JOURNAL OF COMPUTATIONAL PHYSICS **146,** 603–626 (1998)
ARTICLE NO. CP986071

Numerical Solution of Optimal Design Problems for Binary Gratings

J. Elschner and G. Schmidt

Weierstrass Institute for Applied Analysis and Stochastics (WIAS), D-10117 Berlin, Mohrenstrasse 39, Germany
E-mail: elschner@wias-berlin.de, schmidt@wias-berlin.de

Received March 25, 1998; revised July 30, 1998

In this paper we describe recent developments in the application of mathematical and computational techniques to the problem of designing binary gratings on top of a multilayer stack in such a way that the propagating modes have a specified intensity or phase pattern for a chosen range of wavelengths or incidence angles. The diffraction problems are transformed to 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 obtain analytic formulae for the gradients of cost functionals with respect to the parameters of the grating profile and the thickness of the layers, so that the optimal design problems can be solved by minimization algorithms based on gradient descent. For the computation of diffraction efficiencies and gradients the variational problems are solved by using a generalized finite element method with minimal pollution. We provide some numerical examples to demonstrate the convergence properties for evaluating diffraction efficiencies and gradients. The method is applied to optimal design problems for polarisation gratings and beam splitters. © 1998 Academic Press

Key Words: diffraction by periodic structures; Helmholtz equation; transmission problems; nonlocal boundary conditions; optimal design; gradient formulae; generalized FEM with minimal pollution.

1. INTRODUCTION

The practical application of diffractive optics technology has driven the need for mathematical models and numerical codes both to solve 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.

Within the so-called rigorous grating theory, which is based on Maxwell's equations, periodic gratings can be modeled as quasi-periodic transmission problems for the Helmholtz equation in the whole plane. Special difficulties are associated with the numerical solution of these problems due to the highly oscillatory nature of waves and interfaces. Various methods have been proposed to evaluate the solution for a given structure, i.e., to solve the direct diffraction problem. Among the most well known are modal expansion, differential and integral methods (cf. the classical monograph [1] and recent extensions and improvements in, e.g. [2–7]). These methods turned out to be efficient for solving the direct diffraction problem for certain classes of grating structures, but it is difficult to find any mathematical treatment of convergence in the literature. Such a convergence analysis can be found in the case of smooth interfaces between different materials for integral equation methods and the analytical continuation method introduced in [8]. In the case of binary structures, whose surface profile is given by a piecewise constant function, the mathematical complexities are amplified by singularities of the solutions. Recently, a new variational approach was proposed (see [9–10] and the references therein), which appears to be well adapted for the analytical and numerical treatment of very general diffraction structures as well as complex materials and allows straightforward extensions to diffraction problems for conical mounting and crossed gratings [11–12]. In particular, this approach is the basis for the convergence analysis of finite element solution methods provided in [13–15]. We note that in [16] a similar method was used for photolithography simulation on nonplanar substrate.

But it is more important that the variational approach leads to effective formulae for the gradient of cost functionals arising in optimal design problems as shown in [17, 15], such that gradient-based minimization methods can be used to find gratings with specified optical functions. There have been a number of papers from the engineering community that are concerned with the optimal design of periodic gratings. By far the greatest activity has been in optimization for ray-tracing and phase-reconstruction techniques which are valid within the domain of Fourier optics. A few of these papers (e.g. [18–20]) are devoted to optimization problems using rigorous diffraction theory. However, the optimization procedures used there are based only on the values of certain cost functionals; i.e., they require the solution of a large number of direct problems and are therefore computationally expensive. More advanced methods to find optimal solutions utilize, besides the values of cost functionals, also its gradients or even properties of higher order differentials. The simplest example are descent-type algorithms, which are computationally effective if explicit gradient formulae are available. But so far gradient formulas were obtained only for the TE case; see [17], where interface mixture problems have been studied. Sometimes the approximation of gradients by simple difference quotients is used, which is, however, very inefficient for a large number of parameters.

In the present paper we apply some mathematical results from [15] to the model problem of designing binary gratings on top of a multilayer stack in such a way that the propagating modes have a specified intensity or phase pattern for a chosen range of wavelengths or incidence angles. First we present the variational formulations of the diffraction problems for TE and TM polarization and give a summary of some existence and uniqueness results. In Section 3 we consider a typical optimal design problem, formulate the cost functional and write down the formulae for the gradients with respect to the parameters of the grating profile and the thicknesses of the layers. Then the optimal design problem can be solved by minimization algorithms based on gradient descent. For the computation of diffraction efficiencies and of the gradients we use a reliable numerical method which originates from the variational formulations. This method, which combines a generalized finite element method in the grating structure with Fourier expansions in the multilayer system, is discussed in Section 4. It is known that the accuracy of the usual Galerkin FEM for the Helmholtz equation deteriorates for large wave numbers. More precisely, the ratio n_{FEM}/n_o goes to infinity with increasing wave number, where n_o is the dimension of the finite element space required to achieve a prescribed accuracy and the Galerkin–FE solution needs the dimension n_{FEM} to get the same accuracy. This nonrobust behavior of the FEM with respect to the wavenumber is called the *pollution effect*. The goal of the generalized FEM (GFEM) is to modify the entries of the FEM system matrix in such a way that the ratio n_{FEM}/n_o increases as slowly as possible. Our construction of the GFEM with minimal pollution extends a recent approach of [21] to problems with piecewise constant wave numbers on rectangular partitions and leads to essentially better numerical results than usual FEM. Finally we provide some numerical examples to demonstrate the convergence properties of this method for evaluating diffraction efficiencies and gradients. The last section includes also several examples of optimal design problems including polarisation gratings and beam splitters.

2. VARIATIONAL FORMULATION OF THE DIRECT SCATTERING PROBLEM

Consider a binary grating of period d, with height H and transition points t_j at the top of a stack of layers of thicknesses h_j. The materials are nonmagnetic with the permeability μ_0 and have the dielectric constants ϵ. The coordinate system is chosen such that the diffraction problem is invariant in the x_3 direction and that the x_1 axis is parallel to the layers. Thus the problem is determined by the function $\epsilon(x_1, x_2)$ which is d-periodic in x_1. We assume that the material above the grating profile Γ is homogeneous with $\epsilon = \epsilon^+ > 0$. Below Γ the material may be inhomogeneous and we assume that the function $\epsilon = \epsilon^-$ is piecewise constant corresponding to the different layers and constant for the substrate. Further, we suppose that the ϵ^- can be complex valued with $\operatorname{Im} \epsilon^- \geq 0$ and $\operatorname{Re} \epsilon^- > 0$ if $\operatorname{Im} \epsilon^- = 0$.

Assume that an incoming plane wave with time dependence $\exp(-i\omega t)$ is incident in the (x_1, x_2)-plane upon the grating from the top with the angle of incidence $\theta \in (-\pi/2, \pi/2)$. Then the electromagnetic field does not depend on x_3. In either case of polarization, one of the fields $\mathbf{E}$ or $\mathbf{H}$ remains parallel to the grooves and is therefore determined by a single scalar quantity $v = v(x_1, x_2)$ (equal to the transverse component of $\mathbf{E}$ in the TE case and to the transverse component of $\mathbf{H}$ in the TM case). The function v satisfies two-dimensional Helmholtz equations

$$\Delta v + \omega^2 \mu_0 \epsilon v = 0 \tag{2.1}$$

in the regions with constant permittivity, together with the usual outgoing wave condition at infinity. At the material interfaces the solutions are subjected to well-known transmission conditions. For TE polarisation the solution and its normal derivative $\partial_n v$ have to cross the interface continuously, whereas in TM polarisation the product $\epsilon^{-1} \partial_n v$ has to be continuous (for more details cf. the classical monograph [1]).

The diffraction problems admit variational formulations in a bounded periodic cell which were introduced in [9, 10]. In the following k denotes the piecewise constant function $\omega(\mu_0 \epsilon)^{1/2}$. Above the profile Γ it takes the constant value $k^+ = \omega(\mu_0 \epsilon^+)^{1/2}$, whereas below Γ it coincides with the piecewise constant function $k^- = \omega(\mu_0 \epsilon^-)^{1/2}$ (k_g, k_1, k_2, and k_3 in

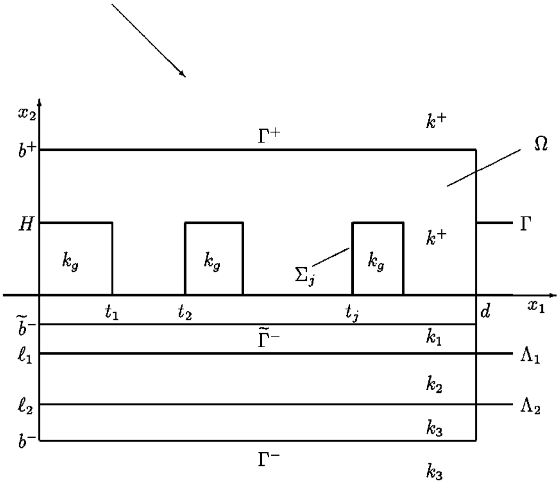

FIG. 2.1. Problem geometry.

