Optimal shape design of blazed diffraction gratings

David C. Dobson*

Abstract. The problem of designing a periodic interface between two materials in such a way that time-harmonic waves diffracted from the interface have a specified far-field pattern is studied.

A minimization problem for the interface is formulated, and it is shown that solutions of constrained bounded variation exist. The differentiability of the cost functional is then studied, with no restrictions on the smoothness of the interface. Some computational issues are discussed, and finally the results of some numerical experiments are presented.

Key Words. Diffraction, periodic structure, optimal shape design.

AMS Classification. 78A45, 49J20, 65K10, 49J50

1. Introduction. The problem considered here is that of finding the shape of a periodic surface profile which separates two materials, in such a way that waves diffracted from the profile have a pre-specified diffraction pattern (scattered field) for a given incident plane wave. This problem is motivated by applications in diffractive optics, a technology in which optical components are micromachined with tiny features. An overview of some of the technological aspects of this field can be found in [16]. One of the most common geometrical configurations for diffractive optical structures is a periodic pattern etched into the surface of an optical substrate, as shown in Figure 1. Such structures are often called *blazed diffraction gratings*.

Our goal is to find structures which diffract light in a given pattern. This problem is formulated as a least-squares minimization. The problem is ill-posed in the sense that there may exist minimizing sequences of surface profiles which converge only in some weak sense. This difficulty can be overcome by "relaxing" the problem, that is, enlarging the class of admissible designs to include "mixtures" of materials. This approach was introduced by Achdou [2] for solving a closely related photocell optimization problem and was used in [9] to design antireflective structures.

Alternatively, the admissible designs can be constrained to some smaller class of curves. It is this latter approach that we pursue here. A key feature of our approach is that nonsmooth profiles present no theoretical or numerical difficulties. This is important from a practical point of view since the mask-etch fabrication process for blazed diffraction gratings generally yields nonsmooth interface profiles.

Several articles carrying out analytical work for determining periodic scattering interfaces in the context of inverse problems have recently appeared. See in particular [3, 6, 15] for uniqueness and stability results. A good introduction to the periodic scattering problem can be found for example in [7, 17]. Descriptions of some additional mathematical problems which arise in diffractive optics and industrial applications may be found in the books [14].

* Department of Mathematics, Texas A&M University, College Station, Texas 77843-3368, USA. dobson@math.tamu.edu

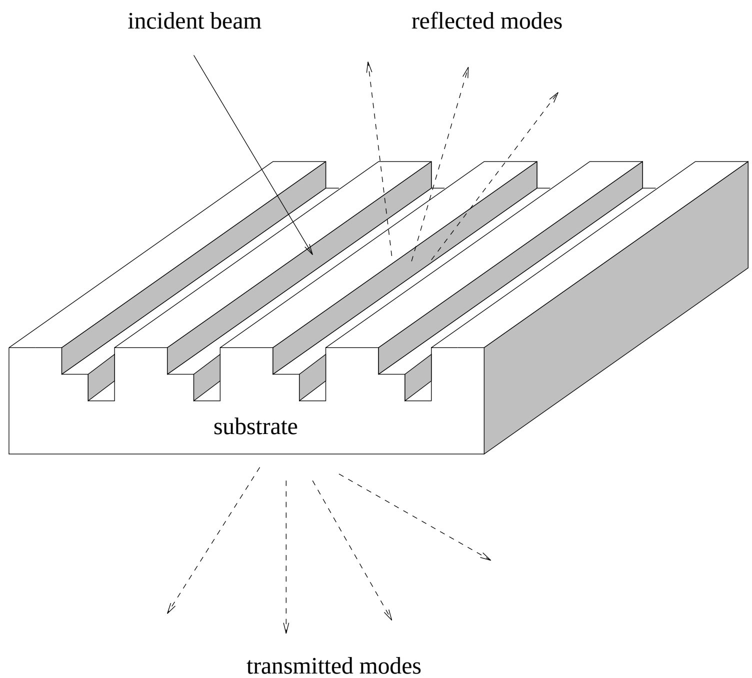

FIG. 1. *Diffraction grating. The structure is assumed to be of infinite extent. The period of the grating is generally comparable to the wavelength of the incident beam.*

The paper proceeds as follows. In Section 2 we briefly review a variational formulation of the Helmholtz equation in a periodic geometry. We then describe a formulation of the optimal design problem as a minimization, taking manufacturing considerations into account. With derivative-based minimization algorithms in mind, in Section 4 we study the continuity and differentiability of the design cost functional. We then discuss some practical issues in the implementation of computational algorithms, and finally present some numerical results in Section 6.

2. The direct diffraction problem. In this section we outline a simple variational formulation for the direct diffraction problem. More detailed accounts can be found for example in [2, 9].

We restrict attention to the simplest geometrical setting, where the diffractive structure is constant in one direction, as in Figure 1. We wish to model the propagation of time-harmonic electromagnetic waves in such a structure. Let $k_1^2 = \omega^2\epsilon_1\mu$ and $k_2^2 = \omega^2\epsilon_2\mu$, where ω is the frequency, μ is the magnetic permeability (assumed to be constant), and ϵ_1 and ϵ_2 are the constant dielectric coefficients in the transmission medium and substrate material, respectively. The refractive index constants k_1, k_2 can be complex in general (nonzero imaginary part corresponds to absorbing material), however for simplicity we assume here that both k_1 and k_2 are real.

Denote points in $\mathbb{R}^3$ by $x = (x_1, x_2, x_3)$ and take the direction of the grating "grooves" to be $\vec{x}_3$. Then the refractive index $k(x)$ is a function of (x_1, x_2) only. The periodicity implies that

$$k(x_1 + nL, x_2) = k(x_1, x_2), \quad \text{for all } (x_1, x_2),$$

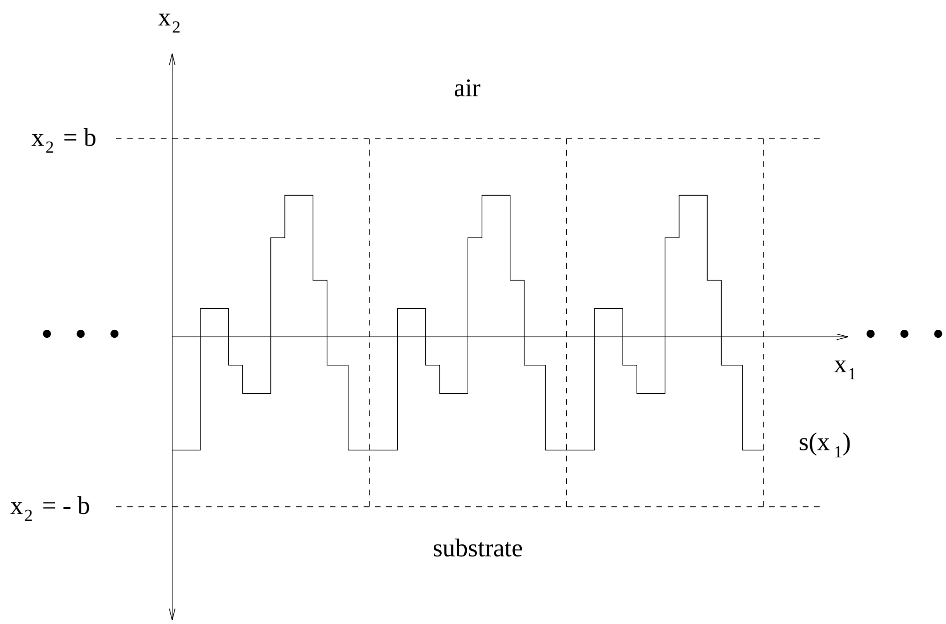

FIG. 2. *Geometrical configuration. The profile $s(x_1)$ is assumed to satisfy $\|s\|_{L^\infty} < b$ for some b.*

where L is the period, and n is any integer. By rescaling the problem if necessary, the period L is taken to be 2π.

