



UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION  
INTERNATIONAL ATOMIC ENERGY AGENCY  
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS  
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE



H4.SMR/1058-8

## WINTER COLLEGE ON OPTICS

9 - 27 February 1998

### *Bases of Diffractive Optics*

**P. Chavel**

**Laboratoire Charles Fabry de l'Institut d'Optique, CNRS, Orsay, France**

**DIFFRACTIVE OPTICS LECTURES AT THE THIRD ICTP/ICO WINTER COLLEGE, TRIESTE,  
FEBRUARY 1998.**

Pierre Chavel, Laboratoire Charles Fabry de l'Institut d'Optique, CNRS  
BP147, 91403 Orsay, France

**Lecture outline**

**Part I - The paraxial model of diffractive optics**

- Some motivations for diffractive optics.
- Transmission of a plane parallel plate, a lens, a prism : validity of the concept.
- Diffraction gratings : thin and thick, reflection and transmission, surface and volume : orders ; the Floquet theorem ; the two versions of the grating law ; diffraction efficiency in the scalar model.
- Gratings fabricated by resist masking.
- Zone plates, applications, difference with Fresnel lenses.
- Application : diffractive achromats.
- The concept of diffraction orders revisited.
- How many zones does it take for a component to be « diffractive? »

**Part II - Diffractive optics in the resonant regime**

- Diffraction efficiency and field calculations : coupled waves, modal theories, and surface coordinate transform method.
- Applications : diffraction efficiency of multiple masks gratings ; polarising beam splitters.

**Part III - Diffractive optics below the resonant regime**

- The concept of effective index. Asymptotic limit.
- Application to reflection control, photonic bandgap structures.
- Polarisation properties.
- Effective index media for wavefront control.

## Lecture documents

### **1) Exercise set :**

- E1 : diffraction efficiency of thin gratings in the paraxial regime.
- E2 : about prisms and blazed gratings (paraxial case).
- E3 : diffraction gratings fabricated by resist masking (paraxial case).
- E4 : achromatic doublets (paraxial case).
- E5 : zone plates (paraxial case).
- E6 : Floquet's theorem (TE case).
- E7 : the Kogelnik coupled wave model for thick transmission holograms (TE case, non slanted fringes)
- E8 : effective indices in the « quasi static » limit.

### **2) The MIT reports on diffractive optics applied to hybrid optical systems :**

- G.J. Swanson, MIT Lincoln Laboratory Technical Report 854, 1989 : « Binary optics technology : the theory and design of multi-level diffractive optical elements. »
- G.J. Swanson, MIT Lincoln Laboratory Technical Report 914, 1991 : « Binary optics technology : theoretical limits on the diffraction efficiency of multilevel diffractive optical elements. »

### **3) Some basic articles on the rigorous calculation of diffraction by thick gratings :**

- N. Chateau and J.P. Hugonin, J. Opt. Soc. Am. A11 (1994) 1321-1331, « Algorithm for the rigorous coupled wave analysis of grating diffraction. »
- F. Montiel and M. Nevière, J. Opt. Soc. Am. A11 (1994) 3241-3250. « Differential theory of gratings : extension to deep gratings of arbitrary profile and permittivity through the R-matrix propagation algorithm. »
- L. Li, J. Opt. Soc. Am. A13 (1996) 1024-1035, « Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings. »

### **4) Some basic articles on the rigorous calculation of diffraction by gratings with discontinuities :**

- Ph. Lalanne and G.M. Morris, J. Opt. Soc. Am. A13 (1996) 779-784, « Highly improved convergence of the coupled-wave method for TM polarization. »
- L. Li, J. Opt. Soc. Am. A13 (1996) 1870-1876, « Use of Fourier series in the analysis of discontinuous periodic structures. »

### **Additional reading : some classical references on grating diffraction : (those marked \* are available from the lecturer at Trieste)**

- two reviews that are still useful :

- R. Petit, ed. Electromagnetic Theory of Gratings, Springer Verlag, Berlin, 1980.
- T.K. Gaylord and M.G. Moharam, Proc. IEEE 73 (1985) 894-937, « Analysis and applications of optical diffraction by gratings. » Easier to find in an Engineering library than in a physics library.

- the C method (coordinate transformation for surface profile gratings) :

- J. Chandezon, M.T. Dupuis, G. Cornet and D. Maystre, J. Opt. Soc. Am. 72 (1982) 839-846, « Multicoated gratings : a differential formalism applicable in the entire optical region. »

- a basic reference on the coupled wave methods :

- M.G. Moharam and T.K. Gaylord, J. Opt. Soc. Am. 71 (1981) 811-818, « Rigorous coupled-wave analysis of planar-grating diffraction. »

\* a short course on hybrid diffractive optics in French :

P. Chavel, « L'optique diffractive au service de la correction des aberrations » in *Optique Instrumentale*, P. Boucharaine, editor, collection de la Société Française d'Optique, Editions de Physique, les Ulis, 1997, pp 249-259.

\* one more modern article on diffraction by thick gratings :

M.G. Moharam, D.A. Pommet, E.B. Grann and T.K. Gaylord, *J. Opt. Soc. Am. A12* (1995) 1075-1086. « Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings : enhanced transmittance matrix approach. »

\* one more modern article on gratings with discontinuities :

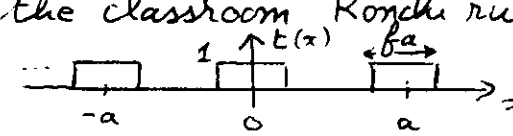
G. Granet and B. Guizal, *J. Opt. Soc. Am. A13* (1996) 1019-1023. « Efficient implementation of the coupled-wave method for metallic gratings in TM polarization. »

Exercise 1

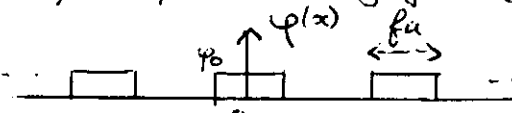
Diffraction efficiency of thin gratings  
in the paraxial regime

Consider a thin diffraction grating of pitch  $a$  with complex amplitude transmission  $t(x)$ . Its diffraction efficiency  $\eta_p$  in order  $p$  is known to be the squared modulus of the  $p$ -th Fourier coefficient in the F. series expansion of function  $t$ :

$$\eta_p = |t_p|^2 = \left| \frac{1}{a} \int_0^a t(x) e^{-2i\pi p \frac{x}{a}} dx \right|^2$$

1) Consider the classroom Ronchi ruling with duty cycle  $f$ : . Find the max. of  $\eta_p$ .

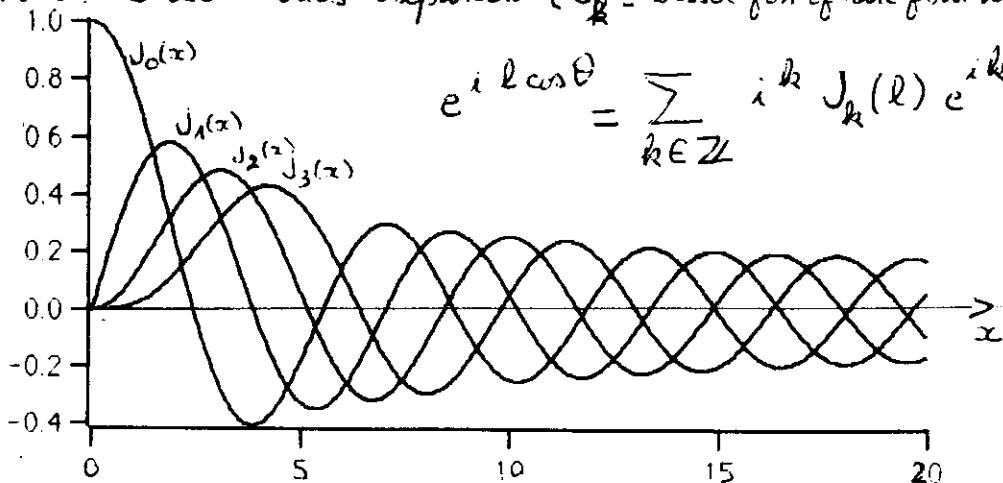
2) Consider the sinusoidal "amplitude only" ( $\varphi = \text{cst.}$ ) grating:  $t(x) = \frac{1}{2} (1 + \cos 2\pi \frac{x}{a})$ . Calculate all  $\eta_p$ .

3) Consider the pure phase binary grating with the following phase: . Find the max. of  $\eta_1$ .

4) Consider the sinusoidal pure phase grating:

$$\varphi(x) = \varphi_0 \cos 2\pi \frac{x}{a} \quad \text{Find the max. of } \eta_1$$

Note: Bessel series expansion ( $J_k$  = Bessel fun of the first kind, order  $k$ )



5) Show that no grating with even  $t(x)$  may exceed  $\eta_1 = 50\%$ . Show that no amplitude only grating may exceed  $\eta_1 = 50\%$

6) Which is the only thin grating that has  $\eta_1 = 100\%$ ?

Solution 1.

1) Note that  $\int_0^a t(x) \exp \frac{-2i\pi x}{a} dx = \int_{x_0}^{x_0+a} \dots$  for any  $x_0$ .

$$\eta_p = \left| \frac{1}{a} \int_{-\frac{fa}{2}}^{\frac{fa}{2}} e^{-2i\pi \frac{px}{a}} dx \right|^2 = \left( \frac{\sin \pi p f}{\pi p} \right)^2$$

$$\eta_{p \text{ max}} = \frac{1}{\pi^2 p^2} \quad \text{Max of all } \eta_p: p = \pm 1, f = \frac{1}{2}, \eta_{\pm 1} = \frac{1}{\pi^2} \approx 10\%$$

2)  $\eta_0 = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$ ,  $\eta_{\pm 1} = \left(\frac{1}{4}\right)^2 = \frac{1}{16} = 6.25\%$ , all other  $\eta_p = 0$ .

3)  $t$  can be expressed as

$$t(x) = 1 + 2 t_{\text{Rouche}}(x) (e^{i\varphi_0} - 1)$$

$$\Rightarrow \eta_1 = \frac{4}{\pi^2} \overset{\text{see question 1}}{|e^{i\varphi_0} - 1|^2} \sin^2 \pi f$$

The maximum,  $\frac{4}{\pi^2} \approx 40.5\%$ , is obtained for  $f = 0.5$ ,  $\varphi_0 = \pi \text{ mod } 2\pi$

4)  $\eta_p = J_p^2(\varphi_0)$ . The max of  $J_1^2$  is  $\approx (0.58)^2 = 34\%$ , for  $\varphi_0 = 1$ .

5)  $t$  even  $\Rightarrow t_p = t_{-p} \Rightarrow \eta_1 = \eta_{-1} \Rightarrow \eta_1 \leq 50\%$

$t$  real  $\Rightarrow t_p = t_{-p}^* \Rightarrow \eta_1 = \eta_{-1} \Rightarrow \eta_1 \leq 50\%$

(in fact, the gratings of questions 1 and 3 can be shown to reach

the max. efficiency  $\eta_1$  of all pure amplitude and real gratings, resp.)

$$6) \eta_1 = \left| \frac{1}{a} \int_0^a t(x) \exp \frac{-2i\pi x}{a} dx \right|^2 \leq \left( \frac{1}{a} \int_0^a |t(x) \exp \frac{-2i\pi x}{a}| dx \right)^2$$

$$\eta_1 \leq \left( \frac{1}{a} \int_0^a |t(x)| dx \right)^2 \leq \left( \frac{1}{a} \int_0^a 1 dx \right)^2$$

For the two equalities to hold:  $|t(x)| = 1$  for all  $x$  and

$t(x) = e^{i\varphi(x)} = e^{2i\pi \frac{x}{a}}$ . This is the "blazed" or "échellette"

grating.

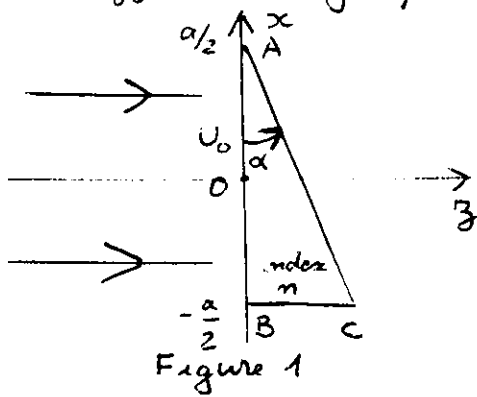
1) Diffraction by a prism:

Figure 1

Consider the prism of figure 1. It is a small angle prism, anti-reflection coated on both sides. It is illuminated under normal incidence by a plane wave with wavelength  $\lambda_0$  and amplitude on the entrance face AB equal to  $U_0$ . Which is the amplitude

on the exit face AC, assuming geometrical propagation? Using the plane wave (= Fourier spectrum) expansion of the latter, find the far field (= Fraunhofer) diffraction pattern. Numerically, try  $\lambda_0 = 0.5 \mu\text{m}$ ,  $n = 1.5$ ,  $a = 1 \text{cm}$ ;  $10 \mu\text{m}$ .

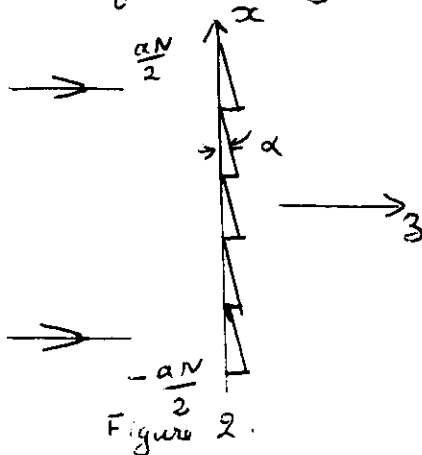
2) Diffraction by a blazed grating consisting of  $N$  micropisms:

Figure 2.

This grating of pitch  $a$  consists of  $N$  micropisms similar to the one of figure 1. Calculate  $\eta_p$ . What happens if  $\alpha$  and  $a$  are such that the phase jump at every discontinuity is exactly  $2\pi$ ? (condition A)

3) Grating dispersion and diffraction efficiency:

In this question, glass dispersion is assumed negligible. Calculate the diffraction efficiency at wavelength  $\lambda$ , given that condition A holds at  $\lambda_0$ . Numerically:  $\lambda \in [0.4, 0.7 \mu\text{m}]$ ,  $\lambda_0 = 0.5 \mu\text{m}$ , then  $\lambda_0 = 20 \mu\text{m}$ . Where does light go?



4) Grating dispersion and glass dispersion.

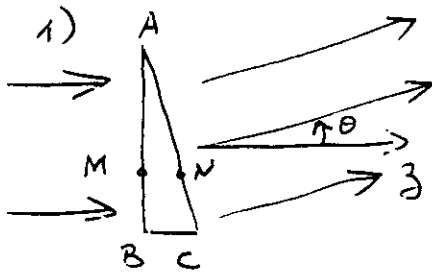
Glass dispersion is now taken into account with the following first order expansion:  $n(\lambda) = n_0 - n_1(\lambda - \lambda_0)$

What changes in question 3? Can the two dispersions cancel out? Numerically:  $n = 1.52$  at  $\lambda = 0.4 \mu\text{m}$  and  $n = 1.50$  at  $\lambda = 0.7 \mu\text{m}$ .

5) Dispersion compensation.

Combining one prism and one grating, how can the two dispersions cancel out?

## Solution 2



$$MN = \left(\frac{a}{2} - x\right) \tan \alpha \approx \left(\frac{a}{2} - x\right) \alpha$$

Incoming wave at N

$$U(N) = U\left(x, \left(\frac{a}{2} - x\right) \alpha\right) = U_0 e^{2i\pi \left(\frac{a}{2} - x\right) \frac{\alpha m}{\lambda_0}}$$

Equation of a plane wave in air, tilted by angle  $\theta$  (see figure):

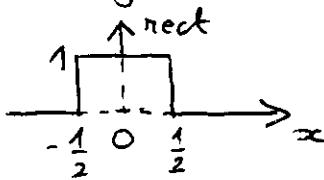
$$U_0' e^{ik_0 z + ik_0 \theta x} \quad \text{at N} \quad U_0' e^{ik_0 \left(\frac{a}{2} - x\right) \alpha + ik_0 \theta x} \quad (k_0 = \frac{2\pi}{\lambda_0})$$

Identifying the incoming wave at N with the latter, it appears that light is refracted according to:

$$-k_0 \alpha x + ik_0 \theta x = k_0 \alpha m x$$

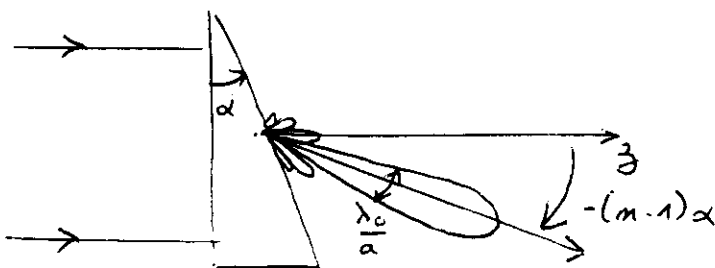
$\Rightarrow \theta = -(m-1)\alpha$ , the usual expression of deviation by a small angle prism.  $U_0' = U_0 e^{i k_0 \alpha x \frac{(m-1)}{2}}$

The complex amplitude of the exiting wave can be rewritten using the Fourier transform of the rect (slit) function.

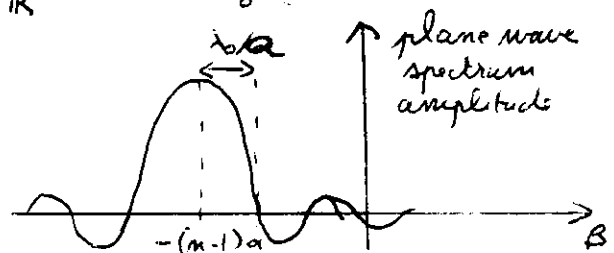


$$U_0' e^{ik_0 z + ik_0 \theta x} \text{rect} \frac{x}{a} = U_0' e^{ik_0 z} \int_{-\infty}^{\infty} a \text{sinc} \mu a \exp i(k_0 \theta + 2\pi \mu) x \, d\mu$$

$$\text{(Let } \beta = \lambda_0 \mu + \theta) \cdot U_0' \frac{a}{\lambda_0} e^{ik_0 z} \int_{-\infty}^{\infty} \text{sinc} \frac{(\beta - \theta)a}{\lambda_0} e^{i k_0 \beta x} \, d\beta$$



polar diffraction diagram



Angular width  $\lambda_0/a$

$$\lambda_0 = 0,5 \mu\text{m}, \alpha = 1 \text{cm} : 5 \cdot 10^{-5} \text{rad}$$

$$10 \mu\text{m} : 5 \cdot 10^{-2} \text{rad} \approx 2^\circ$$

2) Order  $p$  is located at:  $\sin i_p - \sin i_0 = \frac{p\lambda_0}{a} \Rightarrow i_p \approx \frac{p\lambda_0}{a}$

The phase jump at the discontinuities is  $2\pi \frac{a}{\lambda} (m-1)$

Condition A:  $a(m-1)\lambda = \lambda_0$ . In that case,  $\lambda_0$

$i_{-1} = 0$ : order -1 of the grating corresponds to the prism refraction direction.

Diffraction efficiency (refer to exercise 1):

$$t(x) = e^{-2i\pi(m-1)(x-x_0)/\lambda} \quad x_0 - a \leq x < x_0,$$

where  $x_0$  is the apex of one of the microprisms.

$$\eta_p = \left| \frac{1}{a} \int_{x_0-a}^{x_0} t(x) e^{-2i\pi \frac{px}{a}} dx \right|^2.$$

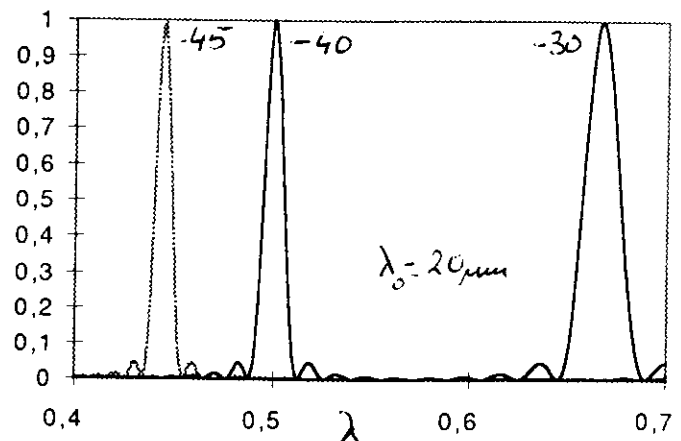
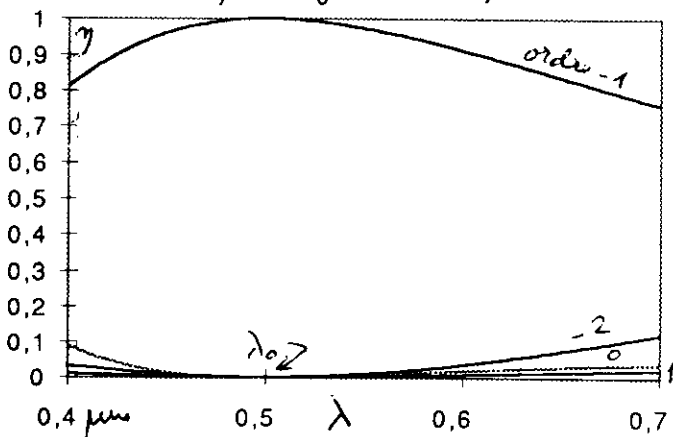
For  $\lambda = \lambda_0$ , condition A yields  $\eta_{-1} = 100\%$  (blazed grating).