Fig. 2.1). We assume that

$$\operatorname{Re} k^+ > 0, \quad \operatorname{Re} k^- > 0, \quad \operatorname{Im} k^- \geq 0. \tag{2.2}$$

Note that $k = \omega\nu(\mu_0\epsilon_0)^{1/2}$, where ϵ_0 is the permittivity of the vacuum and $\nu = (\epsilon/\epsilon_0)^{1/2}$ denotes the optical index. The incoming wave has the form $v^i = \exp(i\alpha x_1 - i\beta x_2)$, where $\alpha = k^+ \sin\theta$, $\beta = k^+ \cos\theta$. If we introduce two artificial boundaries $\Gamma^\pm = \{x_2 = b^\pm\}$ lying above Γ and below the layer structure, respectively, denote by Ω the rectangle $(0, d) \times (b^-, b^+)$ and define the d-periodic function $u = v\exp(-i\alpha x_1)$, then the diffraction problem for TE polarization can be transformed to a variational problem for u in the rectangle Ω. Multiplying the differential equation (2.1) by some smooth function, applying Green's formula, and taking into account the transmission conditions at the material interfaces and the outgoing wave condition on $\Gamma^\pm$, it can be shown (cf. [10, 15]) that the diffraction problem for TE polarization is equivalent to the variational equation

$$\begin{aligned} B_{TE}(u,\varphi) &:= \int_\Omega \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - \int_\Omega k^2 u\overline{\varphi} + \int_{\Gamma^+} (T_\alpha^+ u)\overline{\varphi} + \int_{\Gamma^-} (T_\alpha^- u)\overline{\varphi} \\ &= -\int_{\Gamma^+} 2i\beta \exp(-i\beta b^+)\overline{\varphi}, \quad \forall\varphi, \end{aligned} \tag{2.3}$$

where $\nabla_\alpha = (\partial_{x_1,\alpha}, \partial_{x_2}) := \nabla + i(\alpha, 0)$. The functions $T_\alpha^\pm u$ are defined on $\Gamma^\pm$ as

$$\left(T_\alpha^\pm u\right)\left(x_1, b^\pm\right) := -\sum_{n=-\infty}^{\infty} i\beta_n^\pm \hat{u}_n^\pm \exp(inKx_1), \tag{2.4}$$

where $K = 2\pi/d$ and $\hat{u}_n^\pm$ denote the Fourier coefficients of $u(x_1, b^\pm)$:

$$\hat{u}_n^\pm = \frac{1}{d}\int_0^d u\left(x_1, b^\pm\right) \exp(-inKx_1)\, dx_1.$$

The numbers $\beta_n^{\pm}$ are defined as

$$\beta_n^{\pm} = \beta_n^{\pm}(\alpha) := \left((k^{\pm})^2 - \alpha_n^2\right)^{1/2}, \quad 0 \le \arg \beta_n^{\pm} < \pi,$$

where as usual $\alpha_n = \alpha + nK$ and $k^- = k^-(x_1, b^-)$.

The variational equation (2.3) should be satisfied for all test functions $\varphi \in H_p^1(\Omega)$, that is the function space of all complex-valued functions φ which are d-periodic in x_1 and, together with their first-order partial derivatives, square integrable in Ω; see [22] for the variational approach to classical elliptic boundary value problems.

The variational formulation (2.3) is very useful, because the transmission and outgoing wave conditions are enforced implicitly and it allows to seek the solution in the function space $H_p^1(\Omega)$, which is natural for second-order partial differential equations on nonsmooth domains. Here one can apply well established methods for the analysis and numerical solution of the diffraction problems.

Note that any solution of (2.3) satisfies on $\Gamma^{\pm}$ the boundary conditions

$$\partial_n u|_{\Gamma^+} + T_\alpha^+ u|_{\Gamma^+} = -2i\beta \exp(-i\beta b^+), \quad \partial_n u|_{\Gamma^-} + T_\alpha^- u|_{\Gamma^-} = 0. \tag{2.5}$$

which implies the Fourier series expansion

$$\begin{aligned} u(x_1, b^+) &= \sum_{n=-\infty}^{\infty} A_n^+ \exp(i\beta_n^+ b^+) \exp(inKx_1) + \exp(-i\beta b^+), \\ u(x_1, b^-) &= \sum_{n=-\infty}^{\infty} A_n^- \exp(-i\beta_n^- b^-) \exp(inKx_1). \end{aligned} \tag{2.6}$$

Thus, the operators $T_\alpha^{\pm}$ are the Dirichlet-to-Neumann mappings,

$$\partial_n u^{\pm}|_{\Gamma^{\pm}} = -T_\alpha^{\pm} u^{\pm}|_{\Gamma^{\pm}}, \tag{2.7}$$

for functions of the form

$$u(x_1, x_2) = \sum_{n=-\infty}^{\infty} A_n^{\pm} \exp\left(\pm i\beta_n^{\pm} x_2\right) \exp(inKx_1), \quad x_2 \gtrless b^{\pm}.$$

Similarly, the TM diffraction problem can be formulated as follows (cf. [10], [15]):

$$\begin{aligned} B_{TM}(u, \varphi) :=& \int_\Omega \frac{1}{k^2} \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - \int_\Omega u\overline{\varphi} + \int_{\Gamma^+} \frac{1}{(k^+)^2} (T_\alpha^+ u)\overline{\varphi} + \int_{\Gamma^-} \frac{1}{(k^-)^2} (T_\alpha^- u)\overline{\varphi} \\ =& -\int_{\Gamma^+} \frac{2i\beta}{(k^+)^2} \exp(-i\beta b^+)\overline{\varphi} \quad \forall \varphi \in H_p^1(\Omega). \end{aligned} \tag{2.8}$$

THEOREM 1 [15]. *Suppose that k satisfies condition* (2.2). *Then the sesquilinear forms B_{TE} and B_{TM} are strongly elliptic over $H_p^1(\Omega)$.*

Here we call a bounded sesquilinear form $a(\cdot, \cdot)$, given on some Hilbert space X, strongly elliptic if there exist a complex number ϕ, $|\phi| = 1$, a constant $c > 0$, and a compact form $q(\cdot, \cdot)$ such that

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

Based on the variational formulation of the diffraction problems and Theorem 1 the following existence and uniqueness results can be established ([15], [9]):

1. The TE and TM diffraction problems admit solutions $u \in H_p^1(\Omega)$ for all $\omega > 0$ and θ. These solutions are unique for all but a sequence of countable frequencies ω_j, $\omega_j \to \infty$.

2. For TE polarisation the solution $u(x_1, x_2) = E_{x_3}(x_1, x_2)\exp(-i\alpha x_1)$ has square integrable second-order partial derivatives, $u \in H_p^2(\Omega)$.

3. In the TM case the solution $u(x_1, x_2) = H_{x_3}(x_1, x_2)\exp(-i\alpha x_1)$ may have singularities at the corner points $(t_j, 0)$ and (t_j, H) of the grating. More precisely, near corner points there holds $u = r^\lambda f + g$, where r denotes the distance to the corner point, the exponent λ with $0 < \operatorname{Re}\lambda < 1$ is determined by the optical index of the grating material and f, g are some smoother functions. In particular, if two materials with optical indices ν_1 and ν_2, respectively, meet at some corner then λ is the solution with minimal positive real part of the equation

$$\left(\frac{\sin(\pi\lambda/2)}{\sin(\pi\lambda)}\right)^2 = \left(\frac{\nu_1^2 + \nu_2^2}{\nu_1^2 - \nu_2^2}\right)^2 .$$

Hence, the partial derivatives of u are only of the form $r^{\lambda-1} f + g$, i.e. the electric field components E_{x_1} and E_{x_2} are strongly singular and the normal derivative of u on Γ does not satisfy the Meixner condition, i.e. $\partial_n u \notin L^2(\Gamma)$, in general.

4. Introduce the set of exceptional values (Rayleigh frequencies):

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

If for $(\omega_0, \theta_0) \notin \mathcal{R}(\epsilon)$ one of the diffraction problems is uniquely solvable, then the solution u depends analytically on ω and θ in a neighbourhood of this point.

5. If one of the materials below Γ is absorbing then the TE problem has a unique solution for all frequencies $\omega > 0$.

6. If one of the layer materials is absorbing then the TM problem has a unique solution for all frequencies $\omega > 0$.

7. Let $\epsilon(x) > 0$ for $x \in \Omega$. Suppose that there exists $\tau \in \mathbb{R}$ such that

$$\left((x_2 + \tau)\frac{\partial\epsilon}{\partial x_2}, v\right)_{L^2(\Omega)} \geq 0 \quad \text{for all } v \geq 0.$$

Then the TE diffraction problem is uniquely solvable for $\omega > 0$. (This condition is always satisfied if only two materials are present.)

Note that the variational formulation of the diffraction problems and the validity of the corresponding mathematical results are not restricted to binary or other rectangular grating profiles. They remain valid for general piecewise constant functions k satisfying condition (2.2); hence the presented approach is applicable to rather complex grating structures. But here we focus on the case of binary profiles for which optimal design problems will be considered.

Define the finite sets of indices $P^\pm = \{r \in \mathbf{Z} : \beta_r^\pm \in \mathbb{R}\}$. Then the Rayleigh amplitudes A_r^+ $(r \in P^+)$ (resp. A_r^- $(r \in P^-)$), which are called the reflection (resp. transmission) coefficients, correspond to the propagating modes of u. Note that $P^- = \emptyset$ if $\operatorname{Im} k^-(x_1, b^-) \neq 0$.