We assume that the electric field is parallel to the x_3-axis. Maxwell's equations then reduce to the Helmholtz equation

$$(\triangle + k^2)u = 0 \ , \quad \text{in } \mathbb{R}^2 \ , \tag{1}$$

where u is the electric field vector component in the x_3 direction. This "E-parallel" polarization mode is the simplest case. Other polarization modes are also of interest but will not be treated here.

Let $s \in L^\infty(\mathbb{R})$ be a 2π-periodic function. The graph of s represents the interface between the transmission medium and the substrate material. We define the squared refractive index $a_s(x) = k^2(x)$ a.e. in $\mathbb{R}^2$ by

$$a_s(x) = \begin{cases} k_1^2 & \text{if } x_2 > s(x_1), \\ k_2^2 & \text{if } x_2 \leq s(x_1). \end{cases}$$

See Figure 2. The Helmholtz equation (1) can then be written $(\triangle + a_s)u = 0$. Assume that an incoming plane wave

$$u_* = e^{i\alpha x_1 - i\beta_1 x_2}$$

is incident on the interface from above ($x_2 = +\infty$). Here

$$\alpha = k_1 \sin\theta, \quad \beta_1 = k_1 \cos\theta,$$

and $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$ is the angle of incidence with respect to the x_2-axis.

We are interested in "quasiperiodic" solutions, that is, solutions u such that $u_\alpha \equiv ue^{-i\alpha x_1}$ is 2π-periodic. It is easily seen that if u satisfies $(\triangle + a_s)u = 0$, then u_α satisfies

$$(\triangle_\alpha + a_s)u_\alpha = 0 \quad \text{in } \mathbb{R}^2, \tag{2}$$

where the operator $\triangle_\alpha$ is defined by

$$\triangle_\alpha = \triangle + 2i\alpha\partial_1 - |\alpha|^2.$$

We will henceforth deal only with solutions of equation (2), and so we drop the subscript α. Since both u and a_s are now 2π-periodic in the x_1 variable, the problem is reduced to solving (2) with periodic boundary conditions in x_1. Equivalently, we consider (2) on the quotient space $Q = \mathbb{R}^2/\{2\pi Z \times \{0\}\}$, where $Z = \{0, \pm 1, \pm 2, \ldots\}$. Let $b > \|s\|_{L^\infty}$ and define the periodic strip $\Omega = \{x \in Q : -b < x_2 < b\}$. Define the boundaries $,_1 = \{x \in Q : x_2 = b\}$, and $,_2 = \{x \in Q : x_2 = -b\}$.

Our intention is to reformulate the diffraction problem on the domain Ω. This requires appropriate boundary conditions on $,_1$ and $,_2$. We make use of "exact" nonlocal boundary operators T_j^α, $j = 1, 2$, defined by

$$(T_j^\alpha f)(x_1) = (-1)^{j+1} \sum_{n \in Z} i\beta_j^n(\alpha) f_n e^{inx_1}, \quad j = 1, 2, \tag{3}$$

where f_n are the Fourier coefficients of f and

$$\beta_j^n = \begin{cases} \sqrt{k_j^2 - (n+\alpha)^2} & \text{if } k_j \geq |n+\alpha|, \\ i\sqrt{(n+\alpha)^2 - k_j^2} & \text{if } k_j < |n+\alpha|. \end{cases} \tag{4}$$

For a complete derivation of these operators, see eg. [9]. To exclude "resonances" where the propagation constants β_j^n are zero, we assume that $k_j^2 \neq (n+\alpha)^2$ for all $n \in Z$, $j = 1, 2$. It can be shown that each $T_j^\alpha : H^{\frac{1}{2}}(,_j) \to H^{-\frac{1}{2}}(,_j)$ is continuous, and

$$T_j^\alpha(u|_{\Gamma_j}) = \frac{\partial u}{\partial x_2}|_{\Gamma_j}, \quad j = 1, 2.$$

In other words, T_j^α is a Dirichlet–Neumann map. The operators T_j^α define "transparent" boundary conditions on $\partial\Omega$.

The scattering problem can be formulated as follows: find $u \in H^1(\Omega)$ such that

$$(\triangle_\alpha + a_s)u = 0 \text{ in } \Omega, \tag{5}$$

$$(T_1^\alpha - \frac{\partial}{\partial x_2})u = 2i\beta_1 e^{-i\beta_1 b} \text{ on } ,_1, \tag{6}$$

$$(T_2^\alpha - \frac{\partial}{\partial x_2})u = 0 \text{ on } ,_2. \tag{7}$$

The right-hand side in equation (6) is due to the incident wave. Conditions (6) and (7) implicitly incorporate an "outgoing wave condition" by the construction of the T_j^α operators.

It is convenient to deal with the variational form of (5)–(7). Defining

$$B_s(u,v) \equiv \int_\Omega (\nabla + i\alpha)u \cdot \overline{(\nabla + i\alpha)v} - \int_\Omega a_s\, u\overline{v} - \int_{\Gamma_1} (T_1^\alpha u)\overline{v} - \int_{\Gamma_2} (T_2^\alpha u)\overline{v} \tag{8}$$

where $\int_{\Gamma_j}$ represents the dual pairing of $H^{-\frac{1}{2}}(,_j)$ with $H^{\frac{1}{2}}(,_j)$, and

$$f(v) \equiv -2i\beta_1 e^{-i\beta_1 b} \int_{\Gamma_1} \overline{v},$$

the diffraction problem can be stated: find $u \in H^1(\Omega)$ which satisfies

$$B_s(u,v) = f(v), \quad \text{for all } v \in H^1(\Omega). \tag{9}$$

Theorem 2.1. *For k_1, $k_2 > 0$ sufficiently small, the variational problem (9) admits a unique weak solution $u \in H^2(\Omega)$. In addition, $\|u\|_{H^2(\Omega)}$ is bounded independently of s. More generally, a unique solution $u \in H^2(\Omega)$ exists for all but possibly a discrete set of parameters k_1, k_2.*

Proof. By a perturbation argument, it is proved in [9] that for $\|a_s\|_{L^\infty}$ sufficiently small, a unique solution $u \in H^1(\Omega)$ exists, with $\|u\|_{H^1(\Omega)}$ bounded independently of s. Actually, the slightly stronger statement that the operator A_s associated with the sesquilinear form B_s has a uniformly bounded inverse $A_s^{-1} : H^{-1}(\Omega) \to H^1(\Omega)$, is proved. We will need this fact in Section 4.