3) For an arbitrary  $\lambda$ ,  $\eta_p = \text{sinc}^2 \left( p + \frac{\lambda_0}{\lambda} \right)$

↓  
this is in fact another effect of the sine of question 1.

The curves show the blaze over the visible for  $\lambda_0 = 0.5 \mu\text{m}$  and for  $\lambda_0 = 20 \mu\text{m}$ . In the latter case, it appears that the grating, which by definition is blazed at  $\lambda_0 = 20 \mu\text{m}$  in order -1, is also blazed in order -40 at  $0.5 \mu\text{m}$ , -45 at  $0.4 \mu\text{m}$ , etc. At an arbitrary wavelength  $\lambda$  in the visible, it is not blazed, but it diffracts only a few (non negligible) orders around order  $-\frac{\lambda_0}{\lambda}$ .

Order  $-\frac{\lambda_0}{\lambda}$  is diffracted in direction  $i_p = +\frac{p\lambda}{a} = -\frac{\lambda_0}{a}$ , which according to condition A is just the direction given by geometrical optics for the prism!

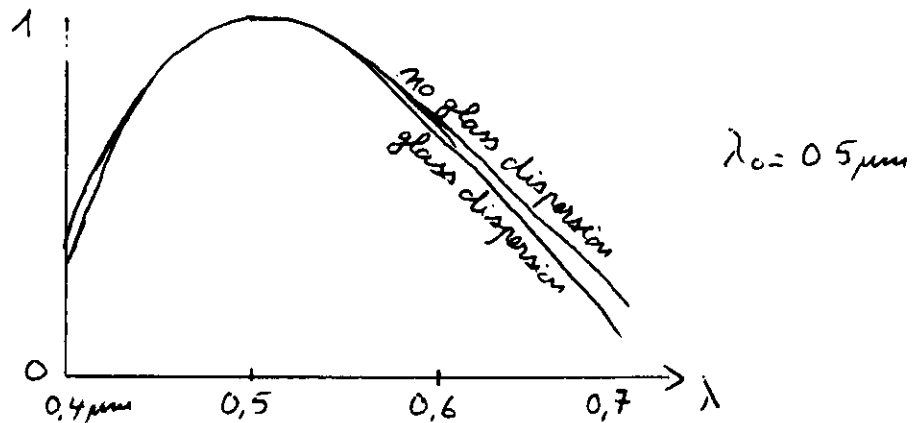


4) With glass dispersion, condition A becomes  $\alpha(m_0 - 1)a = \lambda_0$ ,

while  $t(x) = e^{-2i\pi(m(\lambda) - 1)\alpha(x - x_0)/\lambda}$

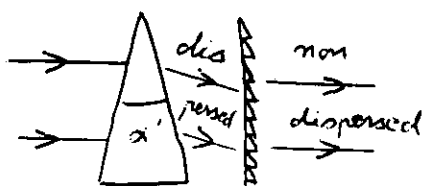
$$\Rightarrow \eta_p = \text{sinc}^2\left(p + \frac{m(\lambda) - 1}{m_0 - 1} \frac{\lambda_0}{\lambda}\right)$$

The grating equation  $i_p = p \frac{\lambda}{a}$  is unchanged: no dispersion occurs. The diffraction efficiency curves are just slightly distorted.



5) The grating sends wavelength  $\lambda$  in direction  $i_{-1}(\lambda) = -\frac{\lambda}{a}$ .

Putting a prism in front of the grating, the incident light is first dispersed by the prism in direction  $-(m(\lambda) - 1)\alpha'$ , where  $\alpha'$  is the prism angle. The grating equation becomes



$$i_{-1}(\lambda) + (m(\lambda) - 1)\alpha' = -\frac{\lambda}{a}$$

With the first order expansion of  $m$

$$i_{-1}(\lambda) = -(m_0 - 1 - m_1(\lambda - \lambda_0))\alpha' - \frac{\lambda \alpha (m_0 - 1)}{\lambda_0}$$

and dispersion cancels out if  $\alpha' m_1 = \alpha(m_0 - 1)$

With the numerical example given,  $\frac{\alpha'}{\alpha} \approx 15 \cdot \frac{\lambda_0}{\lambda_0}$

## Exercise 3. Diffraction gratings fabricated by resist masking (paraxial case).

A grating consists of rectangular steps etched in an antireflection coated plate of glass.

### 1) One groove per period:

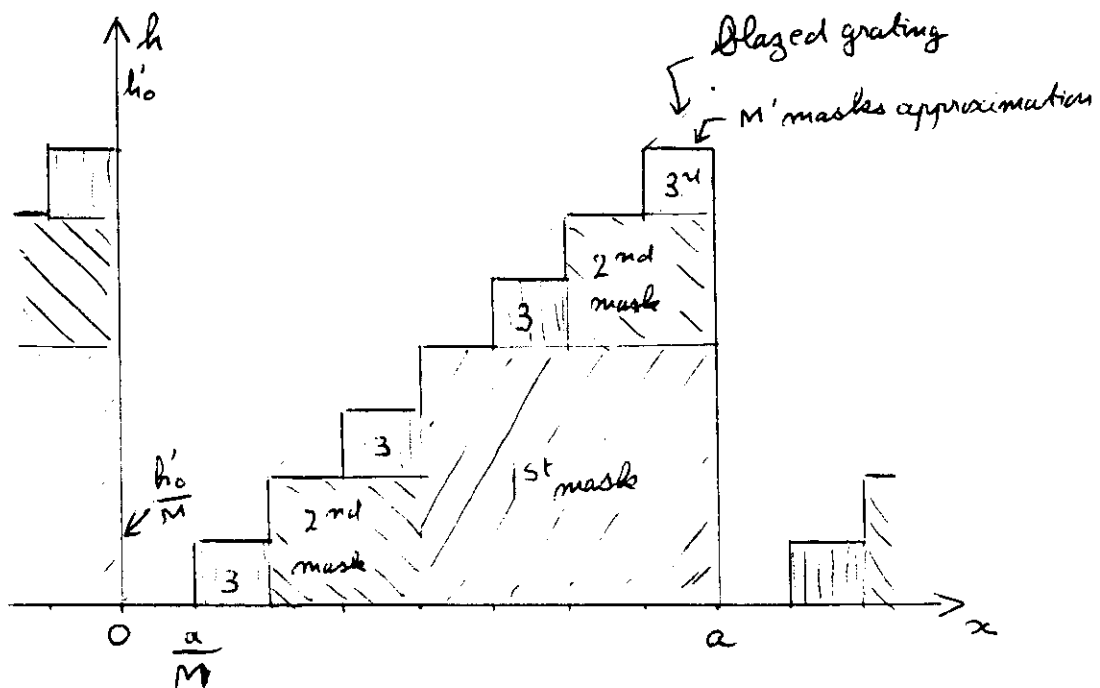
here, the grating is opaque, except that one groove of depth  $h$  is etched in the glass plate (index  $n$ ). The groove extends from abscissa  $x_1$  to abscissa  $x_2$  (modulo  $a$ ). Calculate the Fourier series expansion of transmittance  $t(x)$ .

### 2) M grooves per period:

the grating now consists of  $M$  adjacent grooves. Groove  $j$  ( $j = 0$  through  $M-1$ ) is located between  $x_j$  and  $x_{j+1}$  modulo  $a$  ( $x_0 = 0, x_M = a$ ) and the glass is etched over an height  $h_j$ . Calculate the diffraction efficiencies  $\eta_p$ .

### 3) $M'$ mask approximation of an échellette blazed grating:

Consider the special case of question 2 that best approximates a blazed grating (see figure). Calculate  $\eta_p$ . How large can  $\eta_p$  be at design wavelength if the grating has been fabricated in  $M' = \log_2 M$  masking steps. In the figure,  $M' = 3, M = 8$ .



Solution

$$1) t(x) = \text{rect} \frac{x-x'}{\delta x} e^{-2i\pi \frac{(m-1)h}{\lambda}} * \underset{\substack{\uparrow \\ \text{Dirac comb}}}{\int_a^a} dx$$

$$\text{with } x' = \frac{x_1+x_2}{2}, \delta x = x_2-x_1$$

$$t(x) = \sum_{p \in \mathbb{Z}} t_p e^{+2i\pi \frac{px}{a}} \text{ with } t_p = \frac{1}{a} \int_a^a t(x) e^{-2i\pi \frac{px}{a}} dx$$

$$t_p = \frac{\delta x}{a} \text{sinc} \frac{p\delta x}{a} e^{-2i\pi \left( \frac{x'}{a} p + \frac{(m-1)h}{\lambda} \right)}$$

$$2) t_p = \sum_{j=0}^{M-1} \frac{x_{j+1}-x_j}{a} \text{sinc} p \frac{(x_{j+1}-x_j)}{a} e^{-2i\pi \left( \frac{(x_j+x_{j+1})p}{2a} + \frac{(m-1)h_j}{\lambda} \right)}$$

$$\eta_p = |t_p|^2$$

3) Blaze condition:  $(m-1)h'_0 = \lambda_0$ 

$$x_j = \frac{ja}{M}, h_j = \frac{j h'_0}{M} \Rightarrow t_p \text{ simplifies to}$$

$$\frac{1}{M} \text{sinc} \frac{p}{M} \underset{\substack{\downarrow \\ \text{Some} \\ \text{constant}}}{e^{j\phi_0}} \frac{1 - e^{-2i\pi \left( \frac{p + \frac{\lambda_0}{\lambda} \right) M}}}{1 - e^{-2i\pi \frac{p + \frac{\lambda_0}{\lambda}}{M}}}$$

$$\eta_p = \frac{1}{M^2} \text{sinc}^2 \frac{p}{M} \frac{\sin^2 \pi \left( p + \frac{\lambda_0}{\lambda} \right)}{\sin^2 \pi \frac{p + \frac{\lambda_0}{\lambda}}{M}}$$

For  $p + \frac{\lambda_0}{\lambda} = 0$ ,  $\eta_p = \text{sinc}^2 \frac{p}{M}$  (instead of 100% for a normal blazed gr.)

For a blaze in order -1:

$M=1$	$M=2$	$\eta_{-1}$	40.5%
2	4		81.1
3	8		95.0
4	16		98.7

Mask alignment is a real problem. But the masking method applies to all kinds of diffractive shapes, not just diffraction grating with straight lines and the concept of diffraction efficiency can be generalized (see course).

I Thin lenses:

1.1) Spherical wave:



A spherical wave  $\frac{U_0}{\overline{MC}} e^{-i k \overline{MC}}$  converges at C. It is limited

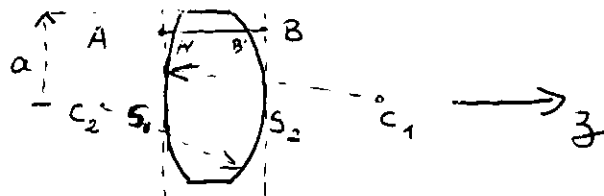
by a circular stop of center O, of axis z, of radius a. Under the assumptions  $a \ll z$  and  $a^4 \ll \lambda z^3$  (the latter is the "Fresnel approximation"), give a simpler expression of the wave complex amplitude.

1.2) Conjugation by thin lenses

A thin lens of index n consists of 2 spherical surfaces of radii

$$R_1 = \overline{C_1 S_1}, R_2 = \overline{C_2 S_2}$$

$$\overline{S_1 S_2} = e$$



a) Assuming a small angle between rays and the z axis, and assuming a Fresnel approximation for the lens face curvature and path delay, the lens pupil radius a, express the propagation from A to B; deduce the amplitude transmittance of the lens.

b) The lens is illuminated by a spherical wave with center P(0,0,p). Express the amplitude at the exit plane and prove the standard lens conjugation formula.

## II Achromatic refractive doublets.

### 2.1) Chromatism of a thin lens.

The lens of q. 1.2 has a nominal image focal length  $f'_d = 50\text{cm}$  at design wavelength  $\lambda_d = 587.6\text{nm}$ . Using the "Abbe number"  $\nu = \frac{n_d - 1}{n_F - n_C}$  ( $\lambda_F = 486.1\text{nm}$ ,  $\lambda_C = 656.3\text{nm}$  and  $\lambda_d$

are standard spectral lines spanning the visible region), express the distance  $\overline{F'F'}$  between the and image foci. For the most common optical glass BK7,

$$n_F = 1.522, n_d = 1.516, n_C = 1.514$$

### 2.2) The basic achromat formula:

Two thin lenses are placed in close contact. The nominal focal length of the doublet is  $f'_d$  at  $\lambda_d$ . One lens is made of BK7, the other of F1 ( $n'_F = 1.632$ ,  $n'_d = 1.620$ ,  $n'_C = 1.615$ ). Calculate the two nominal focal lengths  $f'_{d1}$ ,  $f'_{d2}$  of the two components in such a way that the global focal lengths at  $\lambda_F$  and  $\lambda_C$  coincide (the chromatism has been "folded").

## III Achromatic hybrid (refractive + diffractive) doublet.

### 3.1) Diffractive lens

A glass surface is etched by an amount  $h(x,y)$  in such a way that

- the phase advance  $\varphi(x,y) = -2\pi \frac{(n_d - 1) h(x,y)}{\lambda_d}$  at  $\lambda_d$

never exceeds  $2\pi$

-  $e^{i\varphi(x,y)}$  is equal to the transmittance of a thin lens of

focal distance  $f'_d$ .

Express  $h(x,y)$ .



### 3.2) Chromatism of the diffractive lens

Expressing the transmittance of the diffractive lens as a Fourier series expansion of variable  $r^2 = x^2 + y^2$ , derive the effective Abbe number of its -1 order.

### 3.3) Achromat:

same question as 2.2 with a BK7 + diffractive doublet.

### 3.4) Efficiency:

plot the diffraction efficiency  $\eta_{-1}(\lambda)$  of the diffractive lens for  $\lambda \in [\lambda_F, \lambda_c]$ .

Conclude about the pros and cons of diffractive doublets.

Solution

$$1.1) \left| \frac{U_0}{\bar{m}c} \right| \text{ varies from } \frac{|U_0|}{\sqrt{z^2 + a^2}} \text{ to } \frac{|U_0|}{|z|} \text{ Because } a \ll z,$$

it is essentially constant.

$$-i k \bar{m}c = -i \frac{2\pi}{\lambda} \text{sign}(z) \sqrt{z^2 + r^2} \text{ with } M \begin{vmatrix} x \\ y \\ c \end{vmatrix}, r^2 = x^2 + y^2.$$

$$= -i \frac{2\pi}{\lambda} \left( z + \frac{r^2}{2z} + \frac{r^4}{8z^3} + \dots \right)$$

negligible from Fresnel's approx.

$$\Rightarrow \text{complex amplitude } U_1 e^{-i\pi r^2 / \lambda z}$$

(with  $U_1 = \frac{U_0}{z} e^{-i2\pi z / \lambda}$ ).

1.2) a) Path delay  $(AB) = \overline{AA'} + m \overline{A'B'} + \overline{B'B}$  (terms in  $\cos\theta$ , where  $\theta$  is the angle between the ray and axis  $z$ , are taken as  $\cos\theta$ )

With the same approximation,  $\overline{AA'} = \overline{C_1 S_1} - \overline{C_1 A}$ ; expanding  $\overline{C_1 A}$  as above and neglecting terms in  $r^4$  and higher,

$$\overline{AA'} = -\frac{r^2}{2R_1}, \quad \overline{B'B} = \frac{r^2}{2R_2}, \quad \overline{A'B'} = e + \frac{r^2}{2} \left( \frac{1}{R_1} - \frac{1}{R_2} \right)$$

$$\Rightarrow (AB) = m e + (m-1) \frac{r^2}{2} \left( \frac{1}{R_1} - \frac{1}{R_2} \right)$$

$$t(x, y) = e^{2i\pi \frac{(AB)}{\lambda}}$$

b) Incoming wave:  $U_1 e^{-i\pi r^2 / \lambda p}$

$$\text{Outcoming wave: } U_1 e^{2i\pi \frac{me}{\lambda}} e^{-i\pi \frac{r^2}{\lambda} \left( \frac{1}{p} + (m-1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right) \right)}$$

$$= U_1' e^{-i\pi \frac{r^2}{\lambda p'}}$$

$$\Rightarrow \frac{1}{p} - \frac{1}{p'} = -\frac{1}{f'} \text{ with } \frac{1}{f'} = (m-1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right).$$

## 2.1) Chromatism:

$$f'(m-1) = \text{cst} \Rightarrow$$

$$F'_F F'_C = f'_C - f'_F = f'_d (m_d - 1) \left( \frac{1}{m_c - 1} - \frac{1}{m_F - 1} \right)$$

$$= f'_d \frac{(m_d - 1)(m_F - m_c)}{(m_c - 1)(m_F - 1)} \approx \frac{f'_d}{v} = \frac{500}{\frac{0.516}{0.008}} = 8 \text{ mm}$$

Such chromatism is by far the worst defect for such a lens, unless the aperture is very large.

## 2.2) Basic achromat

linear system in  $\frac{1}{f}$  variables

$$\begin{cases} \frac{1}{f'_{d1}} + \frac{1}{f'_{d2}} = \frac{1}{f'_d} & \text{index 1: BK7; index 2: F2} \\ \frac{1}{f'_{c1}} + \frac{1}{f'_{c2}} = \frac{1}{f'_{F1}} + \frac{1}{f'_{F2}} \end{cases}$$

substitute  $f'_{c1} = f'_{d1} \frac{(m_d - 1)}{(m_c - 1)}$  etc:

linear system in  $\frac{1}{f'_{d1}}, \frac{1}{f'_{d2}}$

$$\begin{cases} \frac{1}{f'_{d1}} + \frac{1}{f'_{d2}} = \frac{1}{f'_d} \\ \frac{m_c - m_F}{f'_{d1}(m_d - 1)} + \frac{m'_c - m'_F}{f'_{d2}(m'_d - 1)} = 0 \end{cases}$$

$$\Rightarrow \begin{aligned} f'_{d1} &= \frac{v - v'}{v'} f'_d = 217 \text{ mm} \\ f'_{d2} &= \frac{v' - v}{v'} f'_d = -384 \text{ mm} \end{aligned}$$

The lens in the more dispersive glass must always be negative (if desired result is positive, and conversely). The other lens must therefore be more convergent (resp divergent) than the desired result. This is a challenge for the geometrical aberrations.

