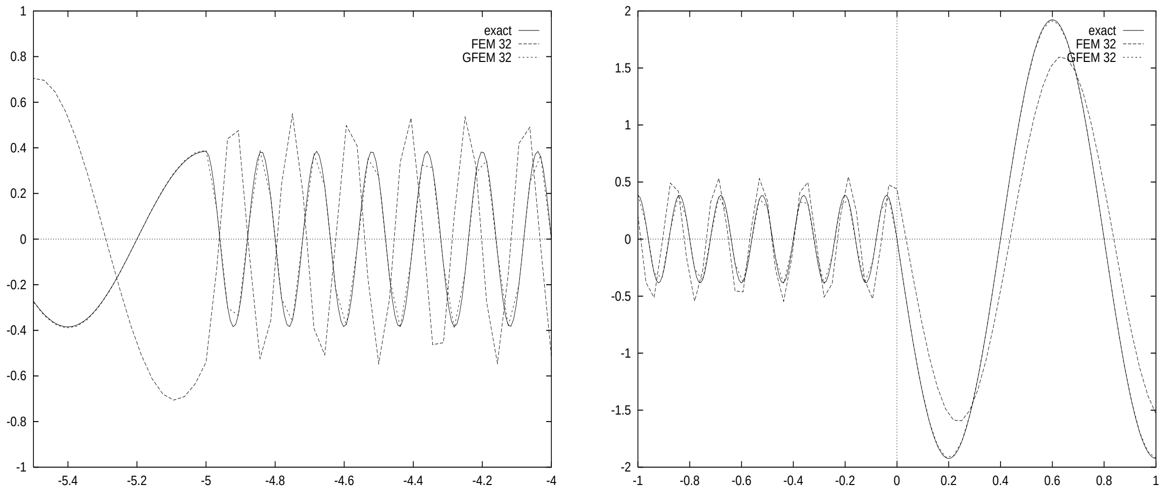


Figure 6: Imaginary part of the solution of (5.9) for $h = 1/32$

In the two-dimensional case the situation is quite different, here the modification of $k$ alone will not lead to an essentially better numerical phase accuracy. Recently in [2] and [3] a new approach has been proposed in the framework of Generalized FEM. The idea here is to replace the stencils of the FEM system matrix by some other stencils, especially adopted to the Helmholtz equation, and to connect the discrete solution with the associated bilinear FE function. The main question concerns the existence and construction of a GFEM, which ensures that the wave number of the discrete solution is as close as possible to the wave number of the analytic solution. For the case of a uniform quadratic mesh on $\Omega$, $h_1 = h_2$, it was shown in [2] and [3] that any GFEM has a phase lag, leading to a pollution effect. The authors define a measure of the approximation quality of the GFEM discretization for the Helmholtz equation and describe the influence of this measure on the error estimates. Further they construct an interior stencil leading to minimal pollution.

In the following we will give an extension of this approach to the case of rectangular meshes on $\Omega$, which also indicates how GFEM with minimal pollution for three-dimensional problems can be designed.

Consider the Helmholtz equation $\Delta u + k^2 u = 0$ in the interior of some domain $\Omega$. We want to find a discrete solution with the wavenumber $k$, i.e. at the grid points $(ph_1, qh_2)$, $p, q \in \mathbb{Z}$, it should be some linear combination of

$$v_\theta(ph_1, qh_2) = \exp(ik_1 ph_1 + ik_2 qh_2) \quad \text{with } k_1 = k\cos\theta\ ,\ k_2 = k\sin\theta\ ,\ \theta \in [0, 2\pi]\ .$$

These discrete functions are solutions of a linear system connected with the Helmholtz equation. For this system we define the interior stencil

$$\begin{pmatrix} a_3 & a_2 & a_3 \\ a_1 & a_0 & a_1 \\ a_3 & a_2 & a_3 \end{pmatrix}$$

with $a_0 > 0$, the form of which is natural for uniform rectangular grids. The application of this stencil to the discrete function $v_\theta$ results in the value

$$\exp(ik_1ph_1 + ik_2qh_2)\Big(a_0 + 2a_1\cos(k_1h_1) + 2a_2\cos(k_2h_2) + 4a_3\cos(k_1h_1)\cos(k_2h_2)\Big)$$

at the grid point $(ph_1, qh_2)$. Hence, the functions $v_\theta$ satisfy the discretized Helmholtz equation if the stencil is chosen such that