Since

$$\triangle u = (-2i\alpha\partial_1 + |\alpha|^2 - a_s)u$$

in Ω, it follows that $\|\triangle u\|_{L^2(\Omega)}$ is also bounded uniformly, proving the H^2 estimate.

The statement that existence and uniqueness of u is obtained for all but possibly a discrete set of parameters follows from Fredholm theory. A proof for the analogous problem involving the full Maxwell equations can be found in [10]. □

In the case of absorbing media, Abboud has shown [1] that the problem has a unique solution for all frequencies.

Since k_1 and k_2 scale with the frequency ω of the incident wave, the assumption that k_1 and k_2 are small can be viewed as a "low-frequency" assumption, which guarantees that the eigenvalue problem $\triangle_\alpha u = -a_s u$ (plus boundary conditions) has no nontrivial solutions. As a practical matter, for typical problems involving blazed diffraction gratings in dielectric materials, k_1 and k_2 are generally "small", so the theorem guarantees that solutions u exist for every s. However, resonances can occur over some parameter ranges. If uniqueness of solutions to (9) fails for some profile s, we will refer to that s as a resonant point. The results which follow require that resonant points are avoided.

3. An optimal shape design problem. A fundamental feature of diffraction in periodic structures is that in homogeneous regions, the scattered field can be expanded as an infinite sum of plane waves. For example, above Ω, the expansion can be written

$$u(x_1,x_2) = \sum_{n=-\infty}^{\infty} a_n e^{i(nx_1 + \beta_1^n x_2)}, \tag{10}$$

(often called the *Rayleigh expansion*), where the a_n are unknown coefficients, and β_1^n is defined in (4). A similar expansion holds in the region below Ω. Since β_1^n is real for at most a finite number of indices n, we see from (10) that only a finite number of plane waves in the sum propagate into the far field, with the remaining modes decaying exponentially as $x_2 \to +\infty$. The number of propagating modes and the direction of propagation for each mode is determined by the frequency of the incident wave, the refractive index of the material, and the period L of the structure.

In the optics literature, the ratio of the energy in a given propagating mode to the energy of the incoming wave is called the *efficiency* of the mode. From an engineering point of view, the key feature of any diffraction grating is the efficiency of each propagating mode.

Suppose that the materials, the period of the structure, and the frequency of the incoming waves are fixed. There are then a fixed number of propagating modes, each of which corresponds to an index n for which the propagation constant β_j^n is real-valued. Let us define two sets of indices of propagating modes

$$\Lambda_j = \{n \text{ integer} : |n + \alpha| < k_j\}, \quad j = 1, 2.$$

Thus Λ_1 contains the indices of the reflected propagating modes; Λ_2 corresponds to the transmitted modes. Up to a constant phase factor which can be safely ignored, the coefficients of each propagating reflected mode are given by the Fourier components of the trace of u on the boundary , ${}_1$:

$$\begin{aligned} r_n &= u_n(b) && \text{for } n \neq 0, \quad n \in \Lambda_1, \\ r_0 &= u_0(b) - const. && \text{for } n = 0, \end{aligned} \tag{11}$$

where $u_n(b) = \frac{1}{2\pi}\int_0^{2\pi} u(x_1, b)e^{-inx_1}dx_1$. Similarly, the coefficients of the propagating transmitted modes are

$$t_m = u_m(-b) \quad \text{for } m \in \Lambda_2. \tag{12}$$

Writing the reflection and transmission coefficients as vectors

$$r = (r_n)_{n\in\Lambda_1}, \quad t = (t_m)_{m\in\Lambda_2},$$

we denote the pair $(r, t) = F$. The coefficients r_n and t_m, and hence F, are functions of the interface profile s. Denote this dependence by $F(s)$. We index the individual reflection and transmission coefficients in F by F_j, $j = 1, \ldots, p$, where p is the total number of indices in Λ_1 and Λ_2.

As mentioned above, in applications one is usually much more concerned with the intensity or energy $|F_j|^2$ of the propagating modes than with the phases. The optimal design problem we wish to consider is to find a profile s such that $F(s)$ is as close as possible to some specified intensity pattern $q \in \mathbb{R}^p$. Asking that the intensities $|F_j(s)|^2$ are close to q in a least-square sense, we attempt to minimize

$$J(s) = \frac{1}{4}\sum_{j=1}^{p}(|F_j(s)|^2 - q_j)^2.$$

over some set of admissible coefficients $s \in \mathcal{A}$. The choice of $\mathcal{A}$ is important. To achieve a well-posed optimization problem, there are two general routes.

The first route is to begin with a large class of admissible curves, and "relax" the problem, enlarging the admissible set to include appropriate "mixtures" of materials. Admissible designs no longer correspond to distinct interface curves. See [2, 9] for a discussion of this approach.

Another natural route is to instead restrict the admissible interfaces s to a smaller class. Simply asking that the profiles stay within the specified domain Ω and away from the boundary suggests the problem

$$\min J(s) = \frac{1}{4}\sum_{j=1}^{p}(|F_j(s)|^2 - q_j)^2, \quad \text{subject to: } \|s\|_{L^\infty} \leq b' < b. \tag{13}$$

Without further constraints on s, one would not expect that a solution to (13) would exist in general, due to the possibility of oscillatory minimizing sequences. Unfortunately, "smoothness" constraints on s would not be appropriate, since mask-etch type manufacturing processes generally produce profiles with corners. Requiring that s is smooth eliminates manufacturable designs from the admissible class.

One convenient quantity which measures oscillations but allows corners is the total variation seminorm

$$TV(s) \equiv \int |s'(x_1)|\, dx_1. \tag{14}$$

Following common practice, in this definition the derivative s' is interpreted in a distributional sense, and does not have to be an L^1 function (see eg. [13]).

We are thus led to consider the problem

$$\begin{aligned} \min J(s) \quad &= \quad \frac{1}{4}\sum_{j=1}^{p}(|F_j(s)|^2 - q_j)^2, \\ \text{subject to:} \quad & \quad TV(s) \leq M, \\ & \quad \|s\|_{L^\infty} \leq b' < b. \end{aligned} \tag{15}$$

The following lemma can be proved by a straightforward calculation.

Lemma 3.1. *For any s_1, s_2 with $\|s_j\|_{L^\infty(0,2\pi)} \leq b'$, and for $1 \leq p < \infty$,*

$$\|a_{s_1} - a_{s_2}\|_{L^p(\Omega)} = C\|s_1 - s_2\|^{1/p}_{L^1(0,2\pi)}.$$

We denote the set of functions of bounded variation by $BV = \{s \in L^1 : \|s\|_{L^1} + TV(s) < \infty\}$.