The reflection and transmission coefficients A_r^+ $(r \in P^+)$ (resp. A_r^- $(r \in P^-)$), which correspond to the propagating modes of u, are determined by the Fourier coefficients of u on the artificial boundaries $\Gamma^\pm$:

$$A_0^+ = -\exp(-2i\beta b^+) + \exp(-i\beta b^+)\hat{u}_0^+, \quad A_r^\pm = \exp\left(\mp i\beta_r^\pm b^\pm\right)\hat{u}_r^\pm, \quad r \in P^+\backslash\{0\}, r \in P^-. \tag{2.9}$$

The reflected and transmitted efficiencies are defined by

$$e_r^{TE,\pm} = \left(\beta_r^\pm/\beta\right)\left|A_r^\pm\right|^2, \quad e_r^{TM,+} = (\beta_r^+/\beta)|A_r^+|^2, \quad e_r^{TM,-} = (k^+/k^-)^2(\beta_r^-/\beta)|A_r^-|^2. \tag{2.10}$$

3. AN OPTIMAL DESIGN PROBLEM

A typical minimization problem occurring in the optimal design of binary gratings on some multilayer system is the following. Assume that the period of the grating and the number of transition points and of thin-film layers are fixed. For given numbers $c_r^{TE,\pm}, c_r^{TM,\pm} \in \{-1, 0, 1\}$, define the functional

$$J(\Xi) = \sum_{r\in P^+}\left(c_r^{TE,+}e_r^{TE,+} + c_r^{TM,+}e_r^{TM,+}\right) + \sum_{r\in P^-}\left(c_r^{TE,-}e_r^{TE,-} + c_r^{TM,-}e_r^{TM,-}\right). \tag{3.1}$$

Note that the efficiencies $e_r^\pm$ are functions of the grating profile Γ and the layer interfaces Λ_j. If we fix one transition point t_0 at the origin the efficiencies $e_r^\pm$ are therefore functions of $t_1, \ldots, t_{m-1}, H, \ell_1, \ldots \ell_p$ (cf. Fig. 2.1). Now the minimization problem reads as follows:

Find transition points $t_1^0, \ldots, t_{m-1}^0$ and the height H^0 of the binary grating profile Γ^0, as well as thicknesses of the layer structure, such that

$$\min_{(t_1,\ldots,t_{m-1},H,\ell_1,\ldots\ell_p)\in K} J(\Xi) = J(\Xi^0), \tag{3.2}$$

where K is some compact set in the parameter space $\mathbb{R}^{m+p}$ reflecting, e.g., natural constraints on the design of the grating and the thin-film layers. Note that the choice $c_r^\pm = -1$ (resp. $c_r^\pm = 1$) in (3.1) amounts to maximizing (resp. minimizing) the efficiency of the corresponding reflected or transmitted propagating mode of order r.

Other minimization problems:

1. If one wants to obtain prescribed values for certain reflection and transmission efficiencies, given by the index sets $I^+ \subset P^+$ and $I^- \subset P^-$, the following smooth functional can be useful:

$$\sum_{r\in I^+}\left(\left|e_r^{TE,+} - c_r^{TE,+}\right|^2 + \left|e_r^{TM,+} - c_r^{TM,+}\right|^2\right) + \sum_{r\in I^-}\left(\left|e_r^{TE,-} - c_r^{TE,-}\right|^2 + \left|e_r^{TM,-} - c_r^{TM,-}\right|^2\right) \to \min.$$

2. The optimal design of a grating providing a given phase shift φ between the rth reflected TE and TM mode can be performed using the functional

$$-e_r^{TE,+} - e_r^{TM,+} + \left|A_r^{TE,+} - \exp(i\varphi)A_r^{TM,+}\right|^2 \to \min. \tag{3.3}$$

Obviously many other functionals are possible, especially if a corresponding optimization over a range of wavelengths or incidence angles is required.

To find local minima of these functionals, the method of gradient descent or other gradient-type methods can be applied. Thus we must calculate the gradient of $J(\Xi)$, for example, which can be easily expressed in terms of the partial derivatives $D_j A_r^{\pm}(\Gamma)$ (with respect to the transition points $t_1, \ldots, t_m - 1$, the height H, and the layer thicknesses, given by the coordinates ℓ_j) of the reflection and transmission coefficients in both the TE and TM case. Here we propose to use gradient formulae based on the solution of the direct problem and its adjoint, instead of simple difference quotients which are very expensive to compute for a large number of parameters.

The gradient of $J(\Xi)$ is given by

$$\begin{aligned} D_j J(\Xi) = \sum_{r\in P^+} & 2(\beta_r^+/\beta)\big\{c_r^{TE,+}\operatorname{Re}\big(\overline{A_r^{TE,+}(\Xi)}D_j A_r^{TE,+}(\Xi)\big) \\ & + c_r^{TM,+}\operatorname{Re}\big(\overline{A_r^{TM,+}(\Xi)}D_j A_r^{TM,+}(\Xi)\big)\big\} \\ & + \sum_{r\in P^-} 2(\beta_r^-/\beta)\big\{c_r^{TE,-}\operatorname{Re}\big(\overline{A_r^{TE,-}(\Xi)}D_j A_r^{TE,-}(\Xi)\big) \\ & + (k^+/k^-)^2 c_r^{TM,-}\operatorname{Re}\big(\overline{A_r^{TM,-}(\Xi)}D_j A_r^{TM,-}(\Xi)\big)\big\}. \end{aligned} \tag{3.4}$$

Once one has derived explicit formulae for those partial derivatives, it is possible to compute also the gradients for a much more general class of functionals involving the Rayleigh coefficients for a given range of incidence angles or wavelengths.

The formulae for all components of the gradient of $A_r^{\pm}$ in the TE case take the form:

$$\begin{aligned} D_j A_r^{\pm}(\Xi) &= (-1)^{j-1}\big(k_g^2 - (k^+)^2\big)\int_{\Sigma_j} u\bar{w}_{\pm}\,dx_2, \quad j = 1, \ldots, m-1, \\ D_m A_r^{\pm}(\Xi) &= \big(k_g^2 - (k^+)^2\big)\int_{\Sigma_m} u\bar{w}_{\pm}\,dx_1, \\ D_{m+j} A_r^{\pm}(\Xi) &= \big(k_j^2 - k_{j+1}^2\big)\int_{\Lambda_j} u\bar{w}_{\pm}\,dx_1, \qquad j = 1, \ldots, p, \end{aligned} \tag{3.5}$$

where u is the solution to the TE diffraction problem (2.3) and the functions $w_{\pm}$ solve the adjoint TE problems,

$$B_{TE}(\varphi, w_{\pm}) = \frac{\exp\big(\mp i\beta_r^{\pm}b^{\pm}\big)}{d}\int_{\Gamma^{\pm}} \varphi\exp(-irKx_1)\,dx_1 \quad \forall\varphi \in H_p^1(\Omega). \tag{3.6}$$

Here Σ_m is the union of all upper horizontal segments of Γ, whereas Σ_j $(j = 1, \ldots, m-1)$ denotes the vertical segment at the transition point t_j. For the derivation of (3.5) and the corresponding formula (3.7) below in the TM case, we refer to [15]. In the latter case the gradient formulae involve the partial derivatives of the solution of direct and adjoint problems at the interfaces. If the optical index of the grating material is such that the solution

u satisfies the Meixner condition then

$$\begin{aligned}
D_j A_r^{\pm}(\Xi) &= (-1)^{j-1}\big(k_g^2 - (k^+)^2\big) \int\limits_{\Sigma_j} gr(u)\cdot \overline{gr(w_{\pm})}\,dx_2, \quad j = 1,\dots,m-1,\\
D_m A_r^{\pm}(\Xi) &= \big(k_g^2 - (k^+)^2\big) \int\limits_{\Sigma_m} gr_H(u)\cdot \overline{gr_H(w_{\pm})}\,dx_1,\\
D_{m+j} A_r^{\pm}(\Xi) &= \big(k_j - k_{j+1}^2\big) \int\limits_{\Sigma_j} gr_j(u)\cdot \overline{gr_j(w_{\pm})}\,dx_2, \qquad j = 1,\dots,p,
\end{aligned} \tag{3.7}$$

Here, u is the solution of the direct TM problem (2.8), the functions $w_{\pm}$ solve the adjoint problem

$$B_{TM}(\varphi, w_{\pm}) = \frac{\exp\big(\mp i\beta_r^{\pm} b^{\pm}\big)}{d} \int\limits_{\Gamma^{\pm}} \varphi \exp(-irKx_1)\,dx_1 \quad \forall \varphi \in H_p^1(\Omega), \tag{3.8}$$

and

$$\begin{aligned}
gr(u) &= \frac{1}{k^+k_g}\left(\frac{k_g}{k^+}\partial_{x_1,\alpha} u\Big|_{\Sigma_j}^{+}, \partial_{x_2} u\Big|_{\Sigma_j}^{+}\right) = \frac{1}{k^+k_g}\left(\frac{k^+}{k_g}\partial_{x_1,\alpha} u\Big|_{\Sigma_j}^{-}, \partial_{x_2} u\Big|_{\Sigma_j}^{-}\right),\\
gr_H(u) &= \frac{1}{k^+k_g}\left(\partial_{x_1,\alpha} u\Big|_{\Sigma_m}^{+}, \frac{k_g}{k^+}\partial_{x_2} u\Big|_{\Sigma_m}^{+}\right) = \frac{1}{k^+k_g}\left(\partial_{x_1,\alpha} u\Big|_{\Sigma_m}^{-}, \frac{k^+}{k_g}\partial_{x_2} u\Big|_{\Sigma_m}^{-}\right),\\
gr_j(u) &= \frac{1}{k_jk_{j+1}}\left(\partial_{x_1,\alpha} u\Big|_{\Lambda_j}^{+}, \frac{k_{j+1}}{k_j}\partial_{x_2} u\Big|_{\Lambda_j}^{+}\right) = \frac{1}{k_jk_{j+1}}\left(\partial_{x_1,\alpha} u\Big|_{\Lambda_j}^{-}, \frac{k_j}{k_{j+1}}\partial_{x_2} u\Big|_{\Lambda_j}^{-}\right),
\end{aligned}$$

where the plus (resp. minus) signs denote the one-sided limits as the interfaces are approached from the region above (resp. below).