$$a_0 + 2a_1\cos(kh_1\cos\theta) + 2a_2\cos(kh_2\sin\theta) + 4a_3\cos(kh_1\cos\theta)\cos(kh_2\sin\theta) = 0$$

for all $\theta \in [0, 2\pi]$. However, there exist no solutions of this equation, i.e. for any choice of the coefficients $a_0, \ldots, a_3$ the ellipse $\mathcal{E}_{h_1h_2}(\theta) = (kh_1\cos\theta, kh_2\sin\theta)$ , $\theta \in [0, 2\pi]$, does not belong to the set of roots of the symbol function associated with the stencil

$$G(\xi_1, \xi_2) := a_0 + 2a_1\cos(\xi_1) + 2a_2\cos(\xi_2) + 4a_3\cos(\xi_1)\cos(\xi_2) \ .$$

Therefore we look for a stencil with the property that the roots of the corresponding symbol are as close as possible to $\mathcal{E}_{h_1h_2}$ as $h_1, h_2 \to 0$. Once $a_0$ is fixed, obviously the coefficients $a_1$, $a_2$ and $a_3$ will depend on $k$, $h_1$ and $h_2$, and we are interested in analytic expressions for them.

Let us denote by $\mathcal{N}_{h_1h_2}$ the set of roots of the symbol $G$ lying in some rectangle $(-kh_1-\varepsilon, kh_1+\varepsilon)\times(-kh_2-\varepsilon, kh_2+\varepsilon)$, where $\varepsilon > 0$ is chosen such that $\mathcal{N}_{h_1h_2}$ is simply connected. Using the results obtained in [3] for the case $h_1 = h_2$ one can show that the distance between $\mathcal{N}_{h_1h_2}$ and $\mathcal{E}_{h_1h_2}$ defined by

$$\mathcal{D}_{h_1h_2} = \max_{\theta\in[0,2\pi]} \ \min_{\xi\in\mathcal{N}_{h_1h_2}} \|\mathcal{E}_{h_1h_2}(\theta) - \xi\|$$

can be taken as a measure for the approximation quality of the GFEM. In particular, given some interior stencil, there exist boundary value problems for the Helmholtz equation such that the error between the exact solution and the GFEM solution can be estimated from below by

$$\|u - u_{GFE}\|^2 \geq c(h_1^2 + h_2^2)\mathcal{D}_{h_1h_2} \ .$$

(detailed proofs will be given elsewhere, compare also [2],[3]).

To find a stencil providing asymptotically the minimal $\mathcal{D}_{h_1h_2}$ we use the fact that for symbol functions $G(\xi_1, \xi_2)$ with roots in a neighbourhood of $\mathcal{E}_{h_1h_2}$ the asymptotics $\max\limits_{\theta\in[0,2\pi]} |G(kh_1\cos\theta, kh_2\sin\theta)| = O((kh)^\ell)$ and $\mathcal{D}_{h_1h_2} = O((kh)^{\ell-1})$ are equivalent. This follows immediately if $\cos(kh_1\cos\theta + r_1)$ and $\cos(kh_2\sin\theta + r_2)$ are expanded with respect to the distance parameters $r_1$ and $r_2$. Therefore we determine the coefficients $a_1$, $a_2$ and $a_3$ such that asymptotically $\max\limits_{\theta\in[0,2\pi]} |g(\theta)|$ is minimal, where $g(\theta) = G(kh_1\cos\theta, kh_2\sin\theta)$. After that we calculate the first coefficients of the asymptotics of $\mathcal{D}_{h_1h_2}$ for this stencil and compare them with the asymptotics expansion of the minimal distance $\mathcal{D}_{h_1h_2}$ obtained by using some special series representation.

The function $g(\theta)$ is $\pi$-periodic, in the case $h_1 = h_2$ it has even the period $\pi/2$. Using the formulas

$$\frac{1}{\pi}\int_0^\pi \cos(a\cos\theta)\cos(2m\theta)\,d\theta = (-1)^m J_{2m}(a)\ ,$$

$$\frac{1}{\pi}\int_0^\pi \cos(b\sin\theta)\cos(2m\theta)\,d\theta = J_{2m}(b)\ ,$$

$$\frac{1}{\pi}\int_0^\pi \cos(a\cos\theta)\cos(b\sin\theta)\cos(2m\theta)\,d\theta = J_{2m}(\sqrt{a^2+b^2})\cos\left(2m\ \arctan\frac{a}{b}\right)\ ,$$