Theorem 3.2. *For k_1 and k_2 sufficiently small, the constrained minimization problem (15) admits a solution $s \in BV$.*

Proof. The proof is straightforward by the direct method. Consider the admissible set

$$\mathcal{A} = \{s : TV(s) \leq M,\ \|s\|_{L^\infty} \leq b'\}.$$

Let $\{s_n\} \in \mathcal{A}$ be a minimizing sequence for $J(s)$. By the compactness properties of BV (see eg. Section 5.2.3 of [13]) there exists a subsequence (which we still denote by $\{s_n\}$) and a function $s \in \mathcal{A}$ such that $s_n \to s$ in L^1. From Lemma 3.1 above, this implies that the corresponding index functions a_{s_n} converge weak $*$ L^∞ to a_s.

Let u_n denote the solution to $B_{s_n}(u_n, v) = f(v)$. By Theorem 2.1, $\|u_n\|_{H^1}$ is uniformly bounded, so $\{u_n\}$ has a subsequence (still denoted u_n) converging weakly in H^1 and strongly in L^2 to some $u \in H^1$. Holding $v \in H^1$ fixed, it follows that

$$B_{s_n}(u, v) \to B_{s_n}(u_n, v) = f(v).$$

The weak$*$ L^∞ convergence of $a_{s_n} \rightharpoonup a_s$ implies $B_{s_n}(u, v) \to B_s(u, v)$. Hence $B_s(u, v) = f(v)$ for all $v \in H^1$, i.e., u solves (9).

Since $u_n \rightharpoonup u$ weakly in H^1, $u_n|_{\Gamma_j} \rightharpoonup u|_{\Gamma_j}$ weakly in $H^{1/2}(, _j)$. Hence any finite set of Fourier coefficients of $u_n|_{\Gamma_j}$ converge strongly to the Fourier coefficients of $u|_{\Gamma_j}$. By the definition of $F(s)$, we thus have $F(s_n) \to F(s)$, and s is a minimizer of $J(s)$. □

4. Evaluation of derivatives. Most simple schemes that we would consider for minimizing $J(s)$ require some derivative information. For the sake of simplicity, assume for now that the total variation constraint is not enforced.

To study the continuity of solutions u with respect to s, we need to be able to evaluate "traces" of functions on the graph of s. Note that the following notion of trace differs from the usual one in that the measure on the graph of s is not the surface measure, but simply Lebesgue measure on the domain $[0, 2\pi)$. If surface measure were used, a Lipschitz assumption on s would be necessary as in standard trace theorems.

Lemma 4.1. *Given $s \in L^\infty(0, 2\pi)$ with $\|s\|_{L^\infty} \le b' < b$, there exists a bounded linear operator*

$$M_s : H^1(\Omega) \to L^p(0, 2\pi),$$

for $2 \le p < \infty$, such that

$$(M_s f)(x_1) = f(x_1, s(x_1)), \quad a.e.\ \ x_1 \in [0, 2\pi), \tag{16}$$

for all $f \in C^1(\Omega) \cap H^1(\Omega)$.

Proof. Assume $f \in C^1(\Omega) \cap H^1(\Omega)$. Define $M_s f$ as in (16). Then

$$\int_0^{2\pi} |M_s f(x_1)|\, dx_1 \le \int_0^{2\pi} |f(x_1, -b')|\, dx_1 + \int_0^{2\pi} \int_{-b'}^{b'} |\partial_2 f(x_1, t)|\, dt\, dx_1.$$

Replacing $|f|$ by $|f|^p$,

$$\begin{aligned}
\int_0^{2\pi} |M_s f(x_1)|^p\, dx_1 &\le \int_0^{2\pi} |f(x_1, -b')|^p\, dx_1 \\
&\quad + C \int_0^{2\pi} \int_{-b'}^{b'} |\partial_2 f(x_1, t)||f(x_1, t)|^{p-1}\, dt\, dx_1 \\
&\le \|T_{-b} f\|^p_{L^p(0,2\pi)} + C\|\partial_2 f\|_{L^2(\Omega)} \|f\|^{p-1}_{L^{2(p-1)}(\Omega)} \\
&\le C\|f\|^p_{H^1(\Omega)}.
\end{aligned}$$

where T_{-b} is the usual trace operator on $\{x_2 = -b'\}$. The last inequality follows from standard trace and Sobolev imbedding theorems. Since C^1 is dense in H^1, M_s extends uniquely to a bounded linear operator from $H^1(\Omega)$ to $L^p(0, 2\pi)$. □

Let $\delta s \in L^\infty$ be a given perturbation to the fixed profile s. For notational convenience, we will henceforth let

$$\begin{aligned} &u_0 \quad \text{satisfy} \quad B_s(u_0, v) = f(v), \\ &u_1 \quad \text{satisfy} \quad B_{s+\delta s}(u_1, v) = f(v), \end{aligned} \tag{17}$$

for all $v \in H^1(\Omega)$, whenever solutions exist.

The solutions u are Lipschitz continuous with respect to s, as we show next.

Theorem 4.2. *Assume that k_1 and k_2 are small enough so that Theorem 2.1 holds. Given s with $\|s\|_\infty \le b' < b$, consider perturbations $\delta s \in L^\infty(0, 2\pi)$ such that $\|s + \delta s\|_\infty \le b'$. Then*

$$\|u_1 - u_0\|_{H^1(\Omega)} \le C\|\delta s\|_{L^2(0,2\pi)}.$$

Proof. One easily finds that

$$B_{s+\delta s}(u_1 - u_0, v) = \int_\Omega (a_{s+\delta s} - a_s) u_0 \overline{v}. \tag{18}$$

As mentioned in the proof of Theorem 2.1, under the low-frequency hypothesis, the operator A_s associated with the quadratic form B_s has a uniformly bounded inverse $A_s^{-1} : H^{-1}(\Omega) \to H^1(\Omega)$ independent of s. Hence a bound on $\|u_1 - u_0\|_{H^1}$ will follow immediately from estimating the H^{-1} norm of the right-hand side of (18), i.e.

$$\sup_{\|v\|_{H^1(\Omega)}=1} \left| \int_\Omega (a_{s+\delta s} - a_s) u_0 \overline{v} \right|$$

Since $u_0 \in H^2(\Omega)$, it is easy to estimate

$$\|u_0 \overline{v}\|_{H^1(\Omega)} \le C\|u_0\|_{H^2(\Omega)} \|v\|_{H^1(\Omega)}.$$