NB. it is easy to calculate a residual, or secondary, chromatism

$$f'_d - f'_c = f'_d - f'_F$$

### III Hybrid doublet

#### 3.1) Diffractive lens

$$e^{i\varphi(x,y)} = e^{-i\pi \frac{r^2}{\lambda d}}$$

Let  $\mathcal{F}$  be the "fractional part" function  $\mathcal{F}(u) = u - \text{integer part}$

$$\varphi = -\mathcal{F}\left(\frac{r^2}{2\lambda d f'_d}\right)$$

$$\Rightarrow h(x,y) = \frac{\lambda d}{2\pi(m_d - 1)} \mathcal{F}\left(\frac{r^2}{2\lambda d f'_d}\right)$$

#### 3.2) Chromaticism:

at wavelength  $\lambda$ ,  $t(x,y) = e^{-2i\pi \frac{(m-1)d}{\lambda} \mathcal{F}\left(\frac{r^2}{2\lambda d f'_d}\right)}$

$$= \exp\left[-i \frac{(m-1)\lambda d}{(m_d - 1)\lambda} \mathcal{F}\left(\frac{r^2}{2\lambda d f'_d}\right)\right]$$

This is a periodic function in  $r^2$  of period  $2\lambda d f'_d \Rightarrow$

$$t(x,y) = \sum_{p \in \mathbb{Z}} t_p e^{i 2\pi p \frac{r^2}{2\lambda d f'_d}}$$

in order  $-1$ , this is equivalent to a lens  $e^{-i\pi \frac{r^2}{\lambda f'}}$   
of focal length  $f' = \frac{\lambda d}{\lambda} f'_d$

$$\Rightarrow \text{effective Abbe number } \nu = \frac{m_d - 1}{m_F - m_C} = \frac{\frac{1}{f'_d}}{\frac{1}{f'_F} - \frac{1}{f'_C}} = \frac{1}{\frac{\lambda_F}{\lambda_d} - \frac{\lambda_C}{\lambda_d}}$$

↑  
these are fictitious indices

A diffractive lens has the same chromatic (dispersive) properties as a glass of refractive index  $\frac{\lambda}{\lambda_d} + 1$ .

$$\text{Effective } \nu = \frac{1}{\frac{486.1}{587.6} - \frac{656.3}{587.6}} = -3.45 \lambda_d \quad \nu \text{ is very small and negative}$$

#### 3.3) Doublet: only numerical values change:

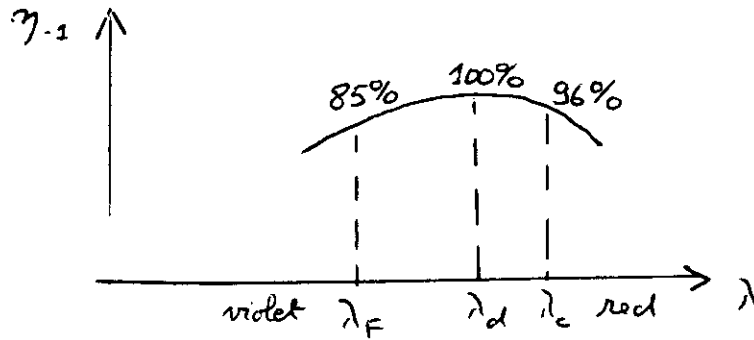
$$\text{BK7: } f'_{d1} = 527 \text{ mm, dif. } f'_{d2} = 9850 \text{ mm}$$

$t_{-1}$  coefficient in Fourier series

$$t_{-1} = \frac{1}{2\lambda_d \theta_d} \int_0^{2\lambda_d \theta_d} e^{-i \frac{(m-1)\lambda_d}{(m_d-1)\lambda} \frac{\pi^2}{2\lambda_d \theta_d} + 2i\pi \frac{\pi^2}{2\lambda_d \theta_d} d(\pi^2)} d(\pi^2)$$

$$= -e^{i\pi \left( \frac{m-1}{m_d-1} \frac{\lambda_d}{\lambda} - 1 \right)} \text{sinc} \left( \frac{m-1}{m_d-1} \frac{\lambda_d}{\lambda} - 1 \right)$$

$$\eta_{-1} = \text{sinc}^2 \left( \frac{m-1}{m_d-1} \frac{\lambda_d}{\lambda} - 1 \right)$$



### 3.4) Discussion:

the system is significantly lighter and suffers less geometrical aberrations because the main (BK7) component has lower convergence. However, there is a non negligible loss of light into the undesired, defocused orders if the spectral range of interest is wide, and it has not been easy until very recently to fabricate diffractive structures.

A zone plate is a diffractive component whose amplitude transmittance is a periodic function of variable  $r^2 = x^2 + y^2$ .

In some spectral ranges where few materials with a refractive index different from unity are available, binary zone plates<sup>(1)</sup> are used for lenses.

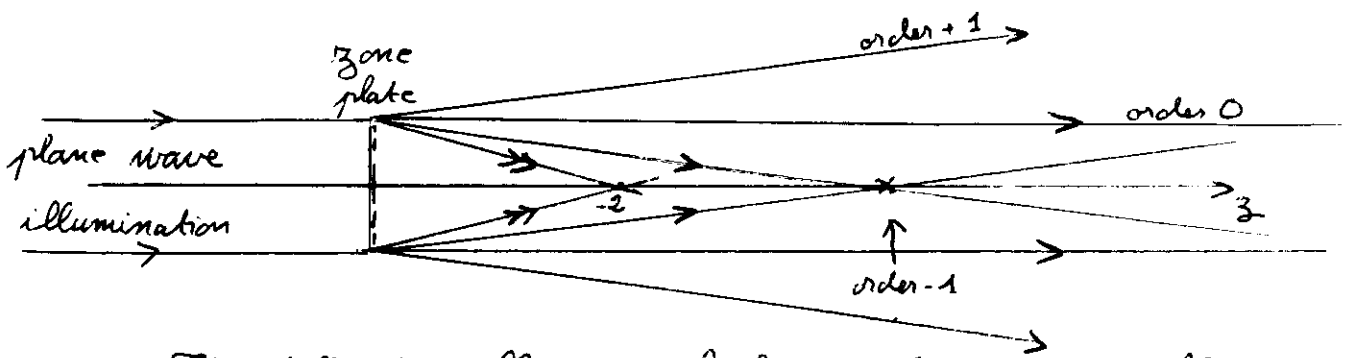
- 1) Show why a zone plate is indeed in some sense equivalent to a lens. Define the concepts of zone plate orders and their diffraction efficiencies.
- 2) Calculate the relation between the width of the outer zone of a zone plate and that of its Airy pattern formed at its  $-1$  order focus when it is illuminated by a plane wave.
- 3) Application:  $\lambda = 3 \text{ nm}$  (soft X-ray range), outer zone radius  $a = 45 \mu\text{m}$ , outer zone width  $50 \text{ nm}$ . Calculate its focal length. Why are those X-ray lenses mostly used with central obstruction? Calculate impulse response for a central obstruction of  $0.4 a$ .
- 4) Is such a lens stigmatic? To answer, calculate how well the Fresnel approximation is satisfied.

(1) They are often called "Fresnel zone plates" because they use the concept of Fresnel zones (Fresnel, 1816, in his historic paper on diffraction) and because they are reminiscent of the (non diffractive!) concept of Fresnel lenses used in lighthouses, overhead projectors, etc. Historically, they were invented as a curiosity by Soret in 1875.

1)  $t$  periodic in  $x^2$  (period  $\cdot r_0^2$ )  $\Rightarrow$

$$t(x, y) = \sum_{p \in \mathbb{Z}} t_p e^{2i\pi p \frac{x^2}{r_0^2}}$$

Each "order"  $p$  is a lens with focal length  $f'_p = -\frac{r_0^2}{2p\lambda}$  and diffraction efficiency  $\eta_p = |t_p|^2$ .  
Compared to a grating, one problem is that the orders strongly overlap. Example:



The diffraction efficiency calculations of exercise 1 all exactly apply to zone plates.

2) Zone plate radius:  $a$ . Airy pattern width (between center and first dark ring):  $1.22 \frac{\lambda f'_{-1}}{2a} = 1.22 \frac{r_0^2}{4a}$

Last zone width:  $a^2 = N r_0^2$ ,  $N$  is the total number of zones (may not be an integer).  $(a - \delta a)^2 = (N-1) r_0^2$ ,  $\delta a$  is the last zone width  $2a \delta a \approx r_0^2 \Rightarrow$

Airy width =  $0.6 \delta a$ .

3)  $f'_{-1} = \frac{r_0^2}{2\lambda} = \frac{2a \delta a}{2\lambda} = \frac{45 \times 0.05}{0.003} = 750 \mu\text{m}$ .

For a very small field of view, central obstruction avoids order overlap. This is satisfactory in a scanning microscope.

We consider the following situation:

- above  $z = h$ ,  $n = n_0$

- below  $z = 0$ ,  $n = n_1$

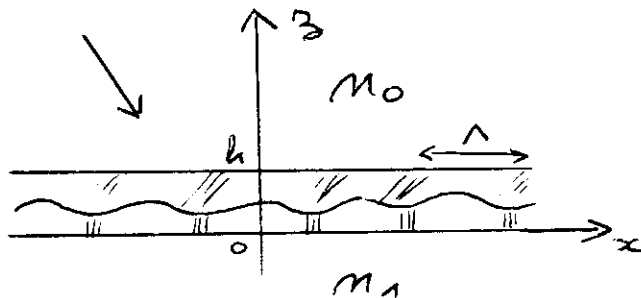
- $0 \leq z \leq h$ ,  $n$  is a periodic function of  $x$ :

$$n(x + \Lambda) = n(x)$$

- only one light source, located at infinity above the grating in such a direction that the  $k_x$  component of the incoming wave is  $2\pi u_0$ ,  $u_0$  real

- two dimensional problem,  $k_y = 0$ , with monochromatic illumination

- TE polarisation, i.e.  $\vec{E} \parallel y$



1) Show that Maxwell's equations reduce to the wave equation

2) Expand  $E$  for  $z > h$  in plane waves using a

1D Fourier transform of  $E(x, z)$  in  $x$ , resulting in

$$z > h \Rightarrow E(x, z) = \int [E_+(u) e^{i\chi_0(u)(z-h)} + E_-(u) e^{-i\chi_0(u)(z-h)}] e^{2i\pi x u} du$$

3) Assume  $E(x, h)$  and  $\frac{\partial E}{\partial z}(x, h)$  is known and solve in  $E_-(u)$

and  $E_+(u)$

4) Use the same principle for  $z < 0$  and solve in  $E'_-(u)$  and

$$E'_+(u) \text{ from } E(x, 0) \text{ and } \frac{\partial E}{\partial z}(x, 0)$$



5) From the assumption about the light source,

$$E_-(u) = A \delta_{u_0}, \quad E_+(u) = 0$$

$\downarrow$   
Constant

Show that the wave equation and the expressions for

$E_-$  and  $E_+$  comply with the property:

if  $E(x, z)$  is a solution, then so is

$$F(x, z) = E(x + \Lambda, z) e^{-2i\pi u_0 \Lambda}$$

6) From the unicity of the solution, deduce Floquet's theorem:

$$E(x, z) e^{-2i\pi u_0 x} \text{ is periodic in } x \text{ with period } \Lambda$$

(NB in the TM case, the demonstration is basically the same using for the wave equation

$$\frac{\partial}{\partial x} \left( \frac{1}{m^2} \frac{\partial B}{\partial x} \right) + \frac{\partial}{\partial z} \left( \frac{1}{m^2} \frac{\partial B}{\partial z} \right) + k_0^2 B = 0.$$

The general case of oblique incidence is more tricky, but the theorem is also valid).

Solution

1) The Maxwell equations reduce to

$$\vec{\nabla} \wedge \vec{E} = i\omega \vec{B} \quad \vec{\nabla} \cdot \vec{B} = 0$$

$$\vec{\nabla} \cdot (m^2 \vec{E}) = 0 \quad \vec{\nabla} \wedge \vec{B} = -i\omega \mu_0 m^2 \vec{E}$$

$$\text{Let } k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda_0}$$

From the two curl equations,  $\vec{\Delta} \vec{E} + m^2(x, z) k_0^2 \vec{E} = 0$ .

With  $\vec{E} = E \vec{y}$ , this reduces to  $\Delta E + m^2(x, z) k_0^2 E = 0$ .

$\vec{B}$  is then given by  $\vec{\Delta} \wedge \vec{E} / i\omega$ .

The  $\vec{\nabla} \cdot \vec{B}$  and  $\vec{\nabla} \cdot (m^2 \vec{E})$  equations vanish using the assumptions of TE polarisation, no  $y$ -dependence.

$$2) E(x, z) = \int \tilde{E}(u, z) e^{2i\pi ux} du$$

$$\Delta E + k_0^2 m^2 E = \int \left( -4\pi^2 u^2 \tilde{E}(u, z) + \frac{\partial^2 \tilde{E}}{\partial z^2} + m^2 k_0^2 \tilde{E} \right) e^{2i\pi ux} du$$

$$\Rightarrow \frac{\partial^2 \tilde{E}}{\partial z^2} = (4\pi^2 u^2 - m^2 k_0^2) \tilde{E}$$

$$\text{Let } \chi_0(u) = \begin{cases} -i\sqrt{4\pi^2 u^2 - m^2 k_0^2} & \text{if } u > \frac{m_0 k_0}{2\pi} \\ -i\sqrt{m^2 k_0^2 - 4\pi^2 u^2} & \text{otherwise} \end{cases}$$

$$\Rightarrow \tilde{E}(u, z) = E_+(u) e^{i\chi_0(u)z} + E_-(u) e^{-i\chi_0(u)z}$$

which can be rewritten for convenience

$$\tilde{E}(u, z) = E_+(u) e^{i\chi_0(u)(z-h)} + E_-(u) e^{-i\chi_0(u)(z-h)}$$

$\Rightarrow$

$$E(x, z) = \int \left[ E_+(u) e^{i\chi_0(u)(z-h)} + E_-(u) e^{-i\chi_0(u)(z-h)} \right] e^{2i\pi ux} du$$

$$3) \frac{\partial}{\partial z} \bar{E}(x, z) = \int i\chi_0(u) \left( E_+(u) e^{i\chi_0(u)(z-h)} - E_-(u) e^{-i\chi_0(u)(z-h)} \right) e^{2i\pi ux} dx$$

By inverse Fourier transform at  $z=h$ .

$$E_-(u) + E_+(u) = \int \bar{E}(x, h) e^{-2i\pi ux} dx$$

$$-E_-(u) + E_+(u) = \frac{1}{i\chi_0(u)} \int \frac{\partial \bar{E}(x, h)}{\partial z} e^{-2i\pi ux} dx$$

$$\Rightarrow E_-(u) = \frac{1}{2} \int \left( \bar{E}(x, h) - \frac{1}{i\chi_0(u)} \frac{\partial \bar{E}(x, h)}{\partial z} \right) e^{-2i\pi ux} dx$$

$$E_+(u) = \dots + \dots$$

$$4) \text{ Similarly, with } \chi_1(u) = \begin{cases} -i \sqrt{4\pi^2 u^2 - m_1^2 k_0^2} & \text{if } u > \frac{m_1 k_0}{2\pi} \\ -i \sqrt{m_1^2 k_0^2 - 4\pi^2 u^2} & \text{otherwise} \end{cases}$$

$$E'_+(u) = \frac{1}{2} \int \left( E(x, 0) + \frac{1}{i\chi_1(u)} \frac{\partial E(x, 0)}{\partial z} \right) e^{-2i\pi ux} dx$$

$$\bar{E}'_-(u) = \dots - \dots$$

5)  $\Delta F + k_0^2 m^2 F = 0$  is immediate from  $\Delta E + k_0^2 m^2 E = 0$

$$E_-(u) = \frac{1}{2} \int \left( E(x, h) - \frac{1}{i\chi_0(u)} \frac{\partial E(x, h)}{\partial z} \right) e^{-2i\pi ux} dx = A \delta_{u_0}$$

$$\int F(x, h) e^{-2i\pi ux} dx = e^{-2i\pi u_0 h} \int E(x+h, h) e^{-2i\pi ux} dx$$

$$= e^{-2i\pi(u_0-u)h} \int E(x, h) e^{-2i\pi ux} dx$$

$$\text{similarly } \int \frac{\partial F(x, h)}{\partial z} e^{-2i\pi ux} dx = e^{-2i\pi(u_0-u)h} \int \frac{\partial E(x, h)}{\partial z} e^{-2i\pi ux} dx$$

$$\Rightarrow \frac{1}{2} \int \left( F(x, h) - \frac{1}{i\chi_0(u)} \frac{\partial F(x, h)}{\partial z} \right) e^{-2i\pi ux} dx = e^{2i\pi(u-u_0)h} A \delta_{u_0}$$

$$= A \delta_{u_0}$$

56.3  
With the same calculation concerning  $E'_+$ , it follows that  $F(x, z)$  is a solution

$$6) \text{ Therefore } F(x, z) = E(x, z) = E(x + \Lambda, z) e^{-2i\pi u_0 \Lambda}$$

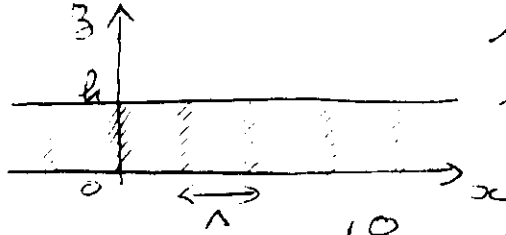
$$\Rightarrow E(x, z) e^{-2i\pi u_0 x} = E(x + \Lambda, z) e^{-2i\pi u_0 (x + \Lambda)}$$

$E(x, z) e^{-2i\pi u_0 x}$  is periodic in  $x$  with period  $\Lambda$ .

The Fourier series expansions that ~~for~~ derive from this theorem are the basis of all rigorous theories of grating diffraction.

# The Kogelnik model of coupled wave in thick transmission holograms

A thick diffraction grating extends between  $z=0$  and  $z=h$ .



The spatially averaged refractive index is  $n = \sqrt{\epsilon}$ . The field at  $z=0$  can be expressed in two ways

$$\vec{E}(x, y, 0) = \begin{pmatrix} 0 \\ E_i e^{i \vec{k}_i \cdot \vec{r}} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ E_i e^{i \vec{k}'_i \cdot \vec{r}} \\ 0 \end{pmatrix}$$

with  $(E_i \text{ a known parameter})$

$$\vec{k}_i = \begin{pmatrix} -\frac{2\pi}{\lambda_0} \sin \theta \\ 0 \\ \frac{2\pi}{\lambda_0} \cos \theta \end{pmatrix} \quad \vec{k}'_i = \begin{pmatrix} -\frac{2\pi m}{\lambda_0} \sin \theta' \\ 0 \\ \frac{2\pi m}{\lambda_0} \cos \theta' \end{pmatrix}$$

- 1) What is the physical interpretation of the compatibility of those two expressions in plane  $z=0$ .

In the following, the dielectric constant of the grating is a non-slanted sinusoid:

$$\epsilon(x, y, z) = \epsilon(x) = \epsilon + \epsilon_1 \cos \frac{2\pi x}{\lambda}$$

Let  $\vec{k}'_i$  designate the vector  $\begin{pmatrix} \frac{2\pi m}{\lambda_0} \sin \theta' \\ 0 \\ \frac{2\pi m}{\lambda_0} \cos \theta' \end{pmatrix}$

- 2) Under which condition does  $\vec{k}'_i$  correspond to the first order diffracted by the grating? Why is this called a (first order) Bragg condition, as opposed to the case of a thin grating?

3) Using the Helmholtz wave equation

$$\Delta E + k^2 \epsilon(x) E(x, z) = 0$$

express the propagation of a  $z$ -varying superposition of the incoming and Bragg-diffracted waves:

$$\vec{E}(\vec{r}, z) = \begin{cases} 0 \\ E_0(z) e^{i\vec{k}_0 \cdot \vec{r}} + E_1(z) e^{i\vec{k}_1 \cdot \vec{r}} \\ 0 \end{cases} \quad 0 < z < h$$

4) Neglecting

- second derivatives of  $E_0, E_1$  and
- non Bragg-matched waves

derive the set of two coupled equations that govern  $E_0(z)$  and  $E_1(z)$

5) Let  $E_+(z) = E_0(z) + E_1(z)$

$$E_-(z) = E_0(z) - E_1(z)$$

Derive the boundary conditions (see question 1) in terms of  $E_+, E_-$ . Solve in  $E_+, E_-$ , and then in  $E_0, E_1$ .

6) A dichromated gelatin hologram has

$$\epsilon = 2.25, \quad \epsilon_1 = 0.1$$

It is reconstructed at wavelength  $\lambda_0 = 0.6 \mu\text{m}$

- Plot the first order diffraction efficiency  $\eta_1$  as a function of  $h$  for  $\Lambda = 1 \mu\text{m}$
- Plot the minimal thickness for which  $\eta_1 = 10\%$  as a function of  $\Lambda$ .

Solution

1) At  $z=0$ ,  $e^{i\vec{k}_0 \cdot \vec{r}} = e^{i\vec{k}'_0 \cdot \vec{r}}$  through Snell's law, that derives from the conservation of the in-plane component of  $\vec{k}$ :  $\sin\theta = m \sin\theta'$

2) Grating orders:  $\vec{k}'_p = \vec{k}'_0 + p\vec{K}$ ,  $\vec{K} = \begin{pmatrix} 2\pi/\Lambda \\ 0 \\ 0 \end{pmatrix}$ . In general,

the relation holds between vectors for thick media and leads to light being diffracted only if  $|\vec{k}'_0| = |\vec{k}'_0 + p\vec{K}|$ , which is the Bragg condition and can be interpreted as a reflection on the equal index planes (in crystallography, reticular planes). For a thin grating, the equal index planes vanish, the Bragg condition disappears and the  $z$  component condition of the above relation vanishes, leaving the usual grating law:  $k'_{px} = k'_{0x} + pK_x$ ,  $\sin\theta'_p - \sin\theta'_0 = p\frac{\lambda}{\Lambda}$ .