one concludes that the coefficients of the Fourier series

$$g(\theta) = \hat{g}_0/2 + \sum_{m=1}^{\infty} \hat{g}_{2m}\cos(2m\theta)$$

have the asymptotics

$$\hat{g}_{2m} \sim \frac{(kh_1)^{2m} + (kh_2)^{2m}}{2^{2m}(2m)!} \qquad \text{for small } kh_1 \text{ and } kh_2$$

if $a_1, a_2, a_3 = O(1)$. Thus for given $a_0$ the values of $a_1$, $a_2$ and $a_3$ can be found from the condition that the first three Fourier coefficients of $g$ vanish, $\hat{g}_0 = \hat{g}_2 = \hat{g}_4 = 0$, which ensures that

$$\max_{\theta\in[0,2\pi]} |g(\theta)| = O((kh)^6)\ , \tag{5.10}$$

recall that $h = \sqrt{h_1 h_2}$. Thus one gets the linear system

$$\begin{aligned}
2\,J_0(kh_1)\,a_1 + 2\,J_0(kh_2)\,a_2 &+ 4\,J_0(k\sqrt{h_1^2+h_2^2})\,a_3 &= -a_0 \\
2\,J_2(kh_1)\,a_1 - 2\,J_2(kh_2)\,a_2 &+ 4\,\frac{h_1^2-h_2^2}{h_1^2+h_2^2}\,J_2(k\sqrt{h_1^2+h_2^2})\,a_3 &= 0 \\
2\,J_4(kh_1)\,a_1 + 2\,J_4(kh_2)\,a_2 &+ \left(4 - \frac{32\,h_1^2\,h_2^2}{(h_1^2+h_2^2)^2}\right)J_4(k\sqrt{h_1^2+h_2^2})\,a_3 &= 0
\end{aligned} \tag{5.11}$$

Note that in the case $h_1 = h_2$ the solution of (5.11) gives $a_1 = a_2$ and there holds

$$\max_{\theta\in[0,2\pi]} |g(\theta)| \le O((kh)^8)\ .$$

Since any other choice of $a_1$, $a_2$ and $a_3$ leads to an asymptotics not better than (5.10), the stencil whose symbol has asymptotically the smallest absolute value on the ellipse $\mathcal{E}_{h_1h_2}$ is uniquely determined. It remains to show the existence of the set of zeroes $\mathcal{N}_{h_1h_2}$ near the ellipse and to estimate $\mathcal{D}_{h_1h_2}$. Similar to the technique of [3] for the case $h_1 = h_2$ one can use an expansion of the zeroes of $G$ in the form

$$\begin{aligned}
\xi_1 &= kh_1\left(1 + \textstyle\sum_{m=1}^{\infty} r_m(\theta, h_1/h_2)(kh)^{2m+1}\right)\cos\theta\ , \\
\xi_2 &= kh_2\left(1 + \textstyle\sum_{m=1}^{\infty} r_m(\theta, h_1/h_2)(kh)^{2m+1}\right)\sin\theta\ ,
\end{aligned} \tag{5.12}$$

which gives for the stencil defined by (5.11)

$$
\begin{aligned}
r_2(\theta,q) &= 0\ ,\\
r_3(\theta,q) &= \frac{(q^2-q^{-2})\,\cos 6\theta}{15360}\ ,\\
r_4(\theta,q) &= \frac{(q^3+q^{-3})\,\cos 8\theta}{1548288}+\frac{(q^3-q^{-3})\,\cos 6\theta}{193536}+\frac{(q-q^{-1})\,(q^2-q^{-2})\,\cos 4\theta}{737280}\ .
\end{aligned}
$$

It is clear that any perturbation of the solution $a_1$, $a_2$, $a_3$ of (5.11) with terms of the order $O((kh)^6)$ determines another stencil with $\max|g(\theta)| = O((kh)^6)$. Since

$$
\mathcal{D}_{h_1h_2} \le \max_{\theta\in[0,2\pi]} \left| \sum_{m=1}^{\infty} r_m(\theta,q)(kh)^{2m+1} \right| ,
$$

the stencil with minimal $\mathcal{D}_{h_1h_2}$ can be determined from the condition that the function values $|r_m(\theta,q)|$ are minimal. Proceeding similarly to [3] expand the elements of the unknown stencil into power series in $(kh)^2$ with coefficients depending on $h_1/h_2$ and take the Taylor expansion of

$$
\begin{aligned}
&\cos\left(kh_1\left(1+\sum_{m=1}^{\infty} r_m(\theta,h_1/h_2)(kh)^{2m+1}\right)\cos\theta\right)\ ,\\
&\cos\left(kh_2\left(1+\sum_{m=1}^{\infty} r_m(\theta,h_1/h_2)(kh)^{2m+1}\right)\sin\theta\right)\ .
\end{aligned}
$$