If the Meixner condition is not fulfilled then the gradient formula has to be modified by an additional term depending on λ.

Concerning the solvability of the adjoint problems (3.6) and (3.8) the same existence and uniqueness results as for the direct problems remain valid. This follows from the fact that the solutions of the adjoint TE or TM problem solve the corresponding diffraction problem with the complex conjugate wave numbers $\overline{k}$ and a special radiation condition. For example, for w_- this condition takes the form

$$\begin{aligned}
w_-(x_1,x_2) &= \sum_{n=-\infty}^{\infty} A_n^+ \exp(-i\overline{\beta_n^+}x_2)\exp(inKx_1), \quad x_2 \ge b^+,\\
w_-(x_1,x_2) &= \sum_{n=-\infty}^{\infty} A_n^- \exp(i\overline{\beta_n^-}x_2)\exp(inKx_1)\\
&\quad + \frac{iC}{4\pi\overline{\beta_r^-}} \exp(-i\overline{\beta_r^-}x_2)\exp(irKx_1), \quad x_2 \le b^-
\end{aligned} \tag{3.9}$$

with $C = 1$ for TE and $C = (k^+/k^-)^2$ for TM.

We emphasis that in order to compute all partial derivatives of functionals arising in the optimal design of binary gratings it is sufficient to solve the direct TE and TM diffraction problem and only one corresponding adjoint problem. We demonstrate this for the functional $J(\Xi)$ defined in (3.1).

From formulae (3.4), (3.5), and (3.7) we obtain by linearity that the components of the gradient of $J(\Xi)$ are equal to

$$D_j J(\Xi) = (-1)^{j-1}\mathrm{Re}\left\{\left(k_g^2 - (k^+)^2\right)\left(\int_{\Sigma_j} u^{TE}\,\overline{w^{TE}}\,dx_2 + \int_{\Sigma_j} gr(u^{TM})\cdot\overline{gr(w^{TM})}\,dx_2\right)\right\}, \quad j = 1,\dots,m-1,$$

$$D_m J(\Xi) = \mathrm{Re}\left\{\left(k_g^2 - (k^+)^2\right)\left(\int_{\Sigma_m} u^{TE}\,\overline{w^{TE}}\,dx_1 + \int_{\Sigma_m} gr_H(u^{TM})\cdot\overline{gr_H(w^{TM})}\,dx_1\right)\right\},$$

$$D_{m+j} J(\Xi) = \mathrm{Re}\left\{\left(k_j^2 - k_{j+1}^2\right)\left(\int_{\Lambda_j} u^{TE}\,\overline{w^{TE}}\,dx_1 + \int_{\Lambda_j} gr_j(u^{TM})\cdot\overline{gr_j(w^{TM})}\,dx_1\right)\right\}, \quad j = 1,\dots,p,$$

where u^{TE} and u^{TM} are the solutions of the direct TE and TM problems, respectively, and w^{TE}, w^{TM} solve the following adjoint problems:

$$B_{TE}(\varphi, w^{TE}) = \sum_{r\in P^+} c_r^{TE,+}\,\overline{A_r^{TE,+}}\,\frac{2\beta_r^+\exp(-i\beta_r^+ b^+)}{d\beta}\int_{\Gamma^+}\varphi\exp(-irKx_1)\,dx_1 + \sum_{r\in P^-} c_r^{TE,-}\,\overline{A_r^{TE,-}}\,\frac{2\beta_r^-\exp(i\beta_r^- b^-)}{d\beta}\int_{\Gamma^-}\varphi\exp(-irKx_1)\,dx_1, \quad (3.10)$$

$$B_{TM}(\varphi, w^{TM}) = \sum_{r\in P^+} c_r^{TM,+}\,\overline{A_r^{TM,+}}\,\frac{2\beta_r^+\exp(-i\beta_r^+ b^+)}{d\beta}\int_{\Gamma^+}\varphi\exp(-irKx_1)\,dx_1 + \sum_{r\in P^-} c_r^{TM,-}\,\overline{A_r^{TM,-}}\,\frac{2\beta_r^-(k^+)^2\exp(i\beta_r^- b^-)}{d\beta(k^-)^2}\int_{\Gamma^-}\varphi\exp(-irKx_1)\,dx_1 \quad \forall\varphi\in H_p^1(\Omega). \quad (3.11)$$

Here $A_r^{TE,\pm}$ and $A_r^{TM,\pm}$ denote the Rayleigh amplitudes of u^{TE} and u^{TM}, respectively.

Note that for simple difference approximations of the gradient the number of the direct problems to be solved is at least equal to the number of optimization parameters, whereas the computational costs for solving adjoint and direct problems are the same.

4. NUMERICAL METHOD AND IMPLEMENTATION

Having described the variational formulation and some basic mathematical properties of the direct diffraction problems as well as the gradient formulae and the variational equations of the adjoint problems, we now consider the numerical solution method of these variational problems.

The proposed method combines a finite element method (FEM) in the grating region, where the solutions are not smooth, with Rayleigh series expansions of the solution within the different layers below the grating.

As discussed in Sections 2 and 3 the direct and adjoint problems (2.3), (2.8), (3.10), and (3.11) have the form: find $u \in H_p^1(\Omega)$ satisfying the equations

$$a(u, \varphi) = (f, \varphi) \quad \text{for all } \varphi \in H_p^1(\Omega), \tag{4.1}$$

where $a(u, \varphi)$ is a strongly elliptic sesquilinear form, and (f, φ) stands for a linear and continuous functional on the function space $H_p^1(\Omega)$.

Choosing a sequence of partitions $\{\Omega_h\}$ of Ω with the discretization parameter h and correspondingly a sequence $\{S^h\}$ of finite-dimensional subspaces of $H_p^1(\Omega)$, the strong ellipticity implies that all invertible problems under consideration lead to uniquely solvable linear systems of the form

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \text{for all } \varphi_h \in S^h, \tag{4.2}$$

if h is sufficiently small. Moreover, the approximate solutions converge to the corresponding exact solution in the norm of the function space $H_p^1(\Omega)$ with optimal order.

In the case that the grating is situated on top of a multilayer stack, one can reduce the integration domain Ω used in the FE solution by taking into account that the solution is smooth within the layers. We introduce a new artificial boundary $\tilde{\Gamma}^- = \{x_2 = \tilde{b}^-\}$ into the first layer, $\ell_1 < \tilde{b}^- < 0$ (cf. Fig. 2.1), and new nonlocal boundary operators $\tilde{T}_\alpha^{TE}$ and $\tilde{T}_\alpha^{TM}$ which model the layer system below $\tilde{\Gamma}^-$, together with the radiation condition for $x_2 < b^-$.

In any layer the solution of the corresponding Helmholtz equation can be written as

$$u(x_1, x_2) = \sum_{n=-\infty}^{\infty} \left(A_n^j \exp\left(-i\beta_n^j x_2\right) + B_n^j \exp\left(i\beta_n^j x_2\right)\right) \exp(inKx_1), \quad \ell_{j-1} \le x_2 \le \ell_j,$$

where $\beta_n^j = (k_j^2 - (n+\alpha)^2)^{1/2}$, and we set $\ell_0 = 0$. The transmission conditions at each of the interfaces connect the coefficients (A_n^j, B_n^j) and (A_n^{j+1}, B_n^{j+1}) via 2×2 transmission matrices. Since the radiation condition (2.6) implies $B_n^- = 0$, one gets explicit formulae for the numbers γ_n connecting the nth Fourier coefficient of a solution and its normal derivative on $\tilde{\Gamma}^-$:

$$\partial_n u|_{\tilde{\Gamma}^-} = -\sum_{n\in\mathbf{Z}} i\gamma_n \hat{u}_n \exp(inKx_1), \quad \text{where } \hat{u}_n = \frac{1}{d}\int_0^{2\pi} u(x_1, \tilde{b}^-) \exp(-inKx_1)\, dx_1.$$

The coefficients γ_n are different for TE and TM polarization, but it can be easily seen that they converge to $\beta_n^1 = (k_1^2 - \alpha_n^2)^{1/2}$, i.e. $|\gamma_n - \beta_n^1| \to 0$ as $|n| \to \infty$. For evaluating these scalars one can use a recursive algorithm which is numerically stable for any number of layers, and there is no limit in layer thickness. Matrix algorithms of this type are widely used in other numerical methods for analyzing layered structures (see [7] and the references therein).