Set $z = u_0\overline{v}$. Assume temporarily that $z \in C^1(\Omega) \cap H^1(\Omega)$. We have

$$\begin{aligned} \left|\int_\Omega (a_{s+\delta s} - a_s)\, z\right| &= C\left|\int_0^{2\pi} \int_{s(x_1)}^{(s+\delta s)(x_1)} z(x_1, x_2)\, dx_2\, dx_1\right| \\ &= C\left|\int_0^{2\pi} \int_{s(x_1)}^{(s+\delta s)(x_1)} \left(z(x_1, s(x_1)) + \int_{s(x_1)}^{x_2} \partial_2 z(x_1, t)\, dt\right) dx_2\, dx_1\right| \\ &\le C\int_0^{2\pi} |\delta s(x_1)| \left(|z(x_1, s(x_1))| + \left|\int_{s(x_1)}^{(s+\delta s)(x_1)} |\partial_2 z(x_1, t)|\, dt\right|\right) dx_1 \\ &\le C\|\delta s\|_{L^2(0,2\pi)}(\|M_s z\|_{L^2(0,2\pi)} + \|z\|_{H^1(\Omega)}) \\ &\le C\|\delta s\|_{L^2(0,2\pi)} \|z\|_{H^1(\Omega)}, \end{aligned}$$

where we applied Lemma 4.1 to get the last inequality. Taking an approximating sequence $z_k \in C^1$ with $\|z_k - z\|_{H^1} \to 0$, the estimate is established for general $z \in H^1$. □

Given $\delta s \in L^2(0, 2\pi)$, define δu as the solution of

$$B_s(\delta u, v) = g_s(\delta s, v), \quad \text{for all } v \in H^1(\Omega), \tag{19}$$

where

$$g_s(\delta s, v) = (k_2^2 - k_1^2) \int_0^{2\pi} \delta s(x_1)(M_s u_0 \overline{v})(x_1)\, dx_1.$$

Since $u_0 \in H^2(\Omega)$, we have $u_0\overline{v} \in H^1(\Omega)$ with $\|u_0\overline{v}\|_{H^1(\Omega)} \le C\|u_0\|_{H^2(\Omega)}\|v\|_{H^1(\Omega)}$. Note that Lemma 4.1 implies that

$$|g_s(\delta s, v)| \le C\|\delta s\|_{L^2}\|v\|_{H^1(\Omega)},$$

where C depends on s. The boundedness of the solution operator associated with B_s then immediately implies that a unique solution $\delta u \in H^1(\Omega)$ exists. The following statement is now proved.

Lemma 4.3. *If s is not a resonance point, then there exists a unique $\delta u \in H^1(\Omega)$ satisfying (19). Furthermore,*

$$\|\delta u\|_{H^1(\Omega)} \le C\|\delta s\|_{L^2(0,2\pi)},$$

where C depends only on s.

We are now in a position to prove the differentiability of the map taking s to the field u. The functions u_0, u_1, and δu are all as previously defined.

Theorem 4.4. *Assume that k_1 and k_2 are small enough so that Theorem 2.1 holds. Given s with $\|s\|_\infty \le b' < b$, consider perturbations $\delta s \in L^\infty(0, 2\pi)$ such that $\|s + \delta s\|_\infty \le b'$. Then*

$$\|u_1 - u_0 - \delta u\|_{H^1(\Omega)} = o(\|\delta s\|_{L^\infty}).$$

Proof. Let $w = u_1 - u_0 - \delta u$. Then w satisfies

$$B_{s+\delta s}(w, v) = h(v) := \int_\Omega (a_{s+\delta s} - a_s)u_0\overline{v} - g_s(\delta s, v) + \int_\Omega (a_{s+\delta s} - a_s)\delta u\, \overline{v}. \tag{20}$$

Because of the uniform boundedness of the solution operators $A^{-1}_{s+\delta s}$ associated with $B_{s+\delta s}$, we can estimate

$$\|w\|_{H^1(\Omega)} \le C \sup_{\|v\|_{H^1(\Omega)}} |h(v)|.$$

Let us examine the last term in $h(v)$ first. We have by Lemmas 3.1, 4.3, and Sobolev imbedding that

$$\begin{aligned} \left|\int_\Omega (a_{s+\delta s} - a_s)\delta u\, \overline{v}\right| &\le \|a_{s+\delta s} - a_s\|_{L^2(\Omega)}\|\delta u\|_{L^4(\Omega)}\|v\|_{L^4(\Omega)} \\ &\le C\|\delta s\|^{3/2}_{L^2(0,2\pi)}\|v\|_{H^1(\Omega)}, \end{aligned} \tag{21}$$

which is $o(\|\delta s\|_{L^\infty})$.

This leaves us to estimate the first and second terms in the right-hand side of (20). Set $z = u_0\overline{v}$, and temporarily assume $z \in C^1(\Omega)\cap H^1(\Omega)$. Then proceeding in a manner similar to the proof of Theorem 4.2,

$$\begin{aligned}
|\int_\Omega (a_{s+\delta s} - a_s)u_0\overline{v} \; - \; g_s(\delta s, v)| & \\
&= |k_2^2 - k_1^2| \left| \int_0^{2\pi} \left(\int_{s(x_1)}^{(s+\delta s)(x_1)} z(x_1,x_2)\, dx_2 - \delta s(x_1)(M_s z)(x_1) \right) dx_1 \right| \\
&= C \left| \int_0^{2\pi} \left(\int_{s(x_1)}^{(s+\delta s)(x_1)} z(x_1,x_2) dx_2 - \delta s(x_1) z(x_1, s(x_1)) \right) dx_1 \right| \\
&= C \left| \int_0^{2\pi} \left(\int_{s(x_1)}^{(s+\delta s)(x_1)} \int_{s(x_1)}^{x_2} \partial_2 z(x_1,t)\, dt\, dx_2 \right) dx_1 \right| \\
&\leq C \int_0^{2\pi} |\delta s(x_1)| \left| \int_{s(x_1)}^{(s+\delta s)(x_1)} |\partial_2 z(x_1,t)|\, dt \right| \, dx_1 \\
&\leq C\|\delta s\|_{L^\infty} \int_\Omega (\chi_{s,\delta s}) |\partial_2 z|,
\end{aligned}$$

where $\chi_{s,\delta s}$ denotes the characteristic function on the set where a_s and $a_{s+\delta s}$ differ. Applying the Schwarz inequality and Lemma 3.1 we then have

$$|\int_\Omega (a_{s+\delta s} - a_s)u_0\overline{v} - g_s(\delta s, v)| \leq C\|\delta s\|_{L^\infty} \|\delta s\|_{L^1}^{1/2} \|z\|_{H^1(\Omega)}.$$

The bound holds for arbitrary $z \in H^1(\Omega)$ by taking an approximating sequence $z_k \in C^1$ converging to z in H^1. Finally, we recall that $\|z\|_{H^1} \leq \|u_0\|_{H^2}\|v\|_{H^1}$. □

Having calculated the derivative of u with respect to s, it is now a simple matter to calculate the gradient of $J(s)$ for use in minimization schemes. Let us view $J(s)$ as a map over some subset D of $L^2(0,2\pi)$. Consider the linearized response $DJ(s)(\delta s)$ of $J(s)$ to a perturbation δs. Formally,

$$DJ(s)(\delta s) = \sum_{j=1}^{p} \left(|F_j(s)|^2 - q_j \right) Re \left(\overline{F_j(s)}\; DF_j(s)(\delta s) \right)$$

where $DF(s)(\delta s)$ denotes the linearization of F in the direction δs. The components of DF are the Fourier modes (Dr, Dt), where

$$\begin{aligned}
Dr_m(s)(\delta s) &= \delta u_m(b), \quad \text{for reflected modes } m \in \Lambda_1 \\
Dt_n(s)(\delta s) &= \delta u_n(-b), \quad \text{for transmitted modes } n \in \Lambda_2.
\end{aligned}$$