The condition here for  $p=1$  is  $2m \sin\theta' = \frac{\lambda}{\Lambda}$ , or  $\sin\theta = m \sin\theta' = \frac{\lambda}{2\Lambda}$ .

3)  $\frac{\partial^2 E}{\partial x^2} = -\left(\frac{2\pi}{\lambda} m \sin\theta'\right)^2 E$

$$\frac{\partial^2 E}{\partial z^2} = -k_3'^2 E + 2ik_3' (E'_0 e^{i\vec{k}'_0 \cdot \vec{r}} + E'_1 e^{i\vec{k}'_1 \cdot \vec{r}} + E''_0 e^{i\vec{k}'_0 \cdot \vec{r}} + E''_1 e^{i\vec{k}'_1 \cdot \vec{r}})$$

Helmholtz wave equation  $\Rightarrow$

$$e^{i\vec{k}'_0 \cdot \vec{r}} \left( 2ik_0 m \cos\theta' E'_0 + \textcircled{E''_0} + k_0^2 \frac{\epsilon_1}{2} E_2 \right) + e^{i\vec{k}'_1 \cdot \vec{r}} \left( 2ik_0 m \cos\theta' E'_1 + \textcircled{E''_1} + k_0^2 \frac{\epsilon_1}{2} E_0 \right) + e^{i\vec{k}'_0 \cdot \vec{r} - i\vec{K} \cdot \vec{r}} k_0^2 \epsilon_1 E_0 + e^{i\vec{k}'_0 \cdot \vec{r} + i\vec{K} \cdot \vec{r}} k_0^2 \epsilon_1 E_1 = 0 \quad \left( k_0 = \frac{2\pi}{\lambda_0} \right)$$

4) In the Kogelnik model, the circled terms above are

neglected  $\Rightarrow$  
$$\begin{cases} 4 i m \cos \theta' E_0' + k_0 \epsilon_1 E_1 = 0 \\ 4 i m \cos \theta' E_1' + k_0 \epsilon_1 E_0 = 0 \end{cases}$$

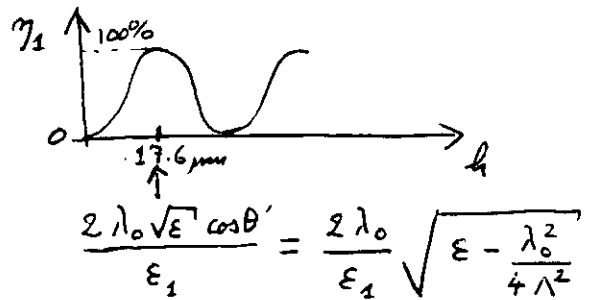
5)  $\beta_3 = 0 \Rightarrow E_0 = E_+$ ,  $E_1 = 0 \Rightarrow E_+ = E_+$ ,  $E_- = E_+$ .

The coupled equations combine into the two eigenmode

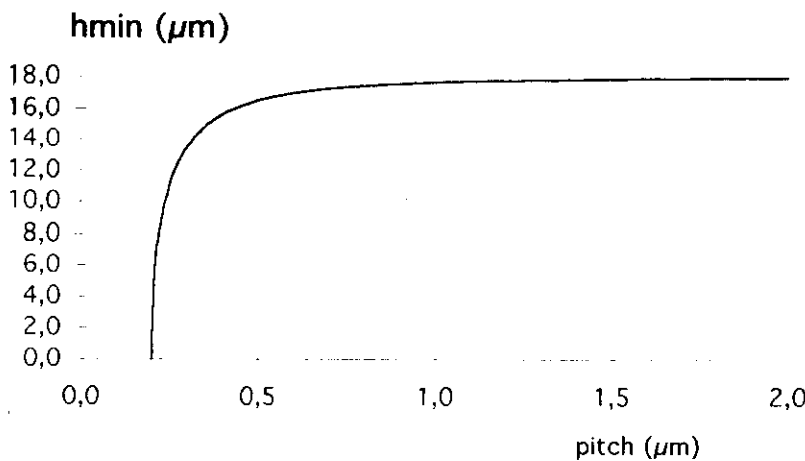
equations. 
$$\begin{cases} E_+ - i A E_- = 0 \\ E_- + i A E_+ = 0 \end{cases}$$
 with  $A = \frac{k_0 \epsilon_1}{4 m \cos \theta'}$

$\Rightarrow \begin{cases} E_+ = E_i e^{-i A \beta_3} \\ E_- = E_i e^{i A \beta_3} \end{cases} \Rightarrow \begin{cases} E_0(\beta_3) = E_i \cos A \beta_3 \\ E_1(\beta_3) = -i E_i \sin A \beta_3 \end{cases}$

6) a)  $\eta_1 = \frac{|E_1|^2}{|E_0|^2} = \sin^2 A h$



b)  $h_{min}(\eta_1 = 100\%) = \frac{2 \lambda_0}{\epsilon_1} \sqrt{\epsilon - \frac{\lambda_0^2}{4 \Lambda^2}}$





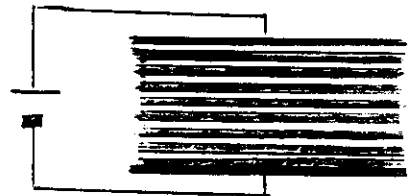
A model of form birefringence due to a grating with a pitch  $\Lambda \ll \lambda$  is given by a capacitor equivalent circuit.

Consider a plane capacitor with surface area  $S$ , thickness  $e$ , and no border effect. From elementary physics courses, its capacity is  $C = \frac{\epsilon_0 \epsilon S}{e}$ . The capacitor, however, has a microscopic

structure. It consists of a stack of homogeneous layers of thickness  $f\Lambda$  with permittivity  $\epsilon$  spaced by a vacuum ( $\epsilon=1$ ) layer of thickness  $(1-f)\Lambda$ .  $f$  is called the "fill factor" (or duty cycle),  $0 < f < 1$ .

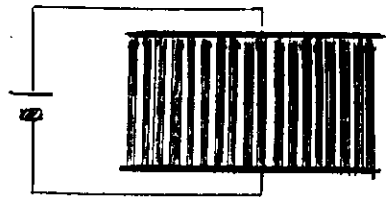
1) The layers run parallel to the electrodes.

$e = m_0 \Lambda$ ,  $m_0$  an integer. Calculate  $\epsilon$ .  
(This is the  $\epsilon_{\perp}$  perpendicular permittivity for  $\vec{E}$  field perpendicular to the layers)



2) The layers are perpendicular to the electrodes.

Calculate  $\epsilon$ , which is defined as the  $\epsilon_{\parallel}$  parallel permittivity.



3) For which  $f$  value is the  $\epsilon$  anisotropy maximal?

↳ For a fill factor  $f \in (0 \rightarrow 1)$ , plot the  $n_o$  and  $n_e$  ordinary and extraordinary indices for a glass ( $n=1.5$ ) and a germanium ( $n=4$ ) grating.

## Solution

The battery is just here to say that there is an  $\vec{E}$  field perpendicular to the "capacitor" planes. The capacitor approximation is valid over distances  $\ll \lambda$  so that no propagation seems to occur.

1) 2  $m_0$  capacitors are placed in series:

$$\frac{1}{C} = \frac{m_0}{C_1} + \frac{m_0}{C_2} \quad \text{with } C_1 \text{ the capacity of an } (\epsilon, f) \text{ capacitor}$$

$$C_2 \text{ " " " } (1, (1-f)) \text{ "}$$

$$\Rightarrow \frac{1}{C} = \frac{m_0 f \Lambda}{\epsilon_0 \epsilon S} + \frac{m_0 (1-f) \Lambda}{\epsilon_0 S} = \frac{e}{\epsilon_0 {}^n \epsilon_{\perp} S} \Rightarrow$$

$${}^n \epsilon_{\perp} = \frac{1}{\frac{f}{\epsilon} + 1-f}$$

2) 2 capacitors are placed in parallel:

$C = C_1 + C_2$ , where  $C_1$  is itself a set of many small capacitors  $(\epsilon, e)$  that together occupy area  $fS$ , and  $C_2 \dots$

$(1, e) \dots \dots \dots (1-f)S$

$$C = \frac{\epsilon_0 {}^n \epsilon_{\parallel} S}{e} = \frac{\epsilon_0 \epsilon f S}{e} + \frac{\epsilon_0 (1-f) S}{e} \Rightarrow$$

$${}^n \epsilon_{\parallel} = \epsilon f + 1-f.$$

$$3) {}^n \epsilon_{\perp} - {}^n \epsilon_{\parallel} = -\frac{(\epsilon-1)^2 f(1-f)}{1+(\epsilon-1)(1-f)}, \text{ the derivative vanishes}$$

$$\text{for } f = \frac{\epsilon \pm \sqrt{\epsilon}}{\epsilon-1}, \text{ only } \frac{\epsilon - \sqrt{\epsilon}}{\epsilon-1} \text{ is physical.}$$

$$f = \frac{\epsilon - \sqrt{\epsilon}}{\epsilon-1} = \frac{n}{n+1} \text{ leads to } {}^n \epsilon_{\perp} - {}^n \epsilon_{\parallel} = -(n-1)^2$$

$$4) m_0 = \sqrt{1-f + \epsilon f} = \sqrt{{}^n \epsilon_{\parallel}}, \quad m_e = \frac{1}{\sqrt{\frac{f}{\epsilon} + 1-f}} = \sqrt{{}^n \epsilon_{\perp}}$$

$$\text{For } f=0.5, m_0 = \sqrt{\frac{n^2+1}{2}}, m_e = n \sqrt{\frac{2}{n^2+1}}, \delta m = m_e - m_0 < 0.$$

$$\text{glass. } \delta m = -0.1; \text{ germanium: } \delta m = -1.54.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
LINCOLN LABORATORY

**BINARY OPTICS TECHNOLOGY: THE THEORY AND DESIGN  
OF MULTI-LEVEL DIFFRACTIVE OPTICAL ELEMENTS**

*G.J. SWANSON*  
*Group 52*

TECHNICAL REPORT 854

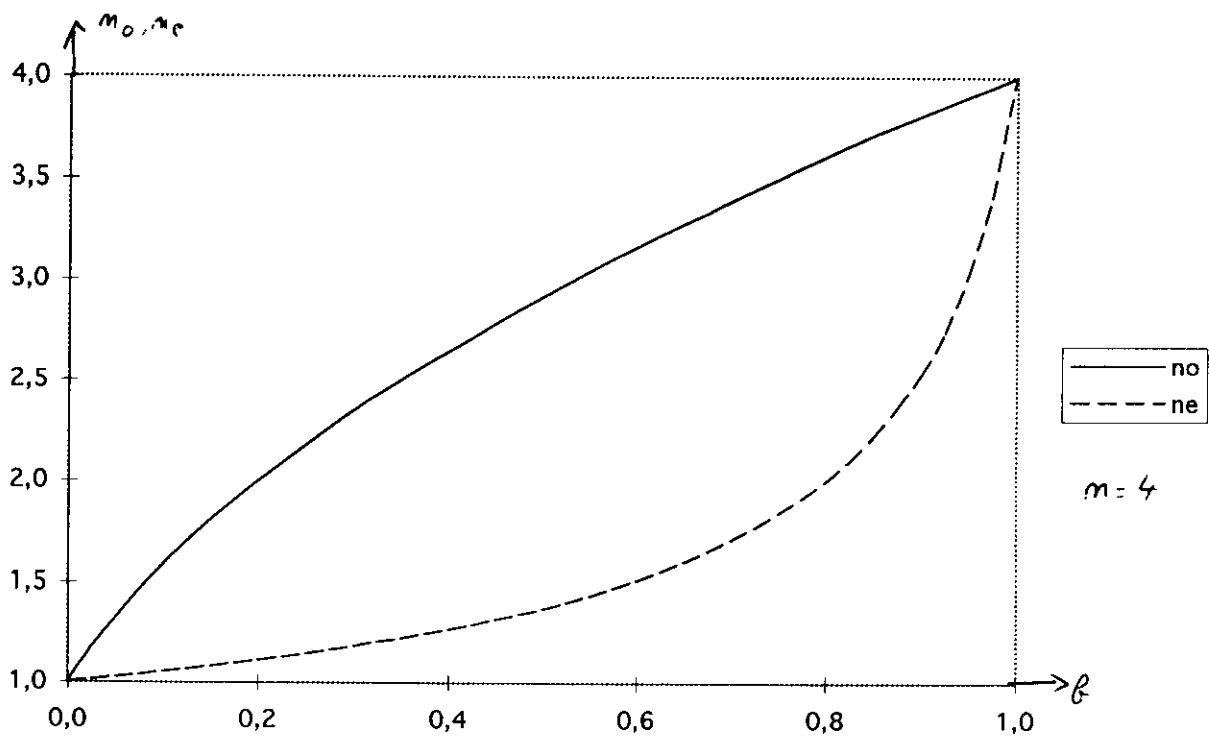
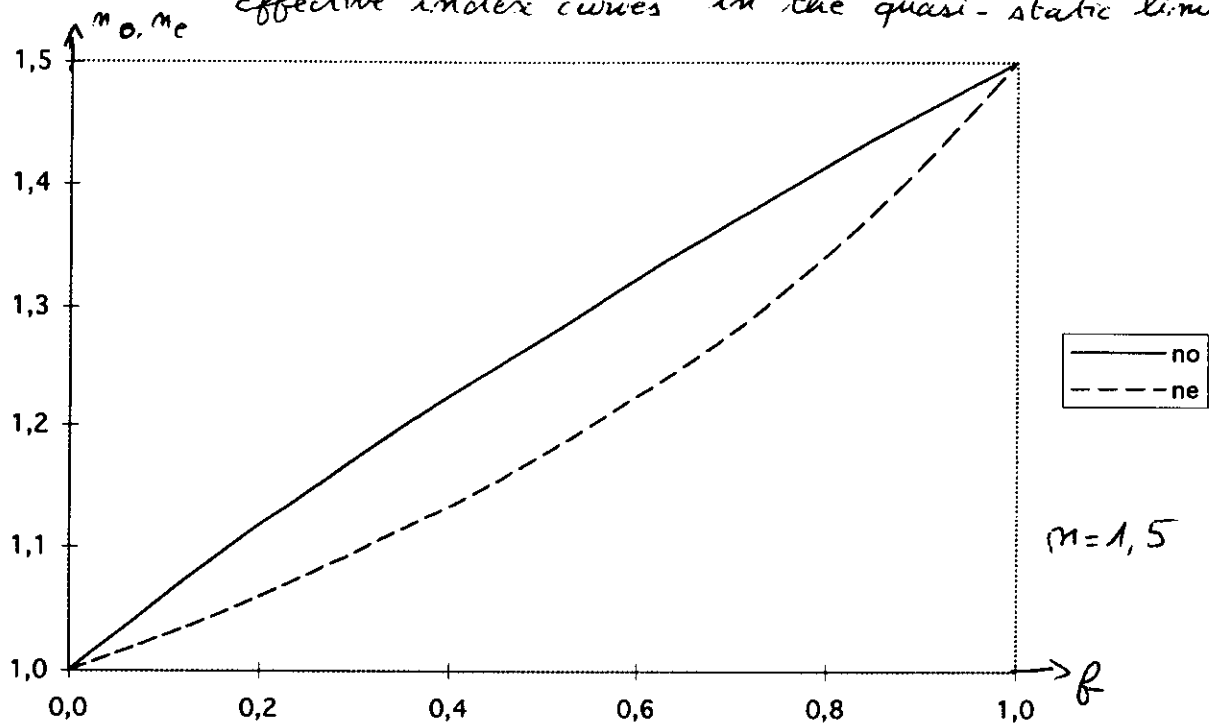
14 AUGUST 1989

Approved for public release; distribution is unlimited.

LEXINGTON

MASSACHUSETTS

Complément: Courbes d'indices effectif dans la limite quasi-statique  
 Effective index curves in the quasi-static limit.





## TABLE OF CONTENTS

ABSTRACT	iii
LIST OF ILLUSTRATIONS	vii
LIST OF TABLES	ix
1. INTRODUCTION	1
2. THEORY	3
2.1 Diffraction Grating	3
2.2 Arbitrary Phase Profile	7
3. MULTI-LEVEL STRUCTURES	13
4. MULTI-LEVEL FABRICATION	17
4.1 Grating Fabrication	18
4.2 Arbitrary Phase Fabrication	21
5. APPLICATIONS OF MULTI-LEVEL DIFFRACTIVE PROFILES	25
5.1 Diffractive Lens	25
5.2 Refractive Diffractive Elements	27
5.3 Spherical Aberration Correction	31
5.4 Limitations of Refractive Diffractive Elements	35
6. DESIGNING DIFFRACTIVE PHASE PROFILES USING CODE V	41
7. SUMMARY	47

<b>Figure No.</b>		<b>Page</b>
5-7	(a) The phase aberration and (b) point spread function of a refractive silicon lens	36
5-8	(a) The phase aberration and (b) point spread function of a diffractively corrected silicon lens	38
5-9	MTF curves for the silicon lens, with and without diffractive correction	39
5-10	Examples of the chromatic and spherical aberration reduction possible by using a diffractive corrector	39
6-1	Recording setup for producing an optically generated holographic element	42

## LIST OF ILLUSTRATIONS

Figure No.		Page
2-1	Surface relief phase grating	4
2-2	Plot of the diffraction efficiency as a function of wavelength	6
2-3	Plot of diffraction efficiency as a function of diffraction order for various wavelengths	8
2-4	Phase functions of a prism and grating	9
2-5	Comparison of the refractive and diffractive phase functions of an arbitrary phase profile	11
2-6	The diffractive phase $\phi(x)$ plotted as a function of the refractive phase $\phi(x)$	12
3-1	A continuous phase grating compared with 2, 4, and 8 discrete phase levels	13
3-2	A multi-level phase structure can be analyzed by representing it as the difference of two continuous phase profiles	14
4-1	Illustration of the fabrication of a binary surface relief grating	19
4-2	Illustration of the fabrication of a 4-level surface relief grating	20
4-3	Example of a phase function that contains a local maximum	23
4-4	Summary of the procedure for determining transition point locations and etch depths	24
5-1	Illustration of a one-dimensional diffractive lens	26
5-2	The phase aberration (a) of a refractive fused silica lens, and (b) of the same lens with diffractive aberration correction	29
5-3	Experimental imaging results of the fused silica lens, with and without diffractive aberration correction	30
5-4	Theoretical phase error due to spherical aberration of a fused silica lens with and without diffractive correction	32
5-5	Experimentally measured focal points of the fused silica lens with and without diffractive correction	34
5-6	Imaging results for the fused silica lens with and without diffractive spherical aberration correction	34



## LIST OF TABLES

Table No.		Page
2-1	Average diffraction efficiency for various fractional bandwidths	7
3-1	Multi-level diffraction efficiency for various numbers of phase levels	15

## 1. INTRODUCTION

The direction of propagation of a light ray can be changed by three basic means: reflection, refraction, and diffraction. Three simple, yet useful, equations that describe these phenomena are the law of reflection, Snell's law, and the grating equation. These three fundamental equations are the foundation for the description of redirecting light rays.

Virtually all optical systems in existence rely on only reflection and refraction to achieve the desired optical transformation. Lens design, based on reflective and refractive elements, is a well-established and refined process. Until recently, diffractive elements have been neglected as viable components of optical systems.

One reason for the lack of interest in using diffractive elements in a lens design is that the process of diffraction does not simply redirect a light ray. Diffraction, unlike reflection and refraction, splits a light ray into many rays — each of which is redirected at a different angle. The percentage of the incident light redirected by the desired angle is referred to as the diffraction efficiency. The diffraction efficiency of a diffractive element is determined by the element's surface profile. If the light that is not redirected by the desired angle is substantial, the result will be an intolerable amount of scatter in the image or output plane of the optical system.

Fortunately, a surface profile exists (in theory) that achieves 100-percent diffraction efficiency at a specified wavelength. The theoretical diffraction efficiency of this surface profile is also relatively insensitive to a change in wavelength. This profile could therefore be used in optical systems operating over finite wavelength bands. Section 2 of this report discusses a theory of this highly efficient diffractive profile. The diffraction efficiency of the simplest example of a diffractive element, a grating, is derived. The section concludes with the extension of the results to diffractive elements having arbitrary phase profiles.