So one gets an infinite series in $(kh)^2$, which roots can be determined from a recursion formula connecting $r_m$ with all $r_j$, $j = 1,\ldots,m-1$, and the power series coefficients of the stencil elements. The condition of minimal $\max|r_m(\theta,q)|$ leads to a unique solution and it turns out, that for this stencil the roots of the corresponding symbol function has the asymptotics (5.12) with

$$
\begin{aligned}
r_3(\theta,q) &= \frac{(q^2-q^{-2})\,\cos 6\theta}{15360}\ ,\\
r_4(\theta,q) &= \frac{(q^3+q^{-3})\,\cos 8\theta}{1548288}+\frac{(q^3-q^{-3})\,\cos 6\theta}{193536}\ .
\end{aligned}
$$

Hence the analytically given stencil (5.11) can be considered as almost optimal and one obtains the estimate

$$
\mathcal{D}_{h_1h_2} \le \frac{1}{15360}\,k^5\,|h_1^2-h_2^2|\,(h_1^2+h_2^2)^{3/2} + O((kh)^7)\ .
$$

As mentioned above, if $h_1 = h_2$ then $\mathcal{D}_h$ has the asymptotics $O((kh)^7)$. Note that in [2] another analytical formula for an optimal stencil was given. It is interesting that the asymptotics of $\mathcal{D}_h$ for both stencils differs only beginning with the term $(kh)^{11}$ and has the form

$$
\mathcal{D}_h \le \frac{1}{774144}(kh)^7 + \frac{1}{55296000}(kh)^9 + O((kh)^{11})\ .
$$

## 5.3 Implementation

Here we briefly describe how the optimal interior stencil for solving 2d Helmholtz equations on rectangular meshes can be adapted for solving the direct and adjoint variational TE and TM problems for binary gratings on top of some multilayer system. For those cases the domain $\Omega$ can be partitioned such that the rectangular mesh is uniform in $x_1$- and piecewise uniform in $x_2$-direction such that the discontinuities of $k$ lie on mesh lines. Since our problems contain the differential operator $\Delta + 2i\alpha\partial_{x_1} + (k^2 - \alpha^2)$ the optimal stencil has to be modified. For a solution $u$ of the TE or TM problem the function $\exp(i\alpha x_1)\, u$ solves the Helmholtz equation $\Delta + k^2$. Therefore we expect the discrete solutions to be combinations of the discrete functions

$$v_\theta(ph_1, qh_2) = \exp(i(k_1+\alpha)ph_1 + ik_2qh_2) \quad \text{with } k_1 = k\cos\theta\,,\ k_2 = k\sin\theta$$

and we implemented a GFEM with scaled versions of the stencil

$$\begin{pmatrix} \exp(-i\alpha h_1)\, a_3 & a_2 & \exp(i\alpha h_1)\, a_3 \\ \exp(-i\alpha h_1)\, a_1 & a_0 & \exp(i\alpha h_1)\, a_1 \\ \exp(-i\alpha h_1)\, a_3 & a_2 & \exp(i\alpha h_1)\, a_3 \end{pmatrix}$$

where the coefficients are the solutions of (5.11). The scaling is necessary due to the jumps of $k$ and to the nonlocal boundary operators $T_\alpha^\pm$. The best results were obtained if the scaling is chosen such that the sum of the central row equals the diagonal element of the GFEM with no pollution for the one-dimensional operator $(d/dx)^2 + (k^2 - \alpha^2)$.