Viewing the domain of DF as $L^2(0,2\pi)$, the adjoint $DF^*(s)(\cdot)$ is defined by

$$DF(s)(\delta s) \cdot \overline{\delta q} = \int_0^{2\pi} \delta s \cdot \overline{DF^*(s)(\delta q)}\, dx_1,$$

for $\delta q = (\psi, \phi)$ with $\psi = (\psi_n)_{n\in\Lambda_1}$ and $\phi = (\phi_m)_{m\in\Lambda_2}$. Let $w \in H^1(\Omega)$ solve

$$(\triangle_\alpha + a_s)w = 0 \quad \text{in } \Omega, \tag{22}$$

$$(T_1^* - \frac{\partial}{\partial\nu})w = -\frac{e^{i\beta_1 b}}{2\pi}\sum_{n\in\Lambda_1}\psi_n e^{inx_1} \quad \text{on } ,\ _1, \tag{23}$$

$$(T_2^* - \frac{\partial}{\partial\nu})w = -\frac{e^{i\beta_2 b}}{2\pi}\sum_{m\in\Lambda_2}\phi_m e^{imx_1} \quad \text{on } ,\ _2, \tag{24}$$

where $T_j^* f = -\sum i\bar{\beta}_j^n f_n e^{inx}$. Notice that this adjoint problem for w represents waves propagating *into* Ω. With an integration by parts calculation, one finds that

$$DF(s)(\delta s)\cdot\overline{\delta q} = (k_2^2 - k_1^2)\int_0^{2\pi}\delta s(x_1)\ M_s(\overline{w}u)(x_1)\ dx_1.$$

Hence we make the identification $\overline{DF^*(s)(\delta q)}(x_1) = (k_2^2 - k_1^2)M_s(\overline{w}u)(x_1)$, and the gradient of $J(s)$ is given by $G(s) = Re\ \{DF^*(s)(F_j(|F_j(s)|^2 - q_j))\}$. More explicitly,

$$G(s)(x_1) = Re\{(k_2^2 - k_1^2)M_s(\overline{w}u)\}(x_1),$$

where w solves (22)–(24) with $(\psi,\phi)_n = F_n(|F_n(s)|^2 - q_n)$.

Note that by Lemma 4.1, $G(s) \in L^p(0, 2\pi)$, for $2 \le p < \infty$. In fact, since both w and u are in H^2, $\overline{w}u \in H^2$, and it follows that $G(s) \in L^\infty$.

5. Computational issues. The variational formulation of the diffraction problem described in Section 2 leads naturally to finite element schemes. Bao [4] has studied the convergence of such schemes for problem (9) and has established "optimal" error estimates. A preconditioner to improve the speed of iterative solution schemes for the finite element system is described in [11]. The primary practical advantage of finite element solutions of (9) is that convergence is very good even for extremely rough s.

Discretizing Ω with a uniform rectangular grid, one can approximate the interface profile as the graph of a sum of step functions

$$s(x_1) = \sum_{j=1}^{N} s_j\chi_j \tag{25}$$

where χ_j is the indicator function in the interval $[(j-1)h, jh)$ and h is the grid spacing in the x_1 direction. Notice that the results in the previous section hold for such s, and in fact each such $s \in BV$.

In order to maintain a fixed, uniform grid as s is varied, one can simply average the squared refractive index a_s in cells where s_j does not lie on a grid line (see Figure 3). From the proof of Theorem 3.2 we know that F is continuous with respect to a_s in the weak$*$ $L^\infty(\Omega)$ topology, and it follows that by calculating cell-wise averages of a_s, we still get a convergent approximation to F as the grid is refined.

Using the formulae for derivatives and gradients established in the previous section, one can approach the numerical minimization of $J(s)$ in a variety of ways. The most

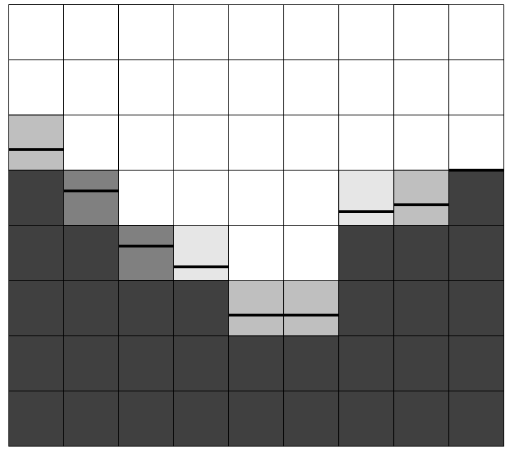

FIG. 3. *Averaging a_s to maintain a uniform grid. Heavy lines indicate the location of $s(x_1)$; thin lines indicate the fixed grid.*

straightforward is the method of projected gradient descent (gradient descent with each step constrained to lie inside the computational box). The method is effective, but quite inefficient: each gradient step requires solving both a direct and an adjoint diffraction problem.

A modified gradient descent algorithm, designed to improve the efficiency of the basic gradient descent scheme for a similar problem, is described in [12]. The basic idea is to view the underlying diffraction problem as a constraint, and then perform inexact solves following infeasible point techniques from constrained optimization. To see the constrained formulation for the present problem, let us modify the statement of the minimization problem (13). Define the linear operator $P : H^1(\Omega) \to \mathbb{R}^p$ by $P(u) = (r, t)$ where the reflection and transmission vectors r and t are defined by (11)–(12). The functional $\sum(|P_j u|^2 - q_j)^2$ is then defined over all $H^1(\Omega)$. One can then reformulate the minimization problem (without the TV constraint) as

$$\min_{(s,u)\in\mathcal{A}\times H^1(\Omega)} \quad \sum_j (|P_j u|^2 - q_j)^2 \tag{26}$$

$$\text{subject to} \quad B_s(u, v) = f(v), \quad \text{for all } v \in H^1(\Omega). \tag{27}$$

Viewing the problem in this way, one can iteratively try to minimize the cost functional while staying outside the constraint set, requiring that the constraints are satisfied only when convergence is achieved. In this way, the direct and adjoint diffraction problems only need to be solved to some tolerance, initially large but controlled by the algorithm and forced to zero as the algorithm converges.

Considerably more sophisticated minimization algorithms for general optimal control and design problems with PDE state equations, based on constrained formulations of the form (26)–(27) have been developed and analyzed by Dennis, Heinkenschloss, and Vicente [8].