Thus, if we define nonlocal boundary operators on $\tilde{\Gamma}^-$,

$$\tilde{T}_\alpha^{TE} u = -\sum_{n\in\mathbf{Z}} i\gamma_n^{TE} \hat{u}_n \exp(inKx_1), \quad \tilde{T}_\alpha^{TM} u = -\sum_{n\in\mathbf{Z}} i\gamma_n^{TM} \hat{u}_n \exp(inKx_1),$$

the direct problems (2.3) and (2.8) are equivalent to the variational equations on the smaller rectangle $\tilde{\Omega} = (0, d) \times (\tilde{b}^-, b^+)$,

$$\begin{aligned}\tilde{B}_{TE}(u, \varphi) &:= \int_{\tilde{\Omega}} \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - \int_{\tilde{\Omega}} k^2 u \overline{\varphi} + \int_{\Gamma^+} (T_\alpha^+ u) \overline{\varphi} + \int_{\tilde{\Gamma}^-} \left(\tilde{T}_\alpha^{TE} u\right) \overline{\varphi} \\ &= -\int_{\Gamma^+} 2i\beta \exp(-i\beta b^+) \overline{\varphi} \quad \forall \varphi \in H_p^1(\tilde{\Omega}),\end{aligned} \tag{4.3}$$

respectively

$$\begin{aligned}\tilde{B}_{TM}(u, \varphi) &:= \int_{\tilde{\Omega}} \frac{1}{k^2} \nabla_\alpha u \cdot \overline{\nabla_\alpha \varphi} - \int_{\tilde{\Omega}} u \overline{\varphi} + \int_{\Gamma^+} \frac{1}{(k^+)^2} (T_\alpha^+ u) \overline{\varphi} + \int_{\tilde{\Gamma}^-} \frac{1}{k_1^2} \left(\tilde{T}_\alpha^{TM} u\right) \overline{\varphi} \\ &= -\int_{\Gamma^+} \frac{2i\beta}{(k^+)^2} \exp(-i\beta b^+) \overline{\varphi} \quad \forall \varphi \in H_p^1(\tilde{\Omega}).\end{aligned} \tag{4.4}$$

The adjoint problems (3.6) and (3.8) are reduced analogously to variational formulations on $\tilde{\Omega}$, but note that for w_- the right-hand sides change according to the layer structure.

Due to the simple geometry of binary gratings it is quite natural to choose as finite elements piecewise bilinear functions on a uniform rectangular partition of $\tilde{\Omega} = (0, d) \times (\tilde{b}^-, b^+)$. This leads to a linear system with a block-tridiagonal matrix. The nonlocal boundary terms in the sesquilinear forms imply that the first and the last block of the main diagonal are fully occupied matrices, whereas the remaining blocks are sparse.

Let us note that the computation of these nonlocal terms can be performed very efficiently with an accuracy comparable with the computer precision. Since the traces of the finite element functions on $\Gamma^\pm$ are piecewise linear periodic functions with uniformly distributed break points, it is possible to use recurrence relations for the Fourier coefficients of spline functions and convergence acceleration methods.

If the artificial boundary Γ^+ is divided into m subintervals of equal length and the basis of hat functions $\{\varphi_j\}$ is used, then the form

$$\int_{\Gamma^+} (T_\alpha^+ \varphi_p) \overline{\varphi_q} \, dx_1 \tag{4.5}$$

corresponds to an $m \times m$ circulant matrix with the eigenvalues

$$\tau_0 = -id\beta, \quad \tau_p = -2id \left(\frac{\sin(\pi p/m)}{\pi} \right)^4 \sum_{r=-\infty}^{\infty} \frac{\beta_{rm+p}^+}{m(r + p/m)^4}, \quad p = 1, \ldots, n-1.$$

Thus, one only has to expand

$$\frac{\beta_{rm+p}^+}{m} = \sqrt{\left(\frac{k^+}{m}\right)^2 - \left(\frac{\alpha}{m} + \left(r + \frac{p}{m}\right) K\right)^2}$$

with respect to powers of $|r + p/m|$ and to use fast computation of the generalized zeta function

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

Since the scalars γ_n converge very fast to β_n^1, for the computation of the forms

$$\int_{\tilde{\Gamma}^-}\left(\tilde{T}_\alpha^{TE}\varphi_p\right)\overline{\varphi_q}\,dx_1,\qquad \int_{\tilde{\Gamma}^-}\left(\tilde{T}_\alpha^{TM}\varphi_p\right)\overline{\varphi_q}\,dx_1,$$

one has to compute only a few of these coefficients and apply the summation method mentioned before.

Thus the discretization error of the direct and adjoint problems is mainly determined by the approximation error of the solution with bilinear finite elements. There hold the following convergence results:

THEOREM 2 [15]. (*a*) *If the TE problem* (2.3) *has a unique solution, then for all sufficiently small h* > 0 *the FE discretization of* (2.3) *and* (3.10) *are uniquely solvable and the approximate solutions converge to the corresponding exact solution in the norm of* $L^2(\Omega)$ *with the optimal rate* $O(h^2)$.

(*b*) *If the TM problem* (2.8) *has a unique solution, then for all sufficiently small h* > 0 *the FE discretization of* (2.8) *and* (3.11) *are uniquely solvable and the approximate solutions converge to the corresponding exact solution with the rate* $O(h^{2\lambda})$.

Together with error estimates in the norm of the function space $H_p^1(\Omega)$ it is easy to derive similar estimates for the approximation of the diffraction efficiencies and the gradients of the minimizing functionals.

As mentioned in the Introduction usual FE approximations of the Helmholtz equation involve besides the approximation error also the pollution error which increases, together with the wave number and enlarging domains. For example, due to [23] piecewise linear FE methods provide the suboptimal discretization error of the form $O(h^2k^{3+\alpha})$, where $\alpha > 0$ depends on the domain and the boundary conditions. Roughly speaking, the pollution error is caused by the well-known fact that the discretization of the Helmholtz equation with the wave number k results in an approximate solution possessing a different wave number k_h. In one-dimensional problems, for example, the usual piecewise linear FE solution of the equation $u'' + k^2u = 0$ on a uniform grid has the discrete wave number

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

It turns out that this “phase lag” leads to the suboptimal error estimate mentioned above. For the one-dimensional case one can easily define a finite element discretization with vanishing phase lag by introducing a modified wave number. So the FE solution of the equation

$$u'' + (\tilde{k}(h))^2u = 0 \quad \text{with } \tilde{k}(h) = \frac{6(1-\cos kh)}{h^2(2+\cos kh)}$$

has the wave number k and, for $h \to 0$, the corresponding solutions converge to the exact solution with the order $O((hk)^2)$ (cf. [21]). Thus, in the one-dimensional case it is possible to construct a generalized FEM without pollution by modifying the evaluation of the sesquilinear form.

Several approaches designed to improve the numerical phase accuracy of FEM in higher dimensions have been proposed in recent years (cf. [24] and the references therein). However, as shown in [25], it is not possible to eliminate the pollution in the FE error by any modification of the evaluation of the sesquilinear form. Therefore we introduce a generalized FEM with minimal phase lag which already for rather poor discretizations of the domain $\tilde{\Omega}$ provides excellent results compared with the usual FEM for both the TE and TM modes. Here we apply and extend the approach of [21] to design a so-called GFEM with minimal pollution ensuring that the wave number of the approximate solution almost coincides with the given k for piecewise uniform rectangular partitions of $\tilde{\Omega}$.

To fix the idea, consider on a rectangular mesh of size (h_1, h_2) the bilinear FE discretization of the Helmholtz equation $\Delta u + k^2 u = 0$ with constant k in some interior node. The corresponding matrix has a stencil of the form

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

with the coefficients

$$\begin{aligned} a_0 &= \frac{4\left(h_1^2 + h_2^2\right)}{3h_1h_2} - \frac{4k^2h_1h_2}{9}, \quad & a_1 &= \frac{h_1^2 - 2h_2^2}{3h_1h_2} - \frac{k^2h_1h_2}{9}, \\ a_2 &= \frac{h_2^2 - 2h_1^2}{3h_1h_2} - \frac{k^2h_1h_2}{9}, & a_3 &= -\frac{h_1^2 + h_2^2}{6h_1h_2} - \frac{k^2h_1h_2}{36}. \end{aligned} \tag{4.6}$$

Since the function $\exp(ik_1x_1 + ik_2x_2)$ with $k_1 = k\cos\theta,\ k_2 = k\sin\theta,\ \theta \in [0, 2\pi]$, solves the Helmholtz equation, we expect that a proper discretization of this equation at inner points should annihilate the grid functions $v_\theta(ph_1, qh_2) = \exp(ik_1ph_1 + ik_2qh_2)$ for all $\theta \in [0, 2\pi]$. But the application of the FEM stencil to the discrete function v_θ results in

$$\begin{aligned} &\exp(ik_1ph_1 + ik_2qh_2)(a_0 + 2a_1\cos(k_1h_1) + 2a_2\cos(k_2h_2) + 4a_3\cos(k_1h_1)\cos(k_2h_2)) \\ &\quad = \exp(ik_1ph_1 + ik_2qh_2)\left(h_1h_2\left(h_1^2k_1^4 + h_2^2k_2^4\right)/12 + O\left(h_1^3h_2^3\right)\right) \end{aligned}$$

at the grid point (ph_1, qh_2). Thus, the grid function v_θ does not satisfy the usual FE discretization of the Helmholtz equation at interior nodes. The idea of [21] was to modify the coefficients a_i of the interior stencil such that its application to the discrete functions v_θ vanishes. This means, the ellipse $\mathcal{E}_{h_1h_2}(\theta) = (kh_1\cos\theta, kh_2\sin\theta),\ \theta \in [0, 2\pi]$, should belong to the set of roots of the symbol function associated with the stencil

$$G(\xi_1, \xi_2) := a_0 + 2a_1\cos(\xi_1) + 2a_2\cos(\xi_2) + 4a_3\cos(\xi_1)\cos(\xi_2). \tag{4.7}$$