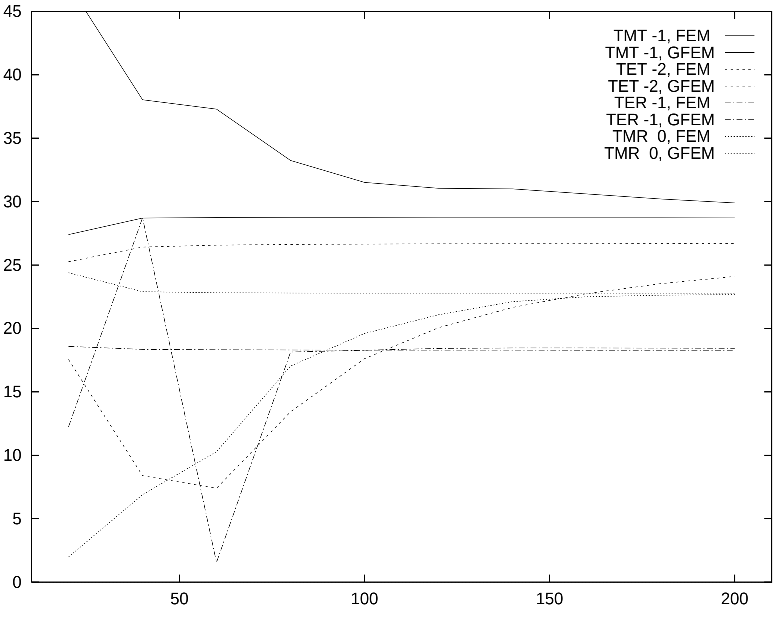


Figure 7: Comparison of some efficiencies computed with FEM and GFEM for a simple binary grating versus the square root $n$ of total grid points.

The method described above was used to evaluate the reflection and transmission efficiencies of binary gratings of different geometries and materials. In any case the method was robust and reliable, for rather poor discretizations the obtained results were excellent compared with the usual FEM. In Figure 6 we compare the numerical values of some reflection and transmission efficiencies versus the square root $n$ of total number of grid points computed with the usual FEM and the GFEM on quadratic meshes for a simple binary grating with the optical index $k = 2.5$ situated on some other material with $k^- = 3.5$. In any case the GFEM results differ already for $n = 40$ only by 2 % from the corresponding values for $n = 200$, whereas the FEM results converge rather slowly to these values. Furthermore, for special binary gratings which can be treated also with other methods (e.g. integral equation or waveguide methods) the performance of our method is of the same or even better quality.

The GFEM for solving direct and adjoint problems was integrated into a computer program for the study of optimal design problems. By using the standard algorithm of gradient descent local minima of functionals are determined, which characterize desired optical properties of binary gratings. These functionals involve the Rayleigh coefficients of the discrete models on a given partition of the domain $\Omega$ for a prescribed range of incidence angles or wavelengths. Of course, the gradients are computed by discretized versions of the formulas given in Sec. 4. Corresponding to the gradients the shape of $\Gamma$ is varied within a class of admissible profiles, which are restricted to the mesh lines and certain technological constraints.

The computer program was used to find the optimal design of large classes of binary gratings for different functionals. As one example we mention the application of metallic subwavelength gratings for polarization devices as considered in [25]. Fig. 7 shows the results for the optimal design of such a zero order grating that should maximize the reflection of TE polarisation and the transmission of TM polarisation over the range of wavelengths from 450 to 633 nm. The grating period is 200 nm, the width of the bar amounts to 60 nm and the height is 150 nm.