Finally we mention that a simple technique to approximately solve the TV-constrained problem (15) using conventional unconstrained minimization techniques involves replacing the constraint $TV(s) \leq M$ with a penalty in the cost functional of the form

$$R_{\beta,\gamma}(s) = \beta \int_0^{2\pi} \sqrt{s'(x_1)^2 + \gamma^2}\, dx_1.$$

For $\gamma > 0$, R is a smooth function of s, so unconstrained gradient and Newton based algorithms can be applied. As $\gamma \to 0$, R approaches the TV measure, and convergence of the minimization algorithm generally deteriorates due to the nonsmoothness of $TV(s)$ with respect to s. Fortunately, acceptable designs can usually be found with γ fixed at some small number. The parameter $\beta \geq 0$ must be adjusted to obtain an acceptable level of oscillation in s, with $TV(s) \leq M$.

6. Numerical experiments. In this section we describe some numerical results. In all experiments, the problem is discretized as described in the previous section, on a 128×128 grid. Some earlier results were presented in [5, 11]. In each example, we solve only for *local* minima of the cost functional $J(s)$. This is appropriate since in engineering applications a good starting guess is often available.

We first demonstrate that some optimal profiles naturally have "corners", and are of bounded variation, without an explicit TV constraint imposed on the problem. Let a plane wave be normally incident on a periodic interface, say with period $L = 1.5$, between air and quartz (glass). Consider the problem of *maximizing* the total energy reflected from the interface. This problem is easily placed in the framework described in the previous sections by minimizing $-J(s)$ instead of $J(s)$, with $q = 0$ and the transmissive orders weighted by zero. Figures 4 through 6 illustrate the solutions obtained for two incident waves of wavelengths 0.71, and 0.48. Each optimization was solved without total variation penalties or constraints, using a sine wave profile s_0 as the starting point. In each case, two reflected modes exist which propagate at (symmetric) angles near horizontal. Nearly all of the reflected energy is carried by these modes. We should note that structures with higher reflectivity can be constructed; this example was chosen merely to illustrate the structure of the solution profiles.

With a final example we wish to illustrate that in some cases explicit regularization is necessary to obtain a solution. As before, we take a plane wave normally incident on an air-glass interface with period $L = 1.5$. The incident wavelength is set at 0.34. We specify that all energy is transmitted through the structure (no reflection), with 50% of the energy in the $+1$ transmitted order, and 50% in the -1 transmitted order. Running the minimization with no regularization generates a minimizing sequence of increasingly oscillatory profiles s_n. The profile s_{200} shown in Figure 7 achieves approximately 48.7% energy in each order. Continuing the minimization monotonically decreases the cost functional, and monotonically increases the total variation as the profiles become rougher and more oscillatory. Presumably, if one were to "relax" the problem by passing to a weak limit, the optimal solution would be a mixture.

Enforcing a total variation constraint by means of a penalty function as described in the previous section leads to a convergent sequence s_n. A solution is shown in Figure 8. When enforcing the constraint, there is a tradeoff between the "simplicity" of the solution (measured by total variation), and the amount of energy achieved in each order. This tradeoff may be considered as a design parameter.

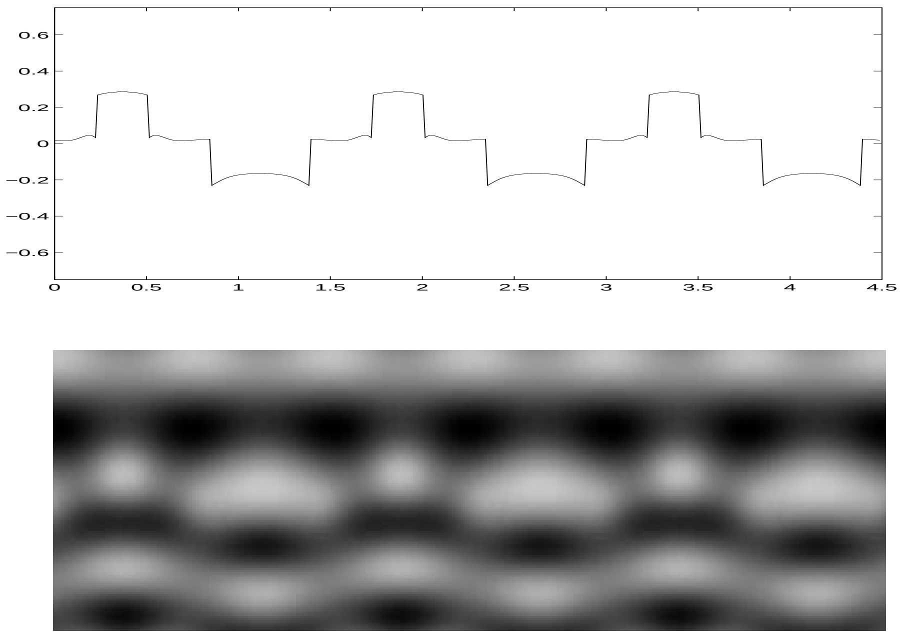

FIG. 4. *Profile producing a local maximum in reflected energy, and the real part of the diffracted field. The incident wavelength is 0.71; total reflected energy is about* 7.6%.

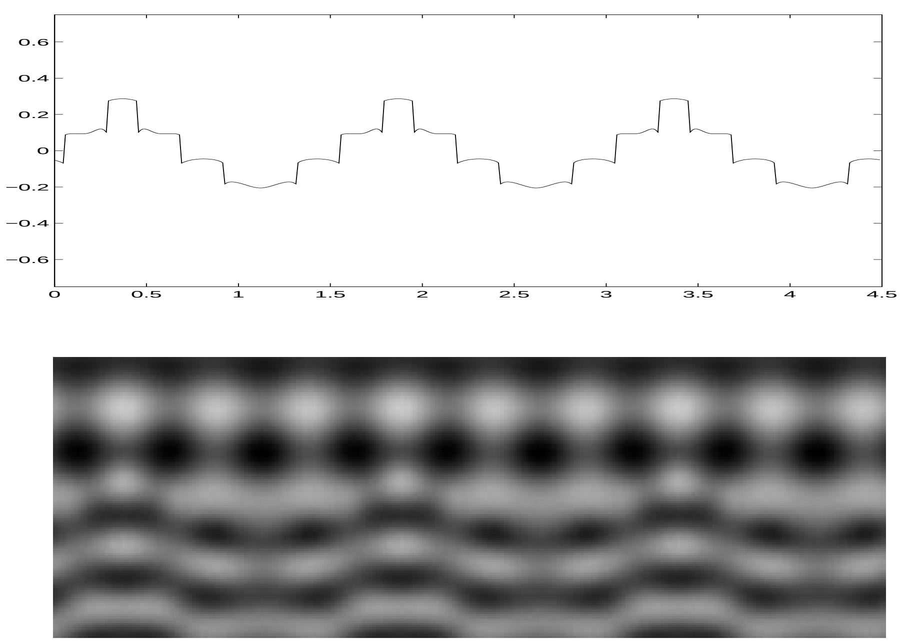

FIG. 5. *Profile producing a local maximum in reflected energy, and the real part of the diffracted field. The incident wavelength is 0.48, total reflected energy is about* 7.3%.