The theoretical existence of a surface profile having high diffraction efficiency has no practical consequences in the design of optical systems unless this profile can be easily determined and readily fabricated. The diffractive surface profile described in Section 2 is not easily fabricated. It is possible, however, to readily fabricate diffractive phase profiles that approximate the ideal diffractive profile. The ideal profile can be approximated in a discrete fashion, similar to the digital representation of an analog function. This discrete representation is called a multi-level profile and is theoretically analyzed in Section 3.

If diffractive surfaces are to become an accepted alternative to reflective and refractive surfaces, a well-defined process of actually fabricating the diffractive surface is needed. Section 4 describes a fabrication process that starts with a mathematical phase description of a diffractive phase profile and results in a fabricated multi-level diffractive surface. The fabrication process is best described in two different steps. The first step, described in detail, is to take the mathematical phase expression and generate from it a set of masks that contains the phase profile information. The second step is to transfer the phase profile information from the masks into the surface of the element specified by the lens design. This particular step is explained in a brief fashion, since the details of this procedure can be found in another report currently in preparation.

A multi-level diffractive phase profile is an additional option that should be seriously considered by lens designers. These profiles are not the solution to all problems; yet, in many instances, they can be used to improve on a design that consists solely of reflective and refractive elements. Section 5 describes some basic examples of cases where a diffractive phase profile can improve on the performance of a completely refractive design. The limitations of these diffractive phase profiles are quantified in order to give the lens designer a sense of the realm of applicability of diffractive profiles.

A lens designer relies heavily on the capabilities of a lens design program in arriving at a suitable solution to a particular problem. If a diffractive phase profile is to be considered in a design, the lens design program must have the capability to insert and optimize these diffractive profiles. Widely distributed lens design programs such as CODE V and ACCOS have the ability to insert diffractive surfaces into lens systems. These programs also have the capability to optimize the phase profiles of the diffractive surfaces in order to attain the best possible performance.

Section 6 describes in detail the process of inserting a diffractive surface in a lens design and the optimization of the diffractive element's phase profile. The format and terminology of the lens design program, CODE V, were chosen as the basis for the description of the procedure. CODE V was chosen because it is not only the most widely available program with the required capability, but also it is the program with which we are most familiar. The reader not familiar with CODE V can gain some insight into the process and peculiarities of designing a lens that contains diffractive surfaces.

## 2. THEORY

### 2.1 DIFFRACTION GRATING

The simplest example of a diffractive optical element is a linear grating. A variety of different types of gratings can be categorized based on the way by which the grating modulates the incident light field. Amplitude gratings, for example, modulate the incident light field by transmitting a certain percentage of the incident light and either absorbing or reflecting the rest. Phase gratings, on the other hand, transmit all of the incident light. The modulation is achieved by imparting to the incident light field a periodic phase delay. This periodic phase delay can be accomplished, as in a volume grating, by periodically modulating the refractive index of a material; or it can be accomplished, as in a surface relief grating, by periodically changing the physical thickness of a material.

Phase gratings have the advantage over amplitude gratings in that they can be made to diffract 100 percent of the incident light (of a given wavelength) into one diffraction order. This is a desirable, if not necessary, condition if a diffractive element is to be used in an optical system. Surface relief gratings have the advantage over volume gratings in that the diffraction efficiency falloff as a function of wavelength is minimized. This is a requirement if the element is to be used in an optical system designed to operate over a finite wavelength band. Furthermore, surface relief gratings can be fabricated in a mass-production environment similar to the integrated circuit fabrication process. For these reasons, we have concluded that surface relief phase gratings have the most potential for finding their way into commercial and military optical systems. The rest of this report will focus only on surface relief structures.

A surface relief phase grating is shown in Figure 2-1. The surface relief pattern, familiar to most people, is that of a conventional blazed grating. In order to analyze this structure, we will assume that the grating period  $T$  is large enough compared with the wavelength of the incident light so that the scalar approximation to Maxwell's equations can be used. The scalar theory is, in general, accurate when the grating period is greater than five wavelengths. The possible applications of diffractive structures, discussed later in this report, fall well within the regime of validity of the scalar approximation.

In the scalar approximation, the transmittance of the surface relief grating in Figure 2-1 can be described by

$$t(x) = \sum_{m=-\infty}^{\infty} \delta(x - mT) * \text{rect}\left(\frac{x}{T}\right) \exp(i2\pi\beta x) \quad (2.1)$$

where  $\beta = (n - 1)d/\lambda T$  and  $*$  represents a convolution.

For an incident plane wave traveling in the  $z$ -direction, the far-field amplitude distribution is given by the Fourier transform  $F(f)$  of the grating transmittance function  $t(x)$

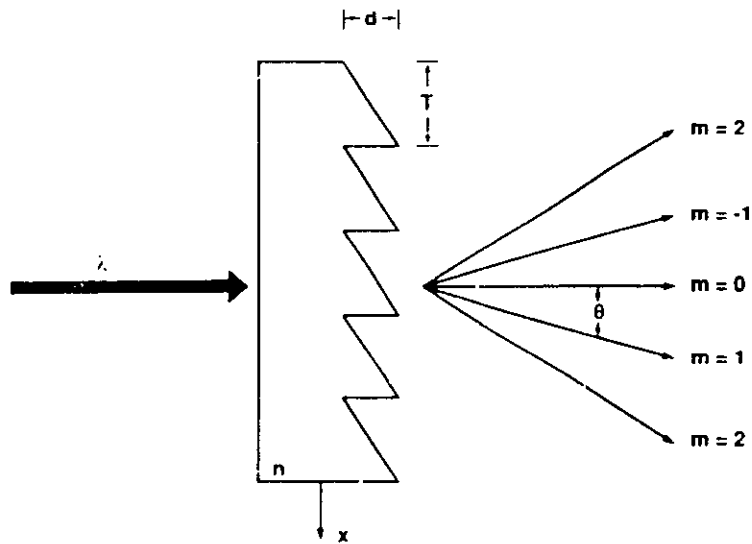


Figure 2-1. Surface relief phase grating.

$$F(f) = \sum_{m=-\infty}^{\infty} \delta\left(f - \frac{m}{T}\right) \frac{\sin(\pi T(\beta - f))}{\pi T(\beta - f)} \quad (2.2)$$

where  $f = \sin(\theta)/\lambda$ . It is apparent from Equation (2.2) that the amplitude of the  $m^{\text{th}}$  diffraction order is given by

$$a_m = \frac{\sin\left(\pi T\left(\beta - \frac{m}{T}\right)\right)}{\pi T\left(\beta - \frac{m}{T}\right)} = \frac{\sin\left(\pi T\left(\beta - \frac{m}{T}\right)\right)}{\pi T\left(\beta - \frac{m}{T}\right)} \quad (2.3)$$

The diffraction efficiency of the  $m^{\text{th}}$  order is the absolute value of the amplitude of the  $m^{\text{th}}$  order squared

$$\eta_m = \left[ \frac{\sin\left(\pi T\left(\beta - \frac{m}{T}\right)\right)}{\pi T\left(\beta - \frac{m}{T}\right)} \right]^2 \quad (2.4)$$

The diffraction order of interest, in general, is the first diffraction order. Setting  $m = 1$  in Equation (2.4), the diffraction efficiency of the first order is given by

$$\eta_1 = \left[ \frac{\sin(\pi(\beta T - 1))}{\pi(\beta T - 1)} \right]^2 \quad (2.5)$$

This equation predicts that, when  $\beta = 1/T$ , the diffraction efficiency of the first order will be 100 percent. Therefore, a properly constructed surface relief phase grating can diffract all of the incident light of a given wavelength into the first diffraction order.

Equation (2.5) also predicts that the first-order diffraction efficiency is both depth and wavelength dependent. A depth error in the fabrication process will result in a lower diffraction efficiency. Likewise, a change in wavelength will result in a diffraction efficiency decrease.

The depth dependence of the diffraction efficiency can be modeled by assuming a depth error of  $\epsilon d$ . The total grating depth is then

$$d = (1 + \epsilon) \frac{\lambda_0}{(n - 1)} \quad (2.6)$$

Substituting this value of  $d$  in Equation (2.5) results in a first-order diffraction efficiency given by

$$\eta_1 = \left[ \frac{\sin(\pi\epsilon)}{\pi\epsilon} \right]^2 \quad (2.7)$$

This equation predicts that a  $\pm 5$ -percent depth error results in a diffraction efficiency falloff of less than 1 percent. In the majority of applications, this can be considered negligible. A depth error of  $\pm 5$  percent corresponds to a physical depth error of approximately  $\pm 500$  Angstroms for visible light. The etching technology used to fabricate these structures can be controlled to achieve depth tolerances of better than  $\pm 500$  Angstroms.

The wavelength dependence of these elements becomes a concern when the element is to be used in an optical system operating over a finite wavelength band. There are, in fact, two wavelength-dependent effects unique to these structures. The first well-known effect, apparent from Equation (2.2) is that the first-order diffraction angle is wavelength dependent. Longer wavelengths are diffracted over larger angles than shorter wavelengths. This is a chromatic dispersion effect that will be discussed later in this report. The second effect is the wavelength dependence of the first-order diffraction efficiency.

Returning to Equation (2.5) and assuming that the diffraction efficiency is maximized for a wavelength  $\lambda_0$  by setting  $d = \lambda_0 / (n - 1)$  result in

$$\eta_1 = \left[ \frac{\sin(\pi(\frac{\lambda_0}{\lambda} - 1))}{\pi(\frac{\lambda_0}{\lambda} - 1)} \right]^2 \quad (2.8)$$

This equation expresses the first-order diffraction efficiency at wavelength  $\lambda$  of an element optimized for wavelength  $\lambda_0$ . It is apparent from Figure 2-2, a plot of the diffraction efficiency as a function of wavelength, that the diffraction efficiency falloff is small for wavelengths close to  $\lambda_0$  and is significant for large wavelength deviations.

For optical systems designed to operate in finite spectral bands, the integrated diffraction efficiency over the spectral band is the parameter of interest. The average diffraction efficiency over a finite bandwidth  $\lambda_0 \pm \Delta\lambda$  is given by

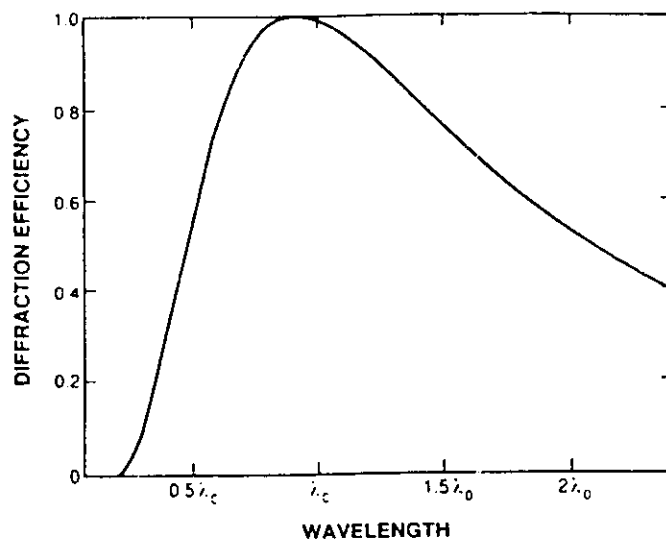


Figure 2-2. Plot of the diffraction efficiency as a function of wavelength.

$$\bar{\eta}_1 = \frac{1}{\Delta\lambda} \int_{\Delta\lambda} \eta_1(\lambda) d\lambda \quad (2.9)$$

which is approximately expressed by

$$\bar{\eta}_1 \approx \left[ 1 - \left( \frac{\pi \Delta\lambda}{6\lambda_0} \right)^2 \right] \quad (2.10)$$

Table 2-1 lists the average diffraction efficiency over various fractional bandwidths. The average diffraction efficiency remains above 95 percent for fractional bandwidths of up to 40 percent, but falls off rapidly for larger bandwidths. This effect is the most limiting constraint in using a diffractive element in a finite bandwidth system. The decrease in efficiency as a function of bandwidth has to be considered in a system design. The residual light that is not diffracted into the desired order is diffracted into different orders. This light manifests itself as a type of scatter at the image plane of an optical system. The amount of tolerable scatter is particular to the optical system's performance requirements. The lens designer has to establish the advantage or disadvantage of introducing a diffractive element into a design based on the performance goals of the optical system.

The scatter as a function of bandwidth introduced by a diffractive element is unlike the more familiar random scatter caused by inadequate surface polishing or surface defects. The scatter caused by the diffractive surface is deterministic. This scatter, or residual light, propagates in different diffraction orders. The amount of light at any wavelength, and in any given order, can be easily calculated from Equation (2.4). Figure 2-3 shows the amount of light in the various diffraction orders, at various wavelengths, for an element designed to have a 100-percent efficient

TABLE 2-1.

Average Diffraction Efficiency for Various Fractional Bandwidths

Fractional Bandwidth ( $\Delta\lambda/\lambda_0$ )	Efficiency $\eta_1$
0.00	1.000
0.10	0.997
0.20	0.989
0.30	0.975
0.40	0.956
0.50	0.931
0.60	0.901

first order at  $\lambda_0$ . As the wavelength increases from  $\lambda_0$ , the residual light appears most strongly in the zero order. For decreasing wavelengths, the light appears in the second order. This residual light can be traced through an optical system to see how it is distributed at the image plane.

The blazed phase grating analyzed in this section was described by the transmittance function given in Equation (2.1). An equivalent way of expressing the transmittance function of a blazed grating is

$$t(x) = e^{i2\pi |x|_0 / \alpha} \quad (2.11)$$

where  $|x|_0 / \alpha$  represents a linear function in  $x$ , modulo  $\alpha$ , limited to values between  $\pm\alpha/2$ .

The transmittance of a prism (the refractive counterpart of a grating) can be expressed as

$$t(x) = e^{i\frac{2\pi}{\lambda} \alpha_0 x} \quad (2.12)$$

The prism and grating phase functions are shown in Figure 2-4.

## 2.2 ARBITRARY PHASE PROFILE

In the previous section, the theory of a blazed phase grating having 100-percent diffraction efficiency in the first order was given. The effects of a change in depth, or a change in wavelength, were analyzed. However, a diffraction grating is of extremely limited usefulness in optical systems which require elements that focus or reshape the wavefront by desired amounts. In conventional optical systems, this is done refractively or reflectively by employing lenses and mirrors. What is required is the diffractive counterpart to these conventional optical elements.



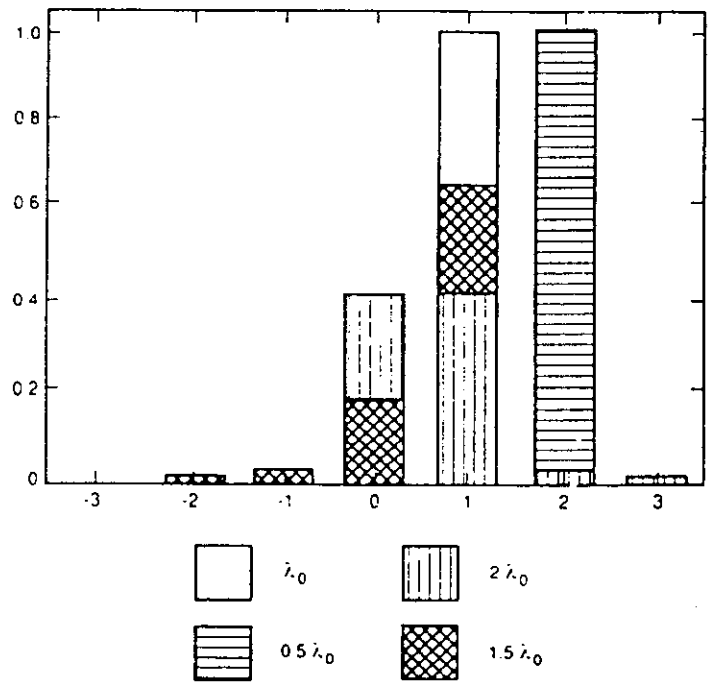
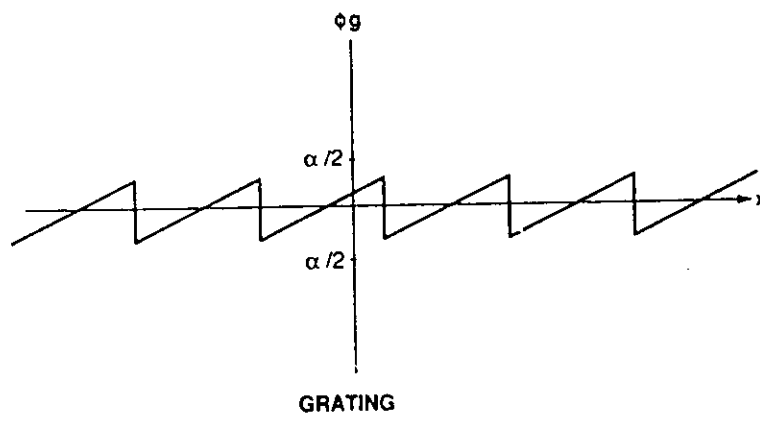
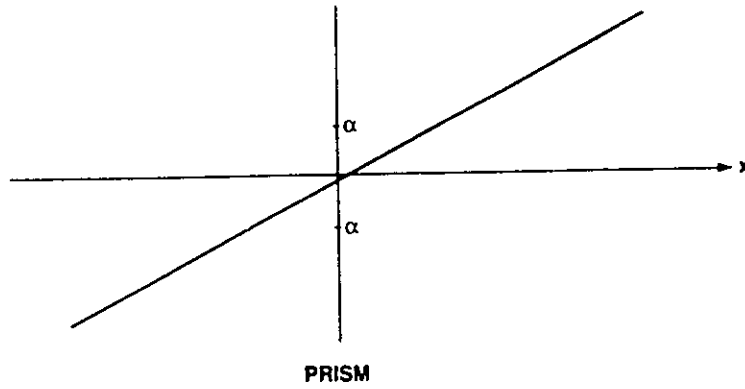


Figure 2-3. Plot of diffraction efficiency as a function of diffraction order for various wavelengths.

124931-9



124931-8

Figure 2-4. Phase functions of a prism and grating.

Consider a refractive element that can be described by a transmittance function

$$t_r(x) = e^{i2\pi\phi(x)} \quad (2.13)$$

where  $\phi(x)$  is an arbitrary function of  $x$ . Can a general analogy, like the grating-prism analogy of the previous section, be made? What is the behavior of the diffractive counterpart with a transmittance function of

$$t_d(x) = e^{i2\pi\phi'(x)} \quad (2.14)$$

where  $\phi'(x) = |\phi(x)|_\alpha$ ? The refractive and diffractive phase functions for an arbitrary phase are plotted in Figure 2-5.

In order to understand the diffractive transmittance function of Equation (2.14), a nonlinear limiter analysis is used. The diffractive phase  $\phi'(x)$  is plotted as a function of the refractive phase  $\phi(x)$  in Figure 2-6. The diffractive phase, for generality, has been limited to values between  $\pm\alpha/2$ . It is apparent from the figure that  $\phi'(x)$  is periodic in  $\phi(x)$  with a period equal to one. It follows that  $\exp\{i2\pi\phi'(x)\}$  is also periodic in  $\phi(x)$  and can therefore be written as a generalized Fourier series

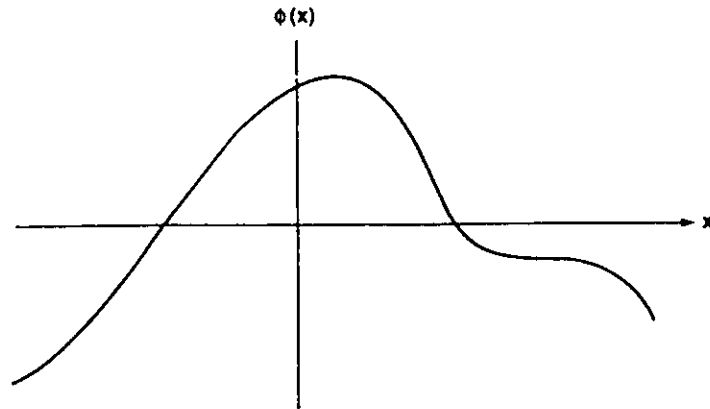
$$e^{i2\pi\phi'(x)} = \sum_{m=-\infty}^{\infty} c_m e^{i2\pi m\phi(x)} \quad (2.15)$$

where the coefficients  $c_m$  are given by

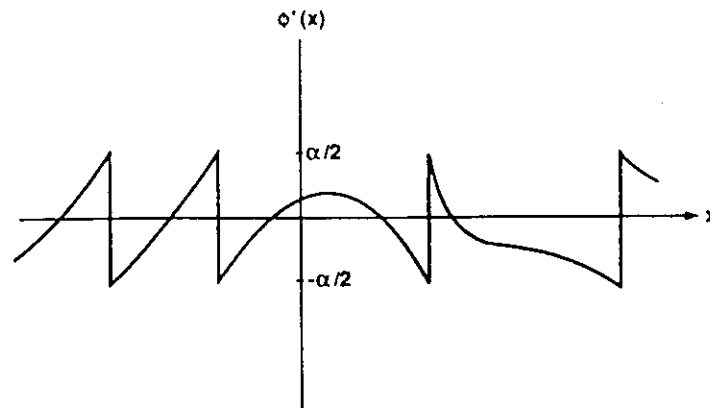
$$c_m = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi(\alpha-m)\phi(x)} d\phi(x) = \frac{\sin(\pi(\alpha-m))}{\pi(\alpha-m)}. \quad (2.16)$$

Therefore, if  $\alpha = 1$ ,  $c_1$  is equal to 1, and all the other  $c_m$  coefficients are zero. The exiting wavefront from this diffractive structure is identical to its refractive counterpart. It is interesting to note that the  $c_m$  coefficients are identical to the  $a_m$  coefficients of Equation (2.3) by setting  $\alpha = (n-1)d/\lambda_0$ .