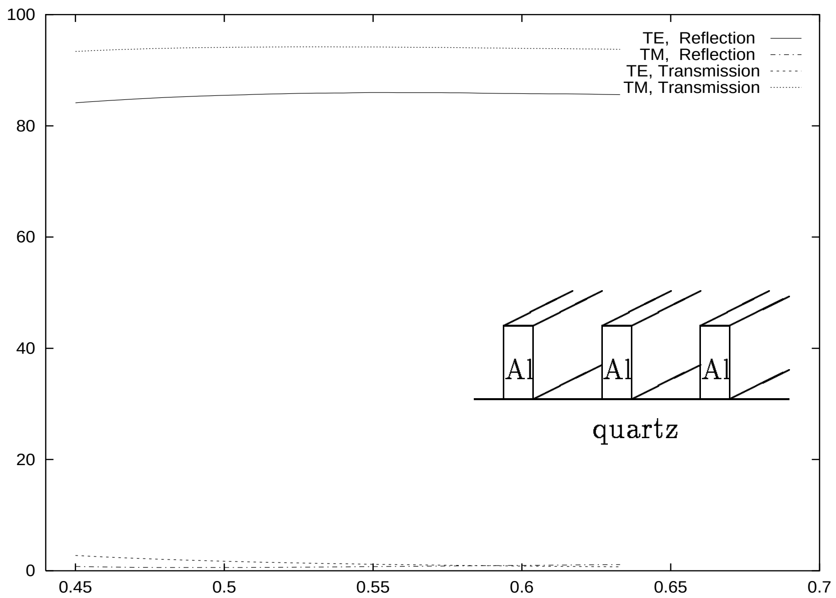


Figure 8: Optimal design for a simple polarisation grating

Certainly better minimization algorithms exist, for example conjugate gradient methods or methods based on higher order derivative information. The design and analysis of different minimization methods for coated binary gratings will be the topic of future research.

# References


[1] T. Abboud, Étude mathématique et numérique de quelques problèmes de diffraction d' ondes électromagnétiques, PhD dissertation, Ecole Polytechnique, Palaiseau, 1991.

[2] I. Babuška, F. Ihlenburg, E. Paik, S. Sauter, A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution, Comp. Meth. Appl. Mech. Eng. **128**, 325–359 (1995).

[3] I. Babuška, S. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers, Technical Report BN-1172, IPST, Univ. of Maryland, 1994.

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

[5] G. Bao, Numerical analysis of diffraction by periodic structures: TM polarization, Num. Math. **75**, 1–16 (1996).

[6] G. Bao, D.C. Dobson, J.A. Cox, Mathematical studies in rigorous grating theory, J. Opt. Soc. Amer. **A 12**, 1029–1042 (1995).

[7] O.P. Bruno, F. Reitich, Solution of a boundary value problem for the Helmholtz equation via variation of the boundary into the complex domain, Proc. Roy. Soc. Edinburgh **A 122**, 317–340 (1992).

[8] O.P. Bruno, F. Reitich, Numerical solution of diffraction problems: a method of variation of boundaries, J. Opt. Soc. Amer. **A 10**, 1168–1175 (1993).

[9] X. Chen, A. Friedman, Maxwell's equation in a periodic structure, Trans. Amer. Math. Soc. **323**, 465–507 (1991).

[10] M. Costabel, A symmetric method for the coupling of finite elements and boundary elements, in *The mathematics of finite elements and applications* VI, J.R. Whiteman (ed.), 281-288, Academic Press, London, 1988.

[11] M. Costabel, E. Stephan, A direct boundary integral equation method for transmission problems, J. Math. Anal. Appl. **106**, 367–413 (1985).

[12] D.C. Dobson, Optimal design of periodic antireflective structures for the Helmholtz equation, European J. Appl. Math. **4**, 321–340 (1993).

[13] D.C. Dobson, A variational method for electromagnetic diffraction in biperiodic structures, Model. Math. Anal. Num. **28**, 419–439 (1994).

[14] J. Elschner, *Singular ordinary differential operators and pseudodifferential equations*, Akademie–Verlag, Berlin, 1985.

[15] A. Friedman, *Mathematics in Industrial Problems, Part* 7, IMA Volume **67**, Springer–Verlag, New York, 1995.

[16] I.C. Gohberg, M.G. Krein, *Introduction to the theory of linear nonselfadjoint operators*, Transl. Math. Monographs Vol. **18**, AMS, Providence, 1969.

[17] B.N. Khoromskij, G. Schmidt, A fast interface solver for the biharmonic Dirichlet problem on polygonal domains, Num. Math., to appear.

[18] A. Kirsch, Diffraction by periodic structures, in *Proceedings of the Lapland Conference on Inverse Problems*, L. Pävarinta, E. Somersalo (eds.), 87–102, Springer-Verlag, Berlin, 1993.

[19] V.A. Kondratiev, Boundary problems for elliptic equations in domains with conical or angular points, Trans. Moscow Math. Soc. **16**, 227–313 (1967).

[20] K. Lemrabet, Régularité de la solution d' un problème de transmission, J. Math. pures et appl. **56**, 1–38 (1977).

[21] D. Maystre, Integral methods, in [22]

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

[23] A. Pomp, The integral method for coated gratings: computational cost, J. Mod. Optics **38**, 109–120 (1991).

[24] S. Prößdorf, B. Silbermann, *Numerical analysis for integral and related operator equations*, Akademie–Verlag, Berlin, 1991.

[25] B. Schnabel, E.-B. Kley, Fabrication and application of subwavelength gratings, Proc. SPIE **3008** (1997).