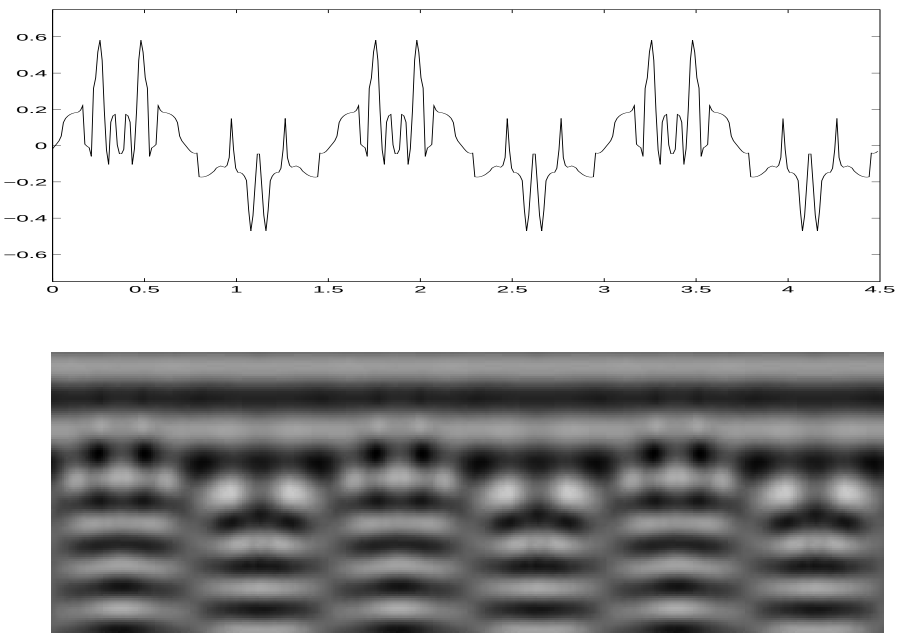

FIG. 6. *One profile in a minimizing sequence generated without a total variation constraint, and the real part of the diffracted field. The profile directs approximately* 97% *energy into the* ±1 *orders. The profile's total variation is 32.9*

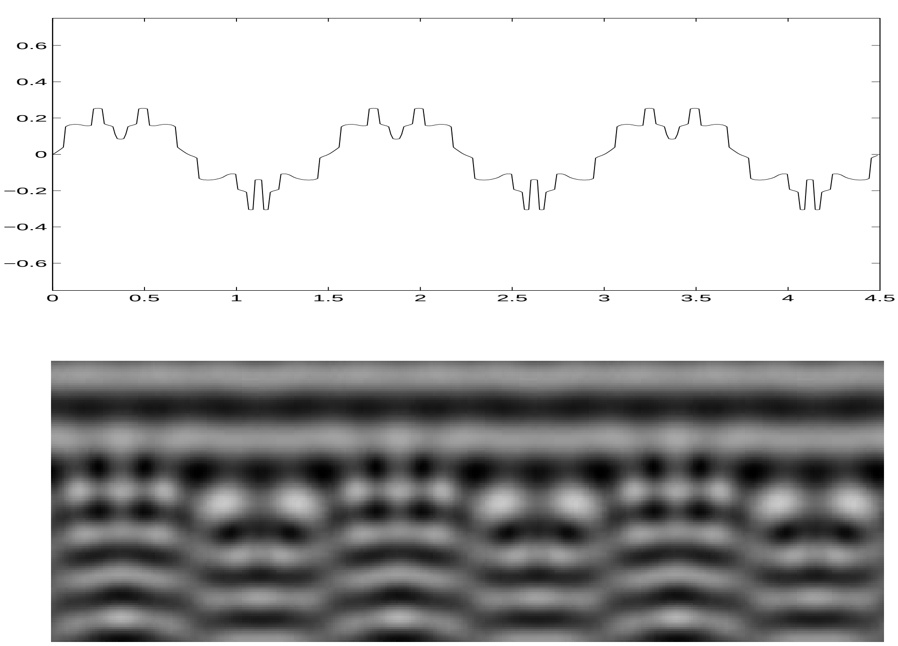

FIG. 7. *A solution obtained with the total variation constraint* $TV(s) \leq 9$ *enforced. The profile directs approximately* 91% *energy into the* ±1 *transmitted orders.*

Several other examples, as well as a discussion on producing "manufacturable" designs, can be found in [5, 11].

Acknowledgement and Disclaimer. Effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0497. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

References

[1] T. Abboud, *Electromagnetic waves in periodic media*, in *Second International Conference on Mathematical and Numerical Aspects of Wave Propagation*, edited by R.Kleinman et al, SIAM, Philadelphia, 1993, 1–9.

[2] Y. Achdou, *Optimization of a photocell*, Opt. Control Appl. Math., 12 (1991), 221–246.

[3] H. Ammari, *Uniqueness theorems for an inverse problem in a doubly periodic structure*, Inverse Problems **11** (1995), 823–833.

[4] G. Bao, *Finite elements approximation of time harmonic waves in periodic structures*, SIAM J. on Numerical Analysis, to appear.

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

[6] G. Bao and A. Friedman, *Inverse problems for scattering by periodic structures*, Arch. Rat. Mech. Anal. **132** (1995), 49–72.

[7] M. Cadilhac, *Rigorous vector theories of diffraction gratings*, in Progress in Optics **21**, E. Wolf, ed., North-Holland, New York, 1984.

[8] J.E. Dennis, M. Heinkenschloss, and L. Vicente, *Trust-region interior-point algorithms for a class of nonlinear programming problems*, Virginia Tech ICAM Report 94-12-01 (Revised April 1995).

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

[10] D.C. Dobson, *A variational method for electromagnetic diffraction in biperiodic structures*, RAIRO Modél. Math. Anal. Numér., **28** (1994), 419–439.

[11] D.C. Dobson, *Controlled scattering of light waves: optimal design of diffractive optics*, in "Control Problems in Industry" edited by I. Lasiecka and B. Morton, Birkhäuser, Boston (1995), pp. 97–118.

[12] D.C. Dobson, *Exploiting ill-posedness in the design of diffractive optical structures*, in "Mathematics in Smart Structures", edited by H.T. Banks, SPIE Proc. **1919** (1993), 248–257.

[13] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton (1992).

[14] A. Friedman, *Mathematics in Industrial Problems*, IMA Volume 16 (1988); Part 3, IMA Volume 38 (1991); Part7 (1994), Springer–Verlag, New York.

[15] A. Kirsch, *Uniqueness theorems in inverse scattering theory for periodic structures*, Inverse Problems **10** (1994), 145-152.

[16] *Diffractive and Miniaturized Optics: proceedings of a conference held 12–13 July 1993, San Diego, California*, SPIE volume CR-49, edited by S.H. Lee, SPIE, Bellingham (1993).

[17] *Electromagnetic Theory of Gratings,* Topics in Current Physics, Vol. 22, edited by R. Petit, Springer-Verlag, Heidelberg, 1980.