However, for any choice of the coefficients $a_0, \ldots, a_3$ the zero set of the function $G(\xi_1, \xi_2)$ does not contain any ellipse. Therefore, one has to 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, the coefficients a_1, a_2, and a_3 will depend on kh_1 and kh_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 [25] 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

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

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 in some special norm from below by

$$\|u - u_{GFE}\|^2 \geq ch^{-1}\,\mathcal{D}_{h_1h_2}, \tag{4.8}$$

where we set $h = \sqrt{h_1h_2}$ (compare [21, 25]). To find a stencil providing asymptotically the minimal distance $\mathcal{D}_{h_1h_2}$, we use the fact that the asymptotics

$$\max_{\theta\in[0,2\pi]} |G(kh_1\cos\theta, kh_2\sin\theta)| = O((kh)^{\ell}), \quad \mathcal{D}_{h_1h_2} = O((kh)^{\ell-1})$$

are equivalent. This can be easily seen 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 in such a way that asymptotically $\max_{\theta\in[0,2\pi]}|g(\theta)|$ is minimal. Here g denotes the π-periodic function $g(\theta) = G(kh_1\cos\theta, kh_2\sin\theta)$ with the Fourier series

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

Note that in the case $h_1 = h_2$ the function g has even the period $\pi/2$. The Fourier coefficients of g can be obtained by using the formulas

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

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

with the first kind Bessel function J_k. Thus, the Fourier coefficients have the asymptotics

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

if $a_1, a_2, a_3 = O(1)$. Consequently, the function g with asymptotically minimal $\max_{\theta\in[0,2\pi]}|g(\theta)|$ is found if, for given a_0, the values of a_1, a_2, and a_3 are chosen such that the first three Fourier coefficients of g vanish, $\hat{g}_0 = \hat{g}_2 = \hat{g}_4 = 0$, which ensures that $\max_{\theta\in[0,2\pi]}|g(\theta)| = O((kh)^6)$. Introducing the function

$$j_k(x) := 2^k J_k(x)/x^k = \sum_{n=0}^{\infty} \frac{(-x)^n}{2^n n!(n+k)!}$$

the corresponding linear system can be written in the form

$$\begin{aligned} j_0(kh_1)a_1 + j_0(kh_2)a_2 + 2j_0\big(k\sqrt{h_1^2+h_2^2}\big)a_3 &= -a_0/2 \\ h_1^2 j_2(kh_1)a_1 - h_2^2 j_2(kh_2)a_2 + 2\big(h_1^2 - h_2^2\big)\, j_2\big(k\sqrt{h_1^2+h_2^2}\big)a_3 &= 0 \\ h_1^4 j_4(kh_1)a_1 + h_2^4 j_4(kh_2)a_2 + \big(2h_1^4 - 12h_1^2h_2^2 + 2h_2^4\big)\, j_4\big(k\sqrt{h_1^2+h_2^2}\big)a_3 &= 0. \end{aligned} \tag{4.9}$$

Note that in the case of a quadratic partition, $h_1 = h_2$, the solution of (4.9) satisfies $a_1 = a_2$, and there holds even $\max_{\theta\in[0,2\pi]}|g(\theta)| \le O((kh)^8)$.

The distance between the ellipse $\mathcal{E}_{h_1h_2}$ and the zero set of the symbol function $G(\xi_1,\xi_2)$ associated with the solution of (4.9) can be estimated similarly to the technique of [25] for the case $h_1 = h_2$. One uses an expansion of the zeroes of G in the form

$$\xi_1 = kh_1\left(1 + \sum_{m=1}^{\infty} r_m(\theta,q)(kh)^{2m}\right)\cos\theta, \quad \xi_2 = kh_2\left(1 + \sum_{m=1}^{\infty} r_m(\theta,q^{-1})(kh)^{2m}\right)\sin\theta,$$

where $q = \sqrt{h_1/h_2}$. From the Taylor series expansion of G one deduces that $G(\xi_1,\xi_2) = 0$ in a neighborhood of the ellipse if

$$\begin{aligned} r_1(\theta,q) &= 0, \\ r_2(\theta,q) &= \frac{(q^4 - q^{-4})\cos 6\theta}{15360}, \\ r_3(\theta,q) &= \frac{(q^6 + q^{-6})\cos 8\theta}{1548288} + \frac{(q^6 - q^{-6})\cos 6\theta}{193536} + \frac{(q^2 - q^{-2})(q^4 - q^{-4})\cos 4\theta}{737280}. \end{aligned}$$

Note that for the usual bilinear FEM stencil with the coefficients (4.6) there holds

$$\begin{aligned} r_1(\theta,q) = {} & \frac{q+q^{-1}}{96} + \frac{q^3+q^{-3}}{64} - \frac{q^2+q^{-2}}{24} + \frac{(q^3+q^{-3} - 2(q+q^{-1}))\cos 4\theta}{192} \\ & + \frac{(q^3 - q^{-3} - 2(q^2 - q^{-2}))\cos 2\theta}{48}. \end{aligned}$$

Since

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

for the stencil of the GFEM one obtains the estimate

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

in the case of rectangular partitions, and for quadratic partitions one even has the improved estimate obtained already in [21]

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

whereas in any case the FEM stencil admits the lower bound

$$\mathcal{D}_{h_1 h_2} \geq \frac{1}{24}(kh)^3 + O((kh)^5).$$

A rigorous analysis for the convergence of GFEM in two-dimensional problems similar to Theorem 2 is not known up to now. For the special case of constant k and Dirichlet boundary conditions one can show by using corresponding results for finite difference methods that the GFE discretization is uniquely solvable for sufficiently small h_1 and h_2 and that it provides h^2 convergence in the L^2-norm. However, the dependence on k of the constant in this error estimate is an open problem, so that estimates from above corresponding to (4.8) are not known.

The GFEM stencil can be adapted for solving the direct and adjoint variational TE and TM problems under consideration which contain the differential operator $\Delta + 2i\alpha\partial_{x_1} + (k^2 - \alpha^2)$. The domain $\tilde{\Omega}$ is partitioned such that the rectangular mesh is uniform in the x_1-direction and piecewise uniform in the x_2-direction and such that the discontinuities of k lie on mesh lines. 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

$$v_\theta(ph_1, qh_2) = \exp(i(k_1 + \alpha)ph_1 + ik_2 qh_2) \quad \text{with } k_1 = k\cos\theta,\ k_2 = k\sin\theta,$$

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 (4.9). The scaling is necessary due to the jumps of k and to the boundary conditions with the nonlocal 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 $(d/dx)^2 + (k^2 - \alpha^2)$.

The sparse structure of the matrix can be used to apply efficient direct or iteration methods for solving linear systems. We use a block version of the so-called sweep method, which utilizes the block-tridiagonal structure of the matrix and additionally the circulant properties of the dense blocks. Since the matrices of the discretized variational problems are nonsymmetric, we apply preconditioned GMRES-type and BiCGstab methods as iterative solvers. For many technological relevant grating materials and wavelengths the optical indices do not strongly jump. Therefore the corresponding equations with averaged wave numbers k are good candidates for the preconditioner, which can be inverted very efficiently using FFT.

After having solved the linear system corresponding to the GFE discretization of the variational equations, the diffraction efficiencies are determined from the Fourier coefficients of the solution on Γ^+ and $\tilde{\Gamma}^-$. For the computation of the transmission efficiencies and the solutions on the layer interfaces, which appear in the gradient formulae, we use a stable recursive algorithm similar to that for evaluating the coefficients γ_n.

5. SOME NUMERICAL RESULTS

The method was used to evaluate the reflection and transmission efficiencies of binary gratings on multilayer systems of different geometries and materials and it turned out to be robust and reliable in both the TE and TM case. Compared with the usual FEM the obtained results were accurate already for rather poor discretizations. In Fig. 4.1 we compare the numerical values of some reflection and transmission efficiencies versus the square root n of the 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 $\nu = 2.5$ situated on a layer with $\nu = 3.5$. In each 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, we compared the results of our method with those obtained with other methods which are known to provide reliable results for binary gratings (e.g., integral equation or modal methods). As an example we give in Table I the zero order reflection efficiencies of TM polarization for a simple binary grating calculated with different methods. The grating consists of aluminium with the optical index $\nu = 0.47 + 4.8i$ for the given wavelength of 436 nm, the grating period d is equal to 1 μm, the fill factor $f = 0.5$, and the angle of incidence $\theta = 0$.

Table I compares the corresponding values of GFEM with an quadratic partitioning of the rectangular domain with $h_1 = h_2 = 10$ nm for different heights H of the binary structure with the results of three other methods, taken form [26]. These methods are two modal methods, AWG (analytic waveguide method), introduced in [3, 27] and the RCWA (rigorous coupled-wave analysis) going back to [2] and essentially improved in recent years (cf. [6]). The third method called IESMP is based on the integral equation method as described in [5, 26].