The wavelength and depth dependence of diffraction efficiency for any arbitrary phase diffractive structure is identical to the linear phase grating. The arbitrary diffractive phase element has *orders* similar to the grating. These *orders*, instead of being plane waves, take on more complicated wavefront profiles represented on the right-hand side of Equation (2.15).



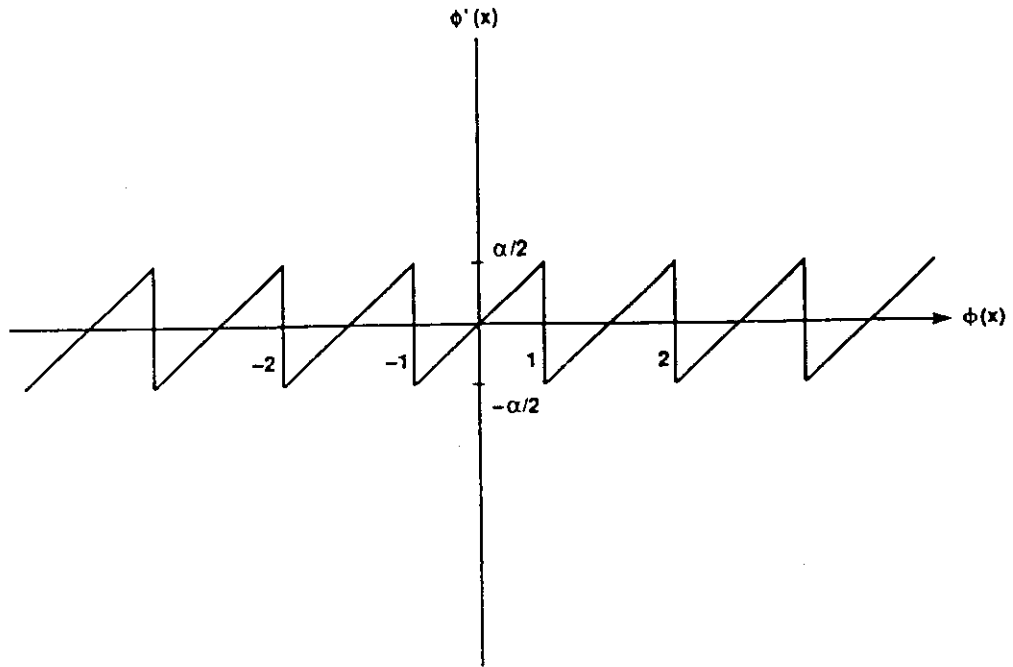
REFRACTIVE PHASE



DIFFRACTIVE PHASE

124931-7

Figure 2-5. Comparison of the refractive and diffractive phase functions of an arbitrary phase profile.



124931-6

Figure 2-6. The diffractive phase  $\phi'(x)$  plotted as a function of the refractive phase  $\phi(x)$ .

### 3. MULTI-LEVEL STRUCTURES

In Section 2 we showed that an arbitrary wavefront can be produced from a diffractive structure with 100-percent diffraction efficiency at the design wavelength. Unfortunately, this diffractive structure has a surface relief depth which varies continuously over every  $2\pi$  phase interval. This phase profile, with a continuous depth, is not easily fabricated with any existing technology. A compromise has to be made between achievable diffraction efficiency and ease of fabrication.

A compromise that results in relatively high diffraction efficiency and ease of fabrication is a multi-level phase structure. Figure 3-1 shows a continuous phase grating profile compared with phase gratings with 2, 4, and 8 discrete phase levels. It is apparent from the figure that the larger the number of discrete phase levels, the better the approximation to the continuous phase profile. These multi-level phase profiles can be fabricated using standard semiconductor fabrication techniques. The fabrication process of multi-level structures will be described later in this report. The first question to be answered is the extent of the sacrifice in diffraction efficiency as a function of the number of discrete phase levels.

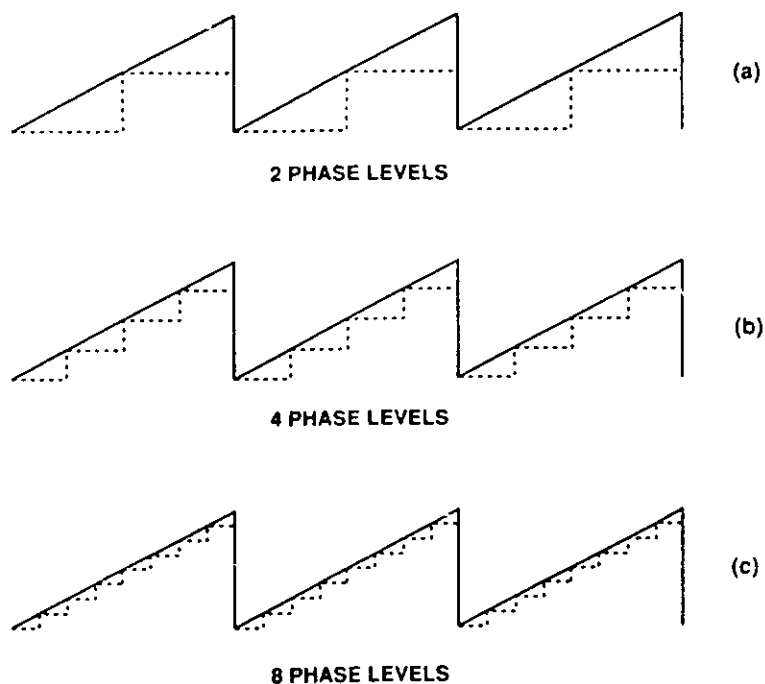


Figure 3-1. A continuous phase grating compared with 2, 4, and 8 discrete phase levels.

The diffraction efficiency of a multi-level structure can be simply derived by considering the multi-level structure as being equal to the desired continuous profile minus an error phase profile. The diffraction efficiency of the multi-level structure is then the efficiency of the desired continuous phase profile multiplied by the zero-order efficiency of the error phase structure. Figure 3-2 illustrates this concept for a 4-level structure. If the number of discrete phase levels of the multi-level structure is  $N$ , then the error phase structure to be subtracted has a depth of  $d/N$  (where  $d$  is the desired continuous phase depth) and a periodicity of  $1/N$  times that of the ideal structure. The resulting diffraction efficiency of the multi-level structure is then easily obtained by using Equation (2.4) and is given by

$$\eta_m^N = \left[ \frac{\sin(\pi(\frac{(n-1)d}{\lambda} - m))}{\pi(\frac{(n-1)d}{\lambda} - m)} \right]^2 \left[ \frac{\sin(\pi(\frac{(n-1)d}{\lambda N}))}{\pi(\frac{(n-1)d}{\lambda N})} \right]^2 \quad (3.1)$$

This equation can be used to determine the diffraction efficiency of any multi-level profile at any wavelength and for any diffraction order.

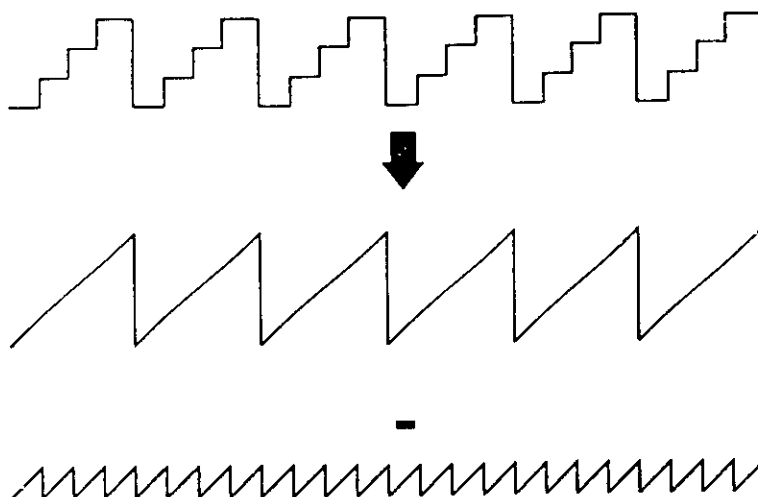


Figure 3-2. A multi-level phase structure can be analyzed by representing it as the difference of two continuous phase profiles.

As an example of how the number of phase levels affects the diffraction efficiency, consider a continuous phase structure designed to achieve 100-percent diffraction efficiency in the first order at the design wavelength. Equation (3.1) then reduces to the expression

$$\eta_1^N = \left[ \frac{\sin(\pi/N)}{\pi/N} \right]^2 \quad (3.2)$$

The continuous phase profile would achieve 100-percent diffraction efficiency, whereas a multi-level structure with  $N$  levels is reduced to that given by Equation (3.2). Table 3-1 lists the diffraction efficiency of a multi-level structure for various values of the number of phase levels. Two things to notice in the table are that for 16 phase levels the diffraction efficiency at the design wavelength is 99 percent, and that values of diffraction efficiency are highlighted for multi-level structures with  $N$  equal to a power of 2. The reason multi-level structures with a number of levels equal to a power of 2 are highlighted will become apparent in the next section.

TABLE 3-1.  
Multi-level Diffraction Efficiency for Various Numbers of Phase Levels

Number of Levels $N$	First-Order Efficiency $\eta_1^N$
2	0.41
3	0.68
4	0.81
5	0.87
6	0.91
8	0.95
12	0.98
16	0.99

A 16-phase level structure achieving 99-percent diffraction efficiency is an element that could have advantageous implications in the design of many optical systems. The residual 1 percent of the light is diffracted into higher orders and manifests itself as scatter. In many optical systems, this is a tolerable amount of scatter. The fabrication of a 16-phase level structure, described in the following section, is relatively efficient due to the fact that only four processing iterations are required to produce the element.



#### 4. MULTI-LEVEL FABRICATION

The fabrication of a multi-level diffractive element requires the same technology used in the production of integrated circuits. This fabrication process will be outlined in order to describe in a general fashion the steps involved in producing a multi-level diffractive element. A separate report describing in detail all the equipment and processing steps used in fabrication of multi-level elements is in preparation.

The first step involved in fabricating a multi-level element is to mathematically describe the ideal diffractive phase profile that is to be approximated in a multi-level fashion. The simplest case, for example, is a grating of period  $T$ , designed to operate at a wavelength  $\lambda_0$ . The phase function for this grating can be mathematically described by

$$\phi(x) = \frac{2\pi}{\lambda_0} ax \quad (4.1)$$

where  $a = \lambda_0 / T$ .

A phase function having more complexity than a simple grating can be mathematically described in a general way by expanding it in a power series

$$\phi(x, y) = \frac{2\pi}{\lambda_0} \sum_{n,m} a_{nm} x^n y^m \quad (4.2)$$

This equation represents a general phase function in the spatial coordinates  $(x, y)$ . The number of terms retained in the power series determines how well of an approximation the series is to the actual phase desired. The values of the  $a_{nm}$  coefficients are optimized to make the series expansion best approximate the desired phase. For example, the grating phase of Equation (4.1) is represented by Equation (4.2) where all the  $a_{nm}$  coefficients are zero, except for  $a_{10}$  which is equal to  $\lambda_0 / T$ .

The majority of cases in optical design require phase functions that are circularly symmetric; these phase functions can also be described by a power series expansion

$$\phi(r) = \frac{2\pi}{\lambda_0} \sum_p a_p r^p \quad (4.3)$$

where  $r$  is the radial coordinate. The optical axis of the lens system is at the radial coordinate  $r = 0$ . The values of the  $a_p$  coefficients determine the functional form of the radially dependent phase.

The next step in the fabrication process, once the phase function is mathematically determined, is to create a set of lithographic masks which are produced by standard pattern generators used in the integrated circuit industry. Pattern generators, either optical or electron beam, expose a thin layer of photoresist which resides on a chrome-covered quartz substrate. The exposed photoresist is then washed off the chrome-coated substrate, leaving the pattern in the remaining unexposed photoresist. The pattern is then transferred to the chrome by etching away the chrome that is not covered by the remaining photoresist. Once the chrome has been patterned, the remaining

photoresist is washed away, resulting in a finished lithographic mask. The final product is a binary amplitude mask that transmits light where the pattern was exposed, and reflects any incident light where there was no exposure.

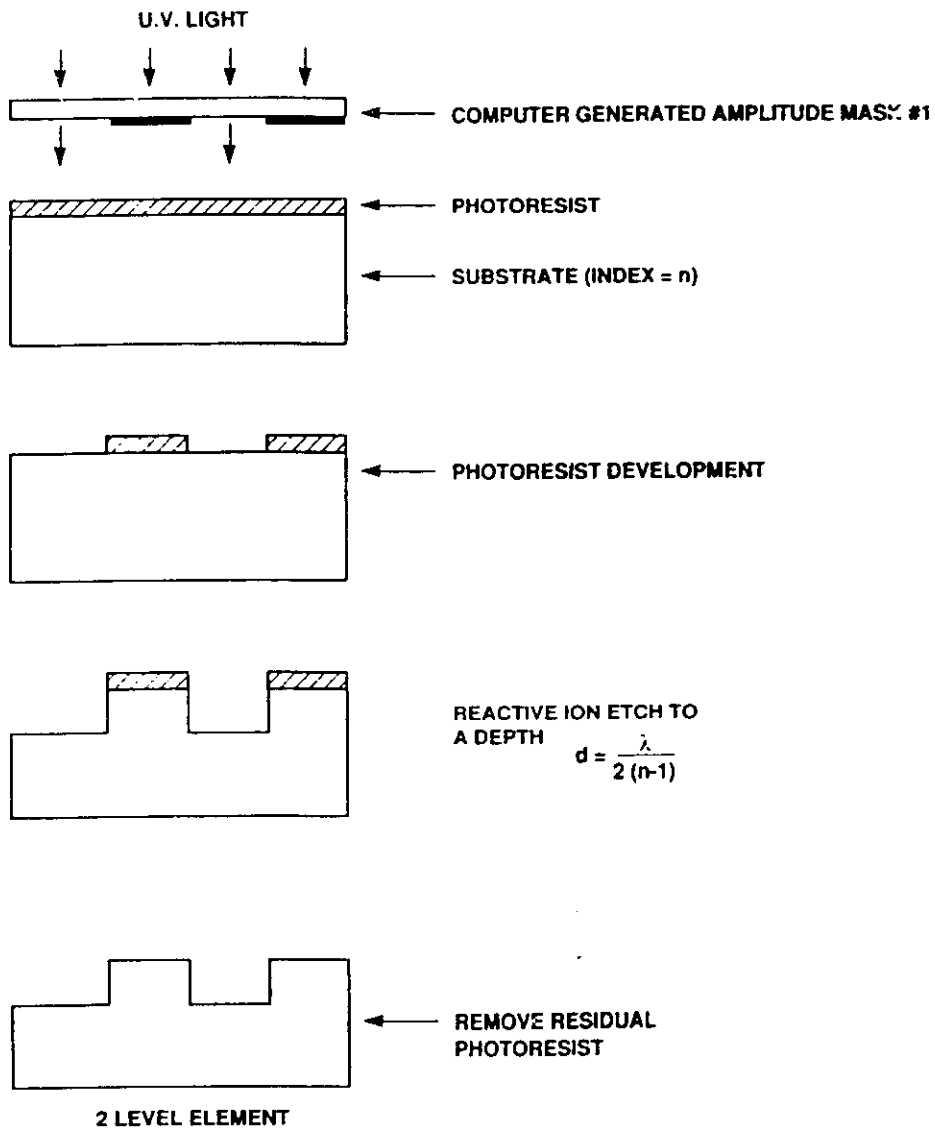
#### 4.1 GRATING FABRICATION

To illustrate the purpose of these lithographic masks in the fabrication of a multi-level element, consider the simplest case of a grating. The final grating that is desired has the surface relief profile shown in Figure 3-1(a). The coarsest approximation to the desired grating profile, also shown in Figure 3-1(a), is a binary phase profile. A lithographic mask can be easily produced with a binary amplitude grating pattern having the desired period and a 50-percent duty cycle (i.e., 50 percent of the light is transmitted). What remains is to transfer the amplitude pattern contained on the lithographic mask onto an optically transmissive substrate, and convert the amplitude pattern into a surface relief pattern.

Figure 4-1 illustrates the process of fabricating a binary surface relief grating, starting with the binary amplitude lithographic mask. A substrate of the desired material is coated with a thin layer of photoresist. The lithographic mask is then placed in intimate contact with the substrate and illuminated from above with an ultraviolet exposure lamp. The photoresist is developed, washing away the exposed resist and leaving the binary grating pattern in the remaining photoresist. This photoresist will act as an etch stop, like in the lithographic mask process, except that now the substrate material has to be etched instead of chrome.

The most reliable and accurate way to etch many optical substrate materials is to use reactive ion etching. The process of reactive ion etching anisotropically etches materials at very repeatable rates. The desired etch depth can be obtained very accurately. The anisotropic nature of the process assures a vertical etch, resulting in a truly binary surface relief profile. Once the substrate has been reactively ion etched to the desired depth, the remaining photoresist is stripped away, leaving a binary phase surface relief grating. In the case of a binary phase profile, Equation (3.2) predicts a maximum first-order diffraction efficiency of 40.5 percent for an etch depth of  $d = \lambda_0/2(n - 1)$ .

Imagine repeating the process described above on the same substrate, except this time using a lithographic mask having twice the period of the first mask. Figure 4-2 illustrates this process. The binary phase element is recoated with photoresist and exposed using the lithographic mask #2 that has a period twice that of the first mask. After developing and washing away the exposed photoresist, the substrate is reactive ion etched to a depth half that of the first etch [i.e.,  $d = \lambda_0/4(n - 1)$ ]. Removal of the remaining photoresist results in a 4-level approximation to the desired profile. The 4-level phase element has a first-order diffraction efficiency, predicted by Equation (3.2), of 81 percent. One can imagine repeating the process a third and fourth time with lithographic masks having periods of one-quarter and one-eighth that of the first mask, and etching the substrate to depths of one-quarter and one-eighth the depth of the first etch. The successive etches result in elements having 8 and 16 phase levels. The first-order diffraction efficiency, after the third and fourth etches, is predicted from Equation (3.2) to be 95 and 99 percent, respectively.



124931-3

Figure 4-1. Illustration of the fabrication of a binary surface relief grating.

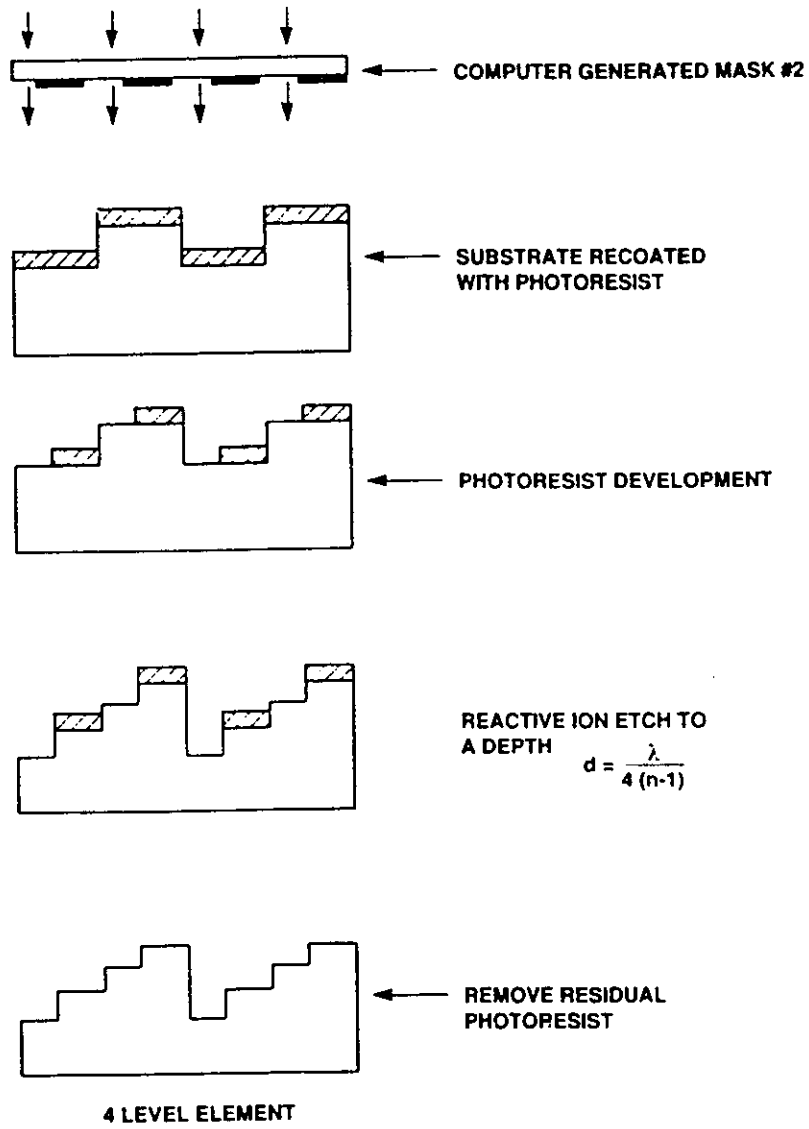


Figure 4-2. Illustration of the fabrication of a 4-level surface relief grating.

124831-2

After only four processing iterations, a 16-phase level approximation to the continuous case can be obtained. A similar iteration process to that described here is used in the fabrication of integrated circuits. The process can be carried out in parallel, producing many elements simultaneously, in a cost-effective manner.

The one major difference between fabricating a binary phase element and a multi-level element is that, after the first etching step, the second and subsequent lithographic masks have to be accurately aligned to the existing pattern on the substrate. Alignment is accomplished using another tool standard to the integrated circuit industry, a mask aligner. Mask aligners are commercially available with varying degrees of sophistication. The majority of aligners can place in registry the mask and substrate to submicron tolerances. This degree of alignment accuracy allows for the fabrication of many useful multi-level diffractive elements.

Some instances in a lens design may require or prefer the diffractive surface to reside on a substrate surface that is not flat. The process, as described, necessitated a flat substrate to obtain intimate contact between the lithographic masks and the substrate. The condition of intimate contact can be relaxed, depending on the feature sizes of the lithographic masks. If the mask and substrate are not in intimate contact, the ultraviolet exposure light will diffract from the mask, blurring the pattern in the photoresist. In many instances, particularly for diffractive elements designed for use in the infrared, the mask's feature sizes are large enough to allow for a significant distance between the mask and the substrate. This is a point the lens designer must be aware of in a system design.

In summary, the fabrication of a multi-level surface relief grating requires a set of lithographic masks and standard integrated circuit fabrication equipment. A set of  $M$  properly designed lithographic masks results in a multi-level surface relief grating with  $2^M$  phase levels. The optimum etch depth for the  $M^{\text{th}}$  mask pattern is  $d_M = \lambda_0 / 2^M (n - 1)$ .

## 4.2 ARBITRARY PHASE FABRICATION

The design of a set of lithographic masks used in the fabrication of a multi-level grating was easy to visualize. Mask # $M$  simply had a 50-percent duty cycle and a period equal to one-half that of mask # $(M - 1)$ . The design of a set of lithographic masks used in the fabrication of multi-level structures approximating the general phase functions described by Equations (4.2) and (4.3) is slightly more involved.

Let us consider the case of designing a set of masks to be used in the fabrication of a circularly symmetric phase function described by Equation (4.3). The coarsest approximation to this phase profile is again a binary phase profile. The lithographic mask needed to produce this binary phase element will have a circularly symmetric amplitude profile (i.e., a set of alternately transmitting and reflecting annuli). The positions and widths of these annuli are determined from Equation (4.3). The phase at the center of the pattern is zero. By stepping out in radius, the phase  $\phi(r)$  either increases or decreases. The magnitude of the phase will reach  $\pi$  at some value of  $r$  which is the first radial position on the lithographic mask where an amplitude transition occurs.

Continuing the process of stepping out in radius results in radial locations where the phase function  $\phi(r)$  takes on values that are integer multiples of  $\pi$ . These are the subsequent radial positions where amplitude transitions occur on the first lithographic mask. The process of stepping out in radius is continued until the maximum radial value of the pattern to be written is reached.

The resulting set of radial values is sufficient information to write the lithographic mask. A computer program called Mann 53, and written by the Binary Optics Group at Lincoln Laboratory, is able to take the set of radial values and properly format the data such that they can be read by a Mebes electron beam pattern generator. The Mann 53 program creates a data tape that can be sent to various lithographic mask vendors for mask fabrication.

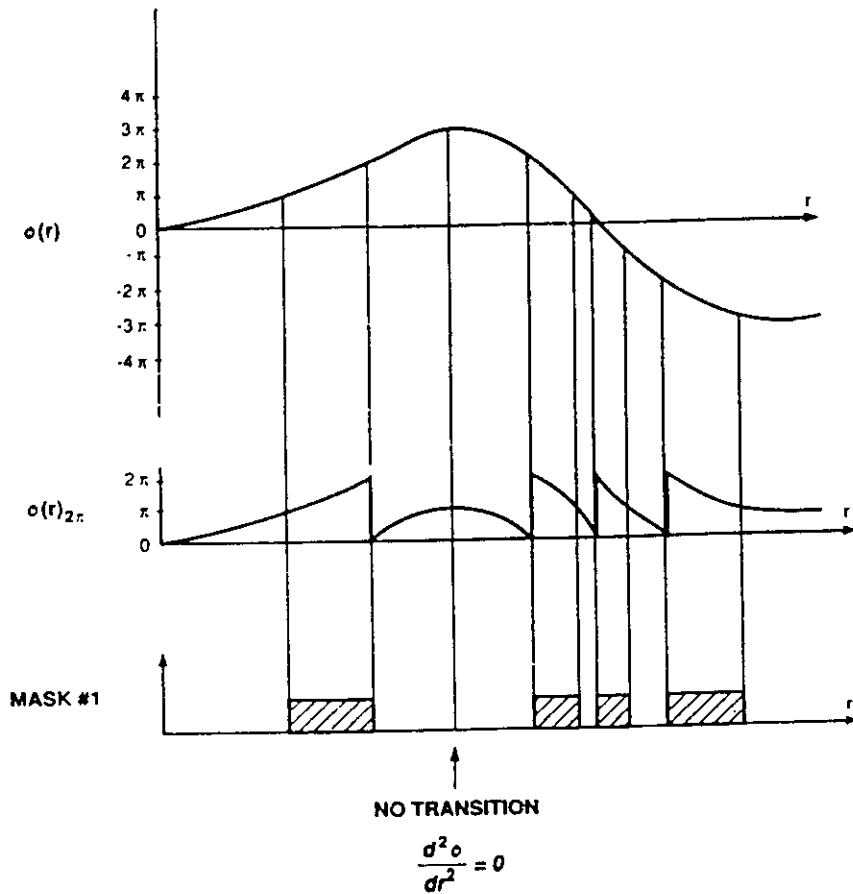
The process of designing mask #M, where M is greater than 1, is carried out in a fashion similar to the first mask. Again, the phase function  $\phi(r)$  is monitored as a function of radius. At some radial distance, the magnitude of the phase will reach the value  $\pi/M$ . This is the radial position where the first amplitude transition will occur on mask #M. Subsequent amplitude transition points for mask #M will occur at radial values where the phase is an integer multiple of  $\pi/M$ . The process of stepping out in radius is continued until the maximum radius of the pattern is reached. The Mann 53 program is used to format these radial values and produce data tapes in a manner similar to the case of mask #1.

The process described above is straightforward and applicable to any radially symmetric phase profile, except for a certain subset of phase profiles where an anomaly can occur. This subset of phase functions is one where the first derivative of the phase with respect to r is zero at some radial position that corresponds to a transition point. For this subset of functions, a little more care has to be taken in locating the transition points. If the first derivative  $d\phi/dr$  is zero at a transition point, the question arises whether or not this point should correspond to a transition. The way to determine whether a transition should or should not occur is to check the second derivative. If  $d^2\phi/dr^2$  is also zero at the radial point in question, the phase is at an inflection point and a transition should be located there. On the other hand, if the second derivative is zero, the point corresponds to a local maximum or minimum of the phase, and a transition should not be located there.

Figure 4-3 illustrates an example of a phase function that contains a local maximum at a radial position  $r_3$ ; this position also happens to correspond to a value of the phase equal to  $3\pi$ . Since the second derivative is not zero at this point, a transition point should not be located there.

The procedure described above is also applicable for noncircularly symmetric phase functions described by Equation (4.2). The basic idea is the same, yet the determination of the transition point locations can become quite computationally intensive. Perkin-Elmer Corp. has devised a software package that can take an arbitrary two-dimensional phase profile and construct from it a set of proper lithographic masks.

Once the proper set of lithographic masks is designed and constructed, the fabrication of the multi-level diffractive phase element is identical to the fabrication process of the multi-level phase grating described above. The first mask pattern is reactively ion etched to a  $\pi$  phase depth.



124931-1

Figure 4-3. Example of a phase function that contains a local maximum.

Subsequent mask patterns are aligned and etched to a phase depth of  $\pi/2^M$ . The procedure for determining transition point location and etch depth is summarized in Figure 4-4.

### TRANSITION POINTS

$$\phi(r) = \frac{\pi l}{2^{(m-1)}} \quad \text{WHERE} \quad \left\{ \begin{array}{l} l = 0, \pm 1, \pm 2, \dots \\ m = \text{MASK \#} \end{array} \right.$$

IF  $\frac{d\phi}{dr} = 0$  AT A TRANSITION POINT  $\Rightarrow$  CHECK  $\frac{d^2\phi}{dr^2}$

(1)  $\frac{d^2\phi}{dr^2} \neq 0 \Rightarrow$  NOT A TRANSITION

(2)  $\frac{d^2\phi}{dr^2} = 0 \Rightarrow$  IS A TRANSITION

### ETCH DEPTHS

$$\text{ETCH DEPTH} = \frac{\lambda_0}{2^{(m)}(n-1)}$$

WHERE  $\left\{ \begin{array}{l} m = \text{MASK \#} \\ n = \text{SUBSTRATE INDEX OF REFRACTION} \end{array} \right.$

124931-12

Figure 4-4. Summary of the procedure for determining transition point locations and etch depths.



## 5. APPLICATIONS OF MULTI-LEVEL DIFFRACTIVE PROFILES

The fabrication of multi-level diffractive phase profiles has been described in the previous section. Here, we attempt to elucidate the potential as well as the limitations of using a diffractive surface in the design of an optical system. Hopefully, a lens designer will be able to determine whether or not a multi-level diffractive surface will be advantageous in any particular design.

We begin with a description of the focusing properties of a completely diffractive lens (Section 5.1). A completely diffractive lens is shown to suffer from severe chromatic aberration, limiting its usefulness in any optical system that has to operate over a finite wavelength band.

In Section 5.2 we show how the chromatic dispersion of the diffractive lens can be used to one's advantage by combining it with a refractive lens element. The combination of a refractive lens and a diffractive profile is shown to be a very powerful concept in the design of optical elements.

The idea of using a diffractive profile to correct for the inherent spherical aberration of a single spherical lens is described in Section 5.3. For a monochromatic system, the spherical aberration can be completely eliminated; for a finite waveband system, it can be reduced. The amount of correction possible is shown to depend on the fractional operating bandwidth of the system.

Finally, in Section 5.4 we show how the diffractive pattern that corrects for spherical aberration can be combined with the diffractive pattern that corrects for chromatic aberration. The result of this combination is shown to be a single diffractive pattern that corrects for both spherical and chromatic aberration.

### 5.1 DIFFRACTIVE LENS

The simplest example of a diffractive phase profile, other than a linear grating, is a quadratic phase profile. In the paraxial approximation, a quadratic phase profile is a lens. A one-dimensional diffractive lens, having a quadratic phase profile, is illustrated in Figure 5-1. The lens has a focal length  $F_0$  for wavelength  $\lambda_0$ , and forms an image at  $z_i$  of an object located at  $z_0$ . The transmittance function for this lens, assuming 100-percent diffraction efficiency in the first order for wavelength  $\lambda_0$ , is described by

$$t(x_0) = e^{-i\pi v x_0^2} \quad (5.1)$$

where  $v = 1/\lambda_0 F_0$  is a constant. By performing a Fresnel diffraction calculation, it is shown that the first-order lens equation for a diffractive lens is

$$\frac{1}{z_i} = \lambda v - \frac{1}{z_0} \quad (5.2)$$

From this equation it is apparent that the image distance  $z_i$  is strongly dependent on wavelength. Setting  $v = 1/\lambda_0 F_0$  in Equation (5.2) results in the expression

$$\frac{1}{F(\lambda)} = \frac{1}{z_0} + \frac{1}{z_i} \quad (5.3)$$

where  $F(\lambda) = \lambda_0 F_0 / \lambda$ . This expression looks conspicuously like the first-order lens equation for refractive lenses. The only difference is that the focal length of the lens, instead of being constant, depends inversely on the wavelength. The result of this wavelength dependence is severe chromatic aberration.

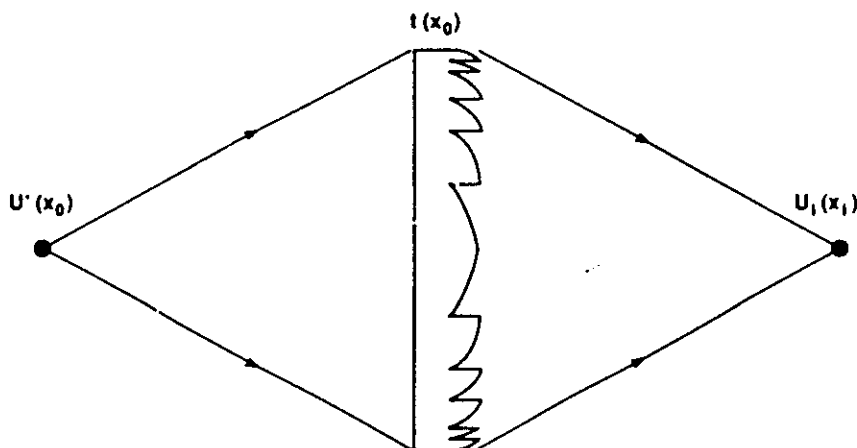


Figure 5-1. Illustration of a one-dimensional diffractive lens.

The amount of chromatic aberration in a diffractive lens can be quantified by setting the object distance to infinity. The output wavefront from the diffractive lens is then described by the lens transmittance function of Equation (5.1). An ideal wavefront, having no chromatic aberration, is given by

$$U_i(x) = e^{-i\frac{\pi x^2}{\lambda F_0}}. \quad (5.4)$$

The output wavefront from the diffractive lens can be rewritten as

$$U_0(x) = U_i(x)U_a(x) = e^{-i\frac{\pi x^2}{\lambda F_0}} e^{-i\frac{\pi}{F_0}(\frac{1}{\lambda_0} - \frac{1}{\lambda})x^2} \quad (5.5)$$

where  $U_i$  is the ideal wavefront and  $U_a$  is the aberration component of the wavefront. The phase function  $\phi_a$  of  $U_a$ ,

$$\phi_a(x) = \frac{1}{2F_0}(\frac{1}{\lambda_0} - \frac{1}{\lambda})x^2 \quad (5.6)$$

is the chromatic phase aberration of a diffractive lens. Note that, for the wavelength  $\lambda = \lambda_0$ , the phase aberration is zero. For any wavelength other than  $\lambda_0$ , the phase aberration is nonzero.

An expression for the maximum amount of chromatic phase aberration present in a diffractive lens of aperture diameter  $A$ , and operating over a bandwidth  $\Delta\lambda$  centered at  $\lambda_0$ , can be written as

$$maro_c = \frac{A}{8(F/\#)} \left( \frac{\Delta\lambda}{\lambda^2} \right) \quad (5.7)$$

where  $F/\#$  is the f-number of the lens. Note that the amount of chromatic aberration is proportional to both the fractional bandwidth and the number of wavelengths across the aperture. As an example, consider an  $F/2$  lens, operating over the 8- to 12- $\mu\text{m}$  wavelength band, and having a 75-mm aperture. The maximum phase error due to chromatic aberration of this lens is 234 waves! This is an intolerable amount of chromatic aberration. Clearly, the usefulness of a completely diffractive lens in an optical system operating over a finite wavelength band is limited.

## 5.2 REFRACTIVE/DIFFRACTIVE ELEMENTS

The previous analysis made it quite apparent that completely diffractive lenses cannot be used in finite wavelength band systems. A solution to this dilemma is to combine a refractive lens with a diffractive lens profile. It should be pointed out that this is not a detraction from using diffractive lens profiles in an optical system — rather, it is an attraction. A completely diffractive lens would have to reside on an optically flat substrate to retain good performance. The cost differential between an optically flat refractive substrate and a refractive lens with spherical surfaces is negligible. There is no cost advantage in using completely diffractive lens elements. Furthermore, the smaller the  $F/\#$  of a diffractive lens, the finer the features become in the diffractive profile and the more difficult the element becomes to construct. By combining a refractive lens and a diffractive lens, the refractive lens can do the majority of the focusing, substantially increasing the feature sizes required in the diffractive lens profile.

The most compelling reason to consider refractive/diffractive elements is that the chromatic dispersion of the diffractive surface can be used to negate the chromatic dispersion of refractive lenses. The etching of a properly designed diffractive lens profile on a surface of a dispersive refractive lens can result in a single-lens element that has virtually no dispersion. This concept is very powerful, especially in wavelength regions where the number of available materials that have suitable transmittance characteristics is limited.

The index of refraction of a refractive lens can be modeled in a linear approximation as

$$n(\lambda) = n_0 - D(\lambda - \lambda_0) \quad (5.8)$$

where  $D$  is the dispersion constant and  $\lambda_0$  is the center wavelength of the wavelength band. The focal length of this dispersive refractive lens,  $F_r$ , is given by

$$\frac{1}{F_r(\lambda)} = \frac{1}{F_{r0}} - \frac{D(\lambda - \lambda_0)}{(n_0 - 1)F_{r0}} \quad (5.9)$$

where  $F_{r0}$  is the focal length of wavelength  $\lambda_0$ . The focal length of a diffractive lens  $F_d$  was previously determined to be

$$F_d(\lambda) = \frac{\lambda_0}{\lambda} F_{d0}. \quad (5.10)$$

A refractive/diffractive combination of the lenses described by Equations (5.9) and (5.10) results in a lens with a focal length  $F(\lambda)$  given by

$$\frac{1}{F(\lambda)} = \frac{1}{F_r(\lambda)} + \frac{1}{F_d(\lambda)}. \quad (5.11)$$

Substituting Equations (5.9) and (5.10) into Equation (5.11) results in

$$\frac{1}{F(\lambda)} = \frac{\lambda}{\lambda_0 F_{d0}} + \frac{1}{F_{r0}} - \frac{D(\lambda - \lambda_0)}{(n_0 - 1)F_{r0}}. \quad (5.12)$$

Now, if the ratio of the focal length of the diffractive surface to that of the refractive lens is set to

$$\frac{F_{d0}}{F_{r0}} = \frac{(n_0 - 1)}{\lambda_0 D} \quad (5.13)$$

the resulting focal length of the combined refractive/diffractive element is given by

$$\frac{1}{F(\lambda)} = \frac{1}{F_{r0}} + \frac{1}{F_{d0}}. \quad (5.14)$$

The result of Equation (5.14) is an element that has a focal length independent of wavelength. By satisfying the condition of Equation (5.13), a refractive/diffractive lens can be made that, in a linear approximation, has no chromatic dispersion. It is important to note that both the refractive and diffractive components of the combination element have focal powers of the same sign. This is unlike a conventional achromatic doublet lens made from two refractive lenses of different materials. In a conventional refractive achromat, the two lenses must have focal powers of opposite sign. This difference between conventional and refractive/diffractive achromats is due to the fact that the focal length of a diffractive lens is shorter for longer wavelengths, while all refractive lenses have focal lengths that are longer for longer wavelengths.

Consider an example of the utility of a refractive/diffractive lens in correcting for chromatic aberration. KrF1 lasers are fast becoming useful tools in microlithography and medicine. KrF1 lasers emit ultraviolet light in a 2-nm wavelength band centered at 248 nm. The only durable material that can be polished into refractive lenses at this wavelength is fused silica. However, fused silica is very wavelength dispersive at 248 nm. A conventional refractive achromatic doublet is difficult to fabricate due to the lack of materials other than fused silica.

A refractive/diffractive achromatic can be readily fabricated. The dispersion constant of fused silica at 248 nm is  $D = 6 \times 10^{-4} \text{ nm}^{-1}$ . Using this value of  $D$  in Equation (5.13) results in

$$\frac{F_{d0}}{F_{r0}} = 3.4. \quad (5.15)$$

Therefore, if a diffractive lens is etched into the surface of a fused silica lens such that Equation (5.15) is satisfied, the resulting combination will have minimum chromatic dispersion. Figure 5-2(a) shows the phase aberration in waves across the aperture of a fused silica lens. The lens has a 1-in-diam. aperture and a 9-in focal length. Approximately 3 waves of chromatic aberration are present at the edge of the aperture over a 2-nm bandwidth. The placement of a diffractive lens profile, that satisfies Equation (5.15), on a surface of the fused silica lens results in the chromatic phase error shown in Figure 5-2(b). The maximum chromatic phase error has been reduced from 3 waves of aberration to less than 0.02 wave. This is a 150-fold improvement in wavefront error!

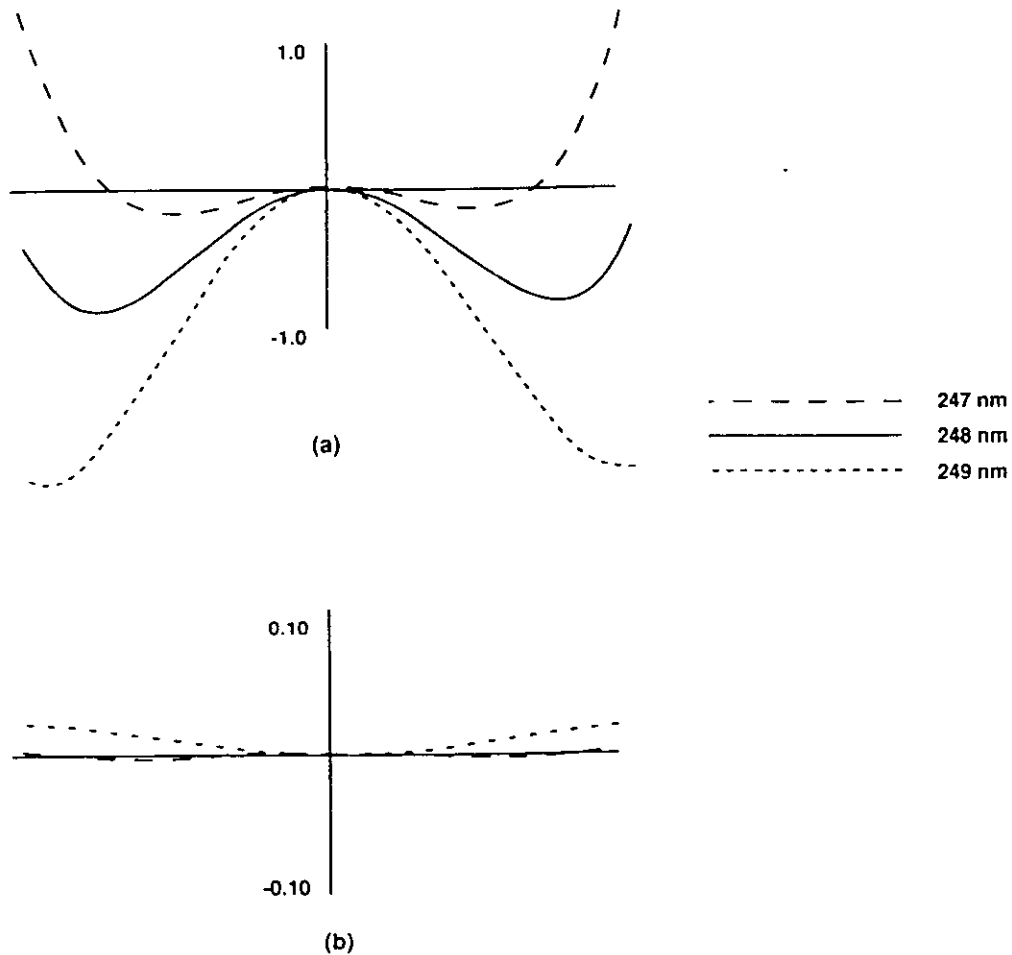
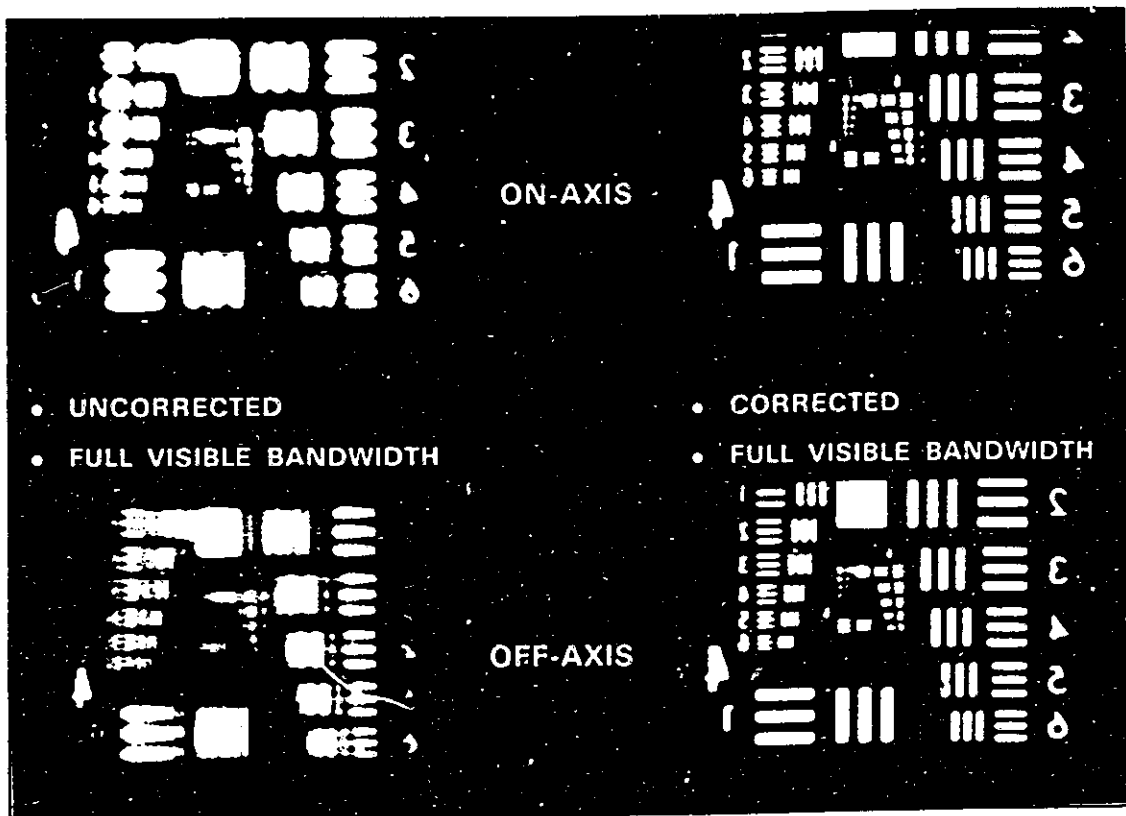


Figure 5-2. The phase aberration (a) of a refractive fused silica lens, and (b) of the same lens with diffractive aberration correction.

The concept of a refractive/diffractive achromat has been experimentally verified in the visible region of the spectrum. A 1-in-diam. fused silica (quartz) lens, with a 6-in focal length, was used to image an Air Force resolution target illuminated with a source emitting from 450 to 700 nm. An identical lens, with the properly designed diffractive profile etched into one of its surfaces, was also tested. The results are shown in Figure 5-3. The diffractively corrected lens is obviously far superior in performance than the refractive lens. The figure also shows that the refractive/diffractive combination is far superior for off-axis points. This can be understood by realizing that the amount of lateral chromatic aberration depends on the separation of the two lens components. In the case of a refractive/diffractive achromat, the two lens components are placed as close in proximity as possible, thus minimizing lateral chromatic effects.



125928-1

Figure 5-3. Experimental imaging results of the fused silica lens, with and without diffractive aberration correction.

Experimental verification of chromatic aberration correction using a refractive/diffractive element has been shown not only in the visible portion of the spectrum, but in the far-infrared (8 to 12  $\mu\text{m}$ ), the mid-infrared (3 to 5  $\mu\text{m}$ ), and the ultraviolet (0.246 to 0.248  $\mu\text{m}$ ) as well.

A refractive/diffractive achromat does not completely eliminate all of the chromatic aberration because the refractive index variation as a function of wavelength is not exactly described by Equation (5.8) which is a linear approximation to the true dispersive properties of refractive materials. In reality, the dispersion has a small nonlinear component that cannot be compensated for by a diffractive element. This nonlinear component is, in terms of lens design, called the secondary spectrum. Fortunately, the secondary spectrum is small in the majority of materials.

### 5.3 SPHERICAL ABERRATION CORRECTION

In the previous section we showed how a properly designed diffractive lens profile could be used to correct for the chromatic aberration of a refractive lens. Diffractive profiles can be used to correct for the monochromatic aberrations of refractive lenses as well. Here we will discuss the particular case of spherical aberration.

In the majority of cases, refractive lenses have spherical surfaces. A lens with spherical surfaces inherently suffers from spherical aberration. The spherical aberration of a single refractive lens element can be minimized by the proper choice of the radii of curvature of the two surfaces of the lens, but cannot be completely eliminated.

Two conventional solutions exist to eliminate spherical aberration. One is to use multiple lenses instead of a single lens. The number of lenses needed depends on the required performance of the lens system. This solution to the problem results in added weight, lower light throughput, and greater system complexity. The other conventional solution to the problem is to place an aspheric surface on the lens. This solution suffers from the fact that, in general, aspheric surfaces are very costly to produce.

The approach described here is to employ a diffractive surface to eliminate the spherical aberration of a refractive lens. For simplicity, the following analysis shows how a diffractive phase profile can eliminate third-order spherical aberration from a lens. A diffractive surface can correct for higher-order spherical aberration as well. Consider the wavefront, of wavelength  $\lambda_0$ , exiting from the back surface of a lens. Ideally, this wavefront would be a spherical wave converging to the focal point  $F$  and described by the phase profile

$$o_i = -\frac{2\pi}{\lambda_0} R \quad (5.16)$$

where  $R = (r^2 + F^2)^{\frac{1}{2}}$ . A second-order approximation to this ideal wavefront can be made by expanding  $R$  in a power series, resulting in

$$o_{i3} = -\frac{\pi r^2}{\lambda_0 F} + \frac{\pi r^4}{4\lambda_0 F^3} \quad (5.17)$$

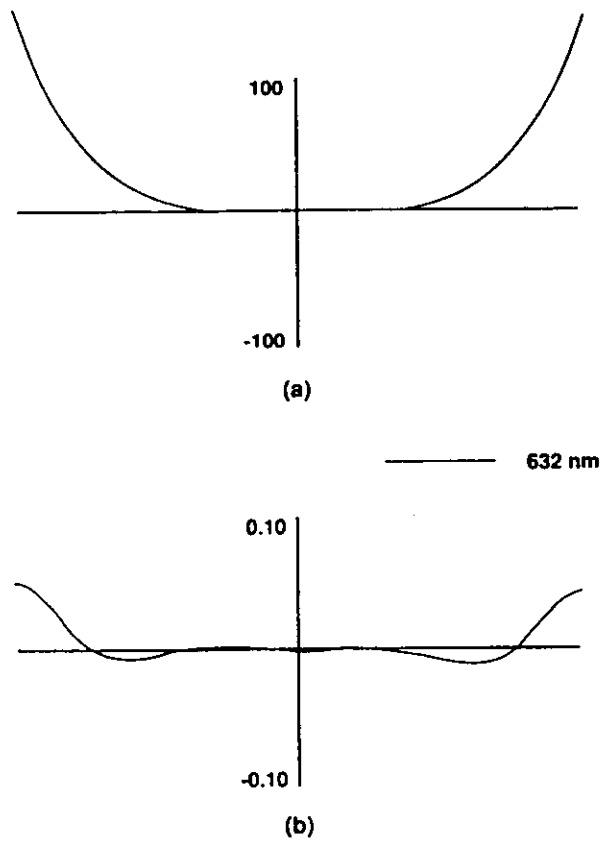


Figure 5-4. Theoretical phase error due to spherical aberration of a fused silica lens with and without diffractive correction.

124931-16



This equation is the expression, to fourth order in  $r$ , of an ideal wavefront. A refractive lens with spherical surfaces cannot produce this wavefront. The wavefront from any particular lens, depending on the design, will vary. To simplify things, let us assume that the wavefront exiting the refractive lens is quadratic and given by

$$O_r = -\frac{\pi r^2}{\lambda_0 F}. \quad (5.18)$$

The third-order spherical aberration of this lens would then be

$$O_a = -\frac{\pi r^4}{4\lambda_0 F^3}. \quad (5.19)$$

This third-order spherical aberration can be negated by simply etching a diffractive phase profile into the back surface of the lens that has a phase profile of

$$O_d = \frac{\pi r^4}{4\lambda_0 F^3}. \quad (5.20)$$

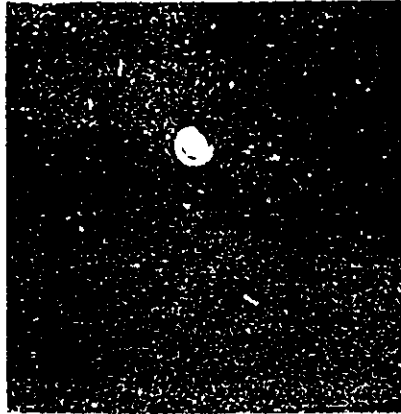
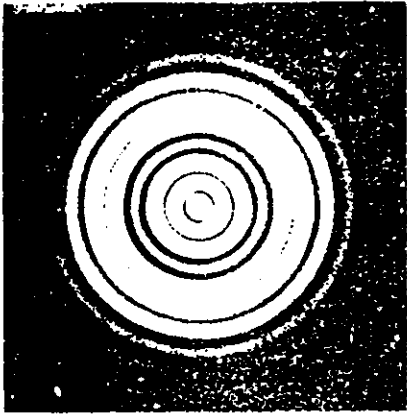
The resultant wavefront would be given by Equation (5.17) and suffer no third-order spherical aberration. This refractive/diffractive element, that has no spherical aberration at the wavelength  $\lambda_0$ , behaves very much like a conventional aspheric element.

A demonstration of the concept of spherical aberration correction has been performed using a fused silica single-element lens at the HeNe laser wavelength of  $0.6328 \mu\text{m}$ . The fused silica lens was plano-convex and had a 1-in aperture and a 2-in focal length. This lens suffered from severe spherical aberration, having a maximum phase error of close to 100 waves, as shown in Figure 5-4(a). When placed on the lens, the properly designed diffractive surface had a theoretical phase error of less than 0.1 wave [see Figure 5-4(b)].

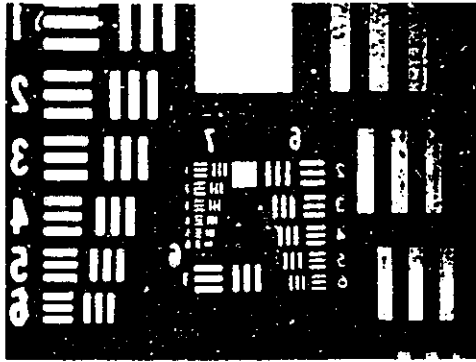
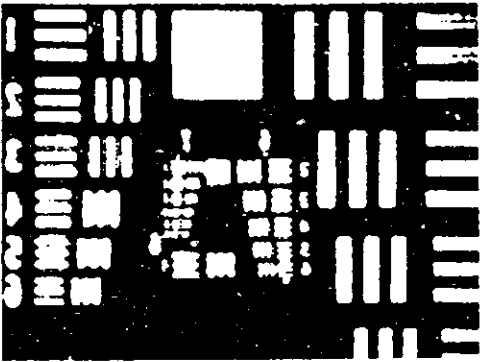
The diffractively corrected refractive lens was fabricated and tested. Figures 5-5(a) and (b) are images of the focal spot produced from the uncorrected and corrected lenses. Figure 5-5(a) clearly shows the expected light distribution associated with spherical aberration. The diffractively corrected focal point of Figure 5-5(b) is essentially diffraction limited. The experimentally verified improvement is enormous. It would take three conventional spherical lenses in tandem to achieve the same performance.

As further verification, the lenses were used to image an Air Force resolution test pattern. The results of the imaging experiments are shown in Figure 5-6. The uncorrected lens [Figure 5-6(a)] has a resolution as expected from theory. The diffractively corrected image [Figure 5-6(b)] has a resolution that is essentially diffraction limited.

The Binary Optics Group at Lincoln Laboratory has also demonstrated spherical aberration correction of lens elements at wavelengths in the far-infrared, the mid-infrared, and the ultraviolet regions of the spectrum.



125029 2



125029 3

## 5.4 LIMITATIONS OF REFRACTIVE/DIFFRACTIVE ELEMENTS

In Sections 5.2 and 5.3 we showed that it is possible to diffractively correct for the primary chromatic aberration and spherical aberration, at a specified wavelength, of a refractive lens. All single-element refractive lenses with spherical surfaces suffer from both chromatic and spherical aberration. The question arises as to how well a diffractive surface can correct for both chromatic and spherical aberrations over a finite wavelength band.

In order to get an estimate on the capability of a diffractive surface to correct for both chromatic and spherical aberration over a finite wavelength band, a model of the phase error of a refractive lens will be assumed to be

$$O_r(r) = \frac{2\pi}{\lambda} \left[ A(\lambda_0 - \lambda)r^2 + B(\lambda_0 - \lambda)^2 r^2 + Cr^4 \right]. \quad (5.21)$$

The first term on the right-hand side of this equation represents the primary chromatic aberration of the lens; the next term is a representation of the secondary spectrum; and the last term represents the spherical aberration of the lens. The values of the constants A, B, and C determine the amounts of primary chromatic aberration, secondary spectrum, and spherical aberration present.

A diffractive profile can be added to the refractive lens that imparts a phase given by

$$O_d(r) = \frac{2\pi}{\lambda} \left[ A\lambda r^2 - C\left(\frac{\lambda}{\lambda_0}\right)r^4 \right]. \quad (5.22)$$

The resulting wavefront, from the refractive-diffractive lens, will have a residual phase error given by

$$O_t(r) = \frac{2\pi}{\lambda} \left[ B(\lambda_0 - \lambda)^2 r^2 + C\left(1 - \frac{\lambda}{\lambda_0}\right)r^4 \right] \quad (5.23)$$

which is the phase error of Equation (5.21) minus the phase correction of Equation (5.22). The residual phase error of Equation (5.23) reveals, as expected, that the secondary spectrum of the refractive lens cannot be corrected. Furthermore, the additional term in Equation (5.23) represents the inability of a diffractive surface to completely correct for spherical aberration over a finite wavelength band. This residual term is commonly referred to as "spherochromatism," which is the amount of spherical aberration present in the image as a function of wavelength. For a center wavelength  $\lambda_0$ , the spherochromatism term in Equation (5.23) is zero, as expected. For wavelengths other than  $\lambda_0$ , the spherochromatism term is nonzero.

A diffractive surface is not able to completely correct for spherical aberration over a finite wavelength band. The amount of correction obtainable, as described in Equation (5.23), is proportional to the fractional bandwidth over which the lens has to operate. In many cases, the amount of correction is sufficient to justify the use of a diffractive surface.

As an example, consider an F/2 single-element silicon lens with a 100-mm focal length and an operating bandwidth from 3 to 5  $\mu\text{m}$ . This bandwidth represents a 50-percent fractional bandwidth