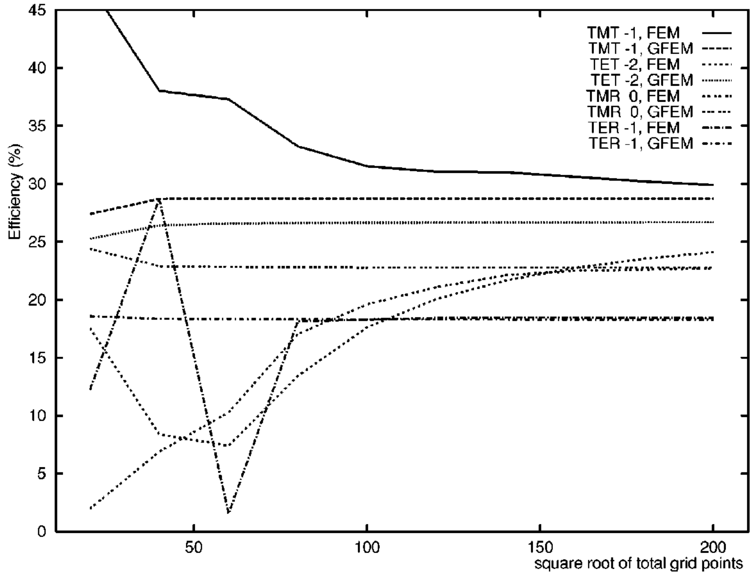

FIG. 4.1. Comparison of some TE and TM reflexion and transmission efficiencies computed with FEM and GFEM for a simple binary grating with $\nu = 3.5$ versus the square root n of total grid points.

TABLE I
Comparison of Zero Order TM Efficiency Computed with Different Methods for Simple Aluminium Gratings for Normal Incidence

H/d	AWG	RCWA	IESPM	GFEM
0.1	0.0186	0.0173	0.0190	0.0190
0.2	0.8532	0.8539	0.8529	0.8533
0.3	0.0095	0.0096	0.0100	0.0098
0.4	0.8079	0.8080	0.8095	0.8095
0.5	0.0440	0.0445	0.0465	0.0452
0.6	0.7000	0.7000	0.7068	0.7027
0.7	0.1497	0.1496	0.1511	0.1506
0.8	0.6250	0.6234	0.6277	0.6257
0.9	0.2500	0.2503	0.2503	0.2504
1.0	0.4810	0.4808	0.4840	0.4816

Note. The parameters are $\lambda = 436$ nm, $d = 1\,\mu$m, $\nu = 0.47 + 4.8i$ and $f = 0.5$.

Note that the convergence of finite element methods is not restricted to the case of binary gratings considered here. These methods can handle very general geometries of the diffraction structures and complex materials. The implementation of effective solvers is simple, especially for problems with polygonal interfaces between different materials and the practical limits are determined only by the computer resources for solving the corresponding linear systems of equations. The implementation of effective solvers for the other methods in the case of more general diffraction problems is very complicated. Moreover, a convergence analysis for these methods is not known at present. However, for smooth interfaces and a small number of different grating materials integral equation methods and the analytical continuation method [8] seem to be advantageous, whereas in the case of rectangular interfaces the methods based on Fourier series or eigenmode expansions give equivalent results.

The GFEM for solving direct and adjoint problems was integrated into a computer program for the study of optimal design problems for binary gratings. By using the standard algorithm of gradient descent local minima of functionals are determined, which characterize desired optical properties. These functionals involve the Rayleigh coefficients of the discrete models on a given partition of the domain Ω for a prescribed range of incidence angles or wavelengths. Of course, the gradients are computed by discretized versions of the formulae given in Section 3. Corresponding to the gradients the thicknesses of the layers and the shape of Γ are varied within a class of admissible parameters, which are restricted by certain technological constraints.

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.

In the following we provide some results of the optimization of a polarisation grating, beam splitters and high reflection mirrors.

The first example concerns the application of metallic subwavelength gratings for polarization devices. Figure 4.2 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. Here the refractive index of

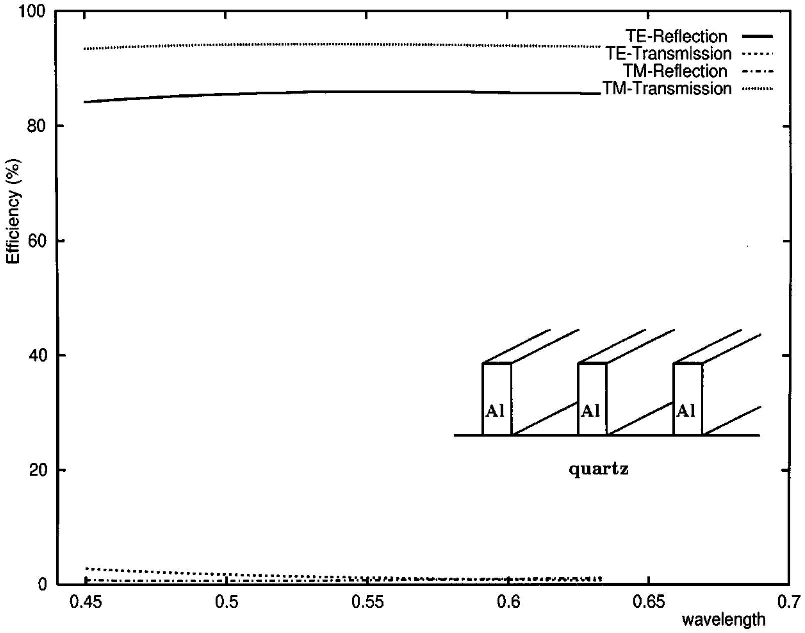

FIG. 4.2. Optimal design for a simple polarisation grating for the range of wavelengths from 450 to 633 nm. Grating parameters are $d = 200$ nm, $H = 150$ nm, and $f = 0.3$.

aluminium is given as a function of the wavelength and the grating period is fixed to 200 nm. The optimization results in the width of the bar of 60 nm and in the height of 150 nm.

Next we provide the optimization results for some beam splitters. The illuminating unpolarized wave with $\lambda = 0.633\,\mu$m is normally incident from a dielectric medium with refractive index $\nu = 1.5315$. Choosing the period $d = 1.266\,\mu$m three diffraction orders propagate with angles 0 and ±30°. The goal is

(a) to maximize the efficiencies of the orders ±1

(b) to obtain maximal and equal efficiencies of all three orders

by optimizing the height H and the fill factor f of the grating with one groove per period. The results are depicted in Figs. 4.3a,b, the obtained values are

(a) $H = 0.734\,\mu$m, $f = 0.72$;

(b) $H = 0.43\,\mu$m, $f = 0.58$.

For the same parameters as before we seek a one-to-four beam splitter with the diffraction angles ±14.5° and ±30°. Choosing the period $d = 2.532\,\mu$m, nine diffraction orders propagate; the goal is to maximize the efficiencies of the orders ±1 and ±2. To obtain a satisfactory solution, it is necessary to use two grooves per period. For the optimal solution the height of these grooves is $H = 1.747\,\mu$m, the scaled transition points are 0.0, 0.24, 0.38, 0.63 (Fig. 4.4).

For the same parameters as before we optimized a one-to-five beam splitter with the diffraction angles 0°, ±20.7°, and ±45°. For the fixed grating period $d = 1.79\,\mu$m the optimization provides the height of the optimal grooves $H = 0.77\,\mu$m and the scaled transition points 0.0, 0.12, 0.36, 0.76 (Fig. 4.5).

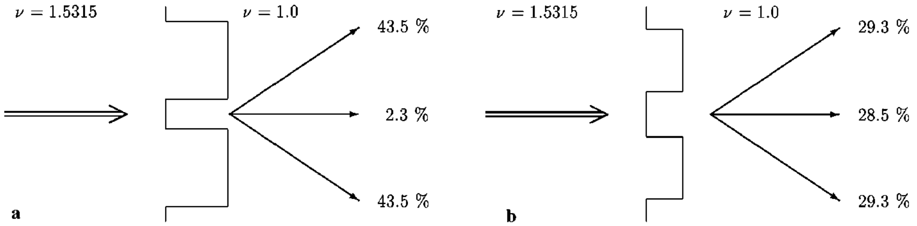

FIG. 4.3. (a) Optimal design of a 1-to-2 beam splitter. The parameters are $\lambda = 0.633\,\mu$m, $d = 1.266\,\mu$m, $H = 0.734\,\mu$m, and $f = 0.72$. (b) Optimal design of a 1-to-3 beam splitter. The parameters are $\lambda = 0.633\,\mu$m, $d = 1.266\,\mu$m, $H = 0.43\,\mu$m, and $f = 0.58$.

The next problem concerns the design of a zero-order copper grating ($\nu = 12.7 + 51.1i$) as circular polarizer for CO_2 laser with $\lambda = 10, 6\,\mu$m such that in the range of incident angles $\theta \in (29^\circ, 31^\circ)$ the efficiencies of the reflected TE and TM polarized wave are maximal and the phase difference between them is close to $\pi/2$. Here one has to minimize the functional (3.3) extended over the range of incident angles, which possesses many local minima. One of the reasonable geometries is $d = 3.0\,\mu$m, $H = 1.65\,\mu$m, and $f = 0.24$. Table II contains the computed values.

Finally we consider a high reflection grating on top of a quarter-wave system of 15 layers for the wavelength $\lambda = 1.45\,\mu$m. The even homogeneous-layer parameters are $\nu = 1.45$ and $h_j = 248$ nm, with the odd homogeneous-layer parameters being $\nu = 2.3$ and $h_j = 157$ nm. The substrate is quartz with $\nu = 1.45$. Without any grating structure the reflection efficiency is almost 100% (99.76% in normal incidence). The problem is to find a grating surface in an additional quartz layer on the top in order to maximize the TE reflection of order -1 in Littrow mounting for $\theta = 20.4^\circ$. Correspondingly, the period of the grating is $d = 2.06\,\mu$m. Optimal values were obtained for the thickness of the additional quartz layer of 866 nm, the binary grating within this layer has the height $H = 804$ nm and the fill factor $f = 0.56$. In that case the efficiency of order -1 amounts to 99.42%.

6. CONCLUSION

In this paper we focused on optimal design problems for binary gratings, using exact formulae for the gradients of the cost functionals and a fast and reliable method for the numerical solution of direct and adjoint diffraction problems. The latter method is based on a variational formulation and combines a finite element method in the grating structure

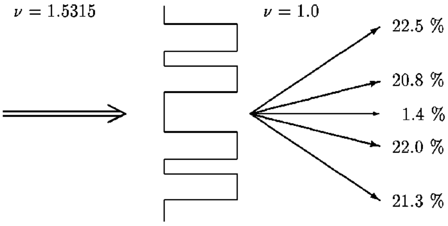

FIG. 4.4. Optimal design of a 1-to-4 beam splitter for the wavelength $\lambda = 0.633\,\mu$m. Grating parameters are $d = 2.532\,\mu$m and $H = 1.747\,\mu$m. The distribution of the transition points is 0., 0.24, 0.38, 0.63.

TABLE II
Zero Order Efficiencies and Phase Difference for Circular Polarizer

θ	TE	TM	Phase
29.0	97.50	95.72	90.72
29.2	97.50	95.72	90.58
29.4	97.51	95.72	90.45
29.6	97.51	95.72	90.32
29.8	97.52	95.72	90.18
30.0	97.52	95.72	90.04
30.2	97.53	95.72	89.91
30.4	97.53	95.72	89.77
30.6	97.54	95.72	89.63
30.8	97.54	95.72	89.49
31.0	97.55	95.72	89.35

Note. The parameters are $\lambda = 10, 6\,\mu$m, $\nu = 12.7 + 51.1i$, $d = 3\,\mu$m, $H = 1.65\,\mu$m, and f $= 0.24$.

with Rayleigh series expansions in the layer system below the grating. This approach is not restricted to binary profiles, but allows the numerical treatment of rather general diffraction structure, together with a complete convergence analysis.

We proposed a generalized finite element method (GFEM) with minimal pollution, which provides highly accurate numerical results in the computation of diffraction efficiencies for both the TE and TM mode. In particular, for TM diffraction problems having a mild singularity of the solution, the convergence performance of our method was comparable with that of the rigorous coupled-wave analysis of [6] and the integral equation method of [5]. Moreover, accurate numerical results can be obtained even in the presence of strong singularities of the solution. We expect that the approach can be also extended to the more general case of conical diffraction and biperiodic gratings.

To solve optimal design problems for binary gratings by gradient descent we presented explicit formulae for the gradients with respect to the parameters of the grating profile and the thicknesses of layers. These formulae involve the solutions of direct and adjoint TE and TM problems and reduce considerably the computational costs compared to simple difference approximations of the gradients. The GFEM and the gradient formulae were integrated into a computer program to find the optimal design of binary gratings with desired phase or intensity pattern for a given range of incidence angles or wavelength. Several numerical

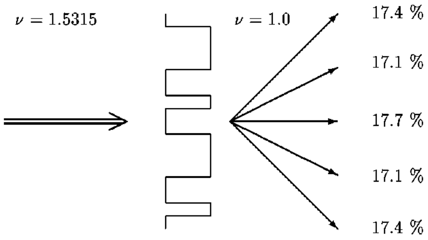

FIG. 4.5. Optimal design of a 1-to-5 beam splitter for the wavelength $\lambda = 0.633\,\mu$m. Grating parameters are $d = 1.79\,\mu$m and $H = 0.77\,\mu$m. The distribution of the transition points is 0., 0.12, 0.36, 0.76.

examples including polarisation gratings and beam splitters successfully demonstrate the efficiency of the algorithm.

ACKNOWLEDGMENTS

We are grateful to Bernd Kleemann (Berlin Institute of Optics) and Andreas Pomp (University of Stuttgart) for fruitful discussions concerning direct solvers for diffraction problems and for providing us several interesting optimal design problems.

REFERENCES

1. R. Petit (Ed.), *Electromagnetic Theory of Gratings*, Topics in Current Physics, Vol. 22 (Springer-Verlag, Berlin, 1980).
2. M. G. Moharam and T. K. Gaylord, Rigorous coupled-wave analysis of planar-grating diffraction, *J. Opt. Soc. Am.* **71**, 811 (1981).
3. L. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams, and J. R. Andrewartha, The dielectric lamellar diffraction grating, *Opt. Acta* **28**, 413 (1981).
4. F. Montiel and M. Nevière, Differential theory of gratings: extension to deep gratings of arbitrary profile and permittivity through the R-matrix propagation algorithm, *J. Opt. Soc. Am. A* **11**, 1321 (1994).
5. A. Pomp, The integral method for coated gratings: computational cost, *J. Mod. Opt.* **38**, 109 (1991).
6. P. Lalanne and G. M. Morris, Highly improved convergence of the coupled-wave method for TM polarization, *J. Opt. Soc. Am. A* **13**, 779 (1996).
7. L. Li, Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings, *J. Opt. Soc. Am. A* **13**, 1024 (1996).
8. O. P. Bruno and F. Reitich, Numerical solution of diffraction problems: a method of variation of boundaries, *J. Opt. Soc. Amer. A* **10**, 1168 (1993).
9. A.-S. Bonnet-Bendhia and F. Starling, Guided waves by electromagnetic gratings and non-uniqueness examples for the diffraction problem, *Math. Methods Appl. Sci.* **17**, 305 (1994).
10. G. Bao, D. C. Dobson, and J. A. Cox, Mathematical studies in rigorous grating theory, *J. Opt. Soc. Amer. A* **12**, 1029 (1995).
11. D. C. Dobson, A variational method for electromagnetic diffraction in biperiodic structures, *Model. Math. Anal. Numer.* **28**, 419 (1994).
12. D. C. Dobson and J. A. Cox, Mathematical modeling for diffractive optics, *Proc. SPIE* **CR49**, 32 (1994).
13. G. Bao, Finite element approximation of time harmonic waves in periodic structures, *SIAM J. Numer. Anal.* **32**, 1155 (1995).
14. G. Bao, Numerical analysis of diffraction by periodic structures: TM polarization, *Numer. Math.* **75**, 1 (1996).
15. J. Elschner and G. Schmidt, Diffraction in periodic structures and optimal design of binary gratings. I. Direct problems and gradient formulas, *Math. Methods Appl. Sci.*, in press.
16. I. H. Park, H. K. Oh, and S. B. Hyun, Photolithography simulation on nonplanar substrate using the finite-element method with absorbing boundary conditions, *Proc. SPIE* **3051**, 578 (1997).
17. D. C. Dobson, Optimal design of periodic antireflective structures for the Helmholtz equation, *Eur. J. Appl. Math.* **4**, 321 (1993).
18. T. Jaaskelainen and M. Kuittinen, Inverse grating diffraction problems, *Proc. SPIE* **1574**, 272 (1991).
19. E. Noponen, J. Turunen, and A. Vasara, Parametric optimization of multilevel diffractive optical elements by electromagnetic theory, *Appl. Opt.* **31**, 5910 (1992).
20. H. Kikuta, Y. Ohira, and K. Iwata, Subwavelength gratings optimized for broadband quarter-wave plates, *Proc. SPIE* **2873**, 218 (1996).
21. I. Babuška, F. Ihlenburg, E. Paik, and S. Sauter, A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution, *Comp. Methods Appl. Mech. Eng.* **128**, 325 (1995).
22. P. Ciarlet, *The Finite Element Method for Elliptic Problems* (North-Holland, Amsterdam/New York, 1978).

23. A. Bayliss, C. I. Goldstein, and E. Turkel, On accuracy conditions for the numerical computation of waves, *J. Comput. Phys.* **59**, 396 (1985).

24. L. J. Thompson and P. M. Pinsky, A Galerkin least-squares finite element method for the twodimensional Helmholtz equation, *Int. J. Numer. Methods Eng.* **38**, 371 (1995).

25. I. Babuška and S. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers, *SIAM J. Numer. Anal.* **34**, 2392 (1997).

26. M. Davidson, B. Kleemann, and J. Bischoff, A comparison between rigorous light scattering methods, *Proc. SPIE* **3051**, 606 (1997).

27. M. Davidson, K. Monahan, and R. Monteverde, Linearity of coherence probe metrology: simulation and experiment, *Proc. SPIE* **1464**, 155 (1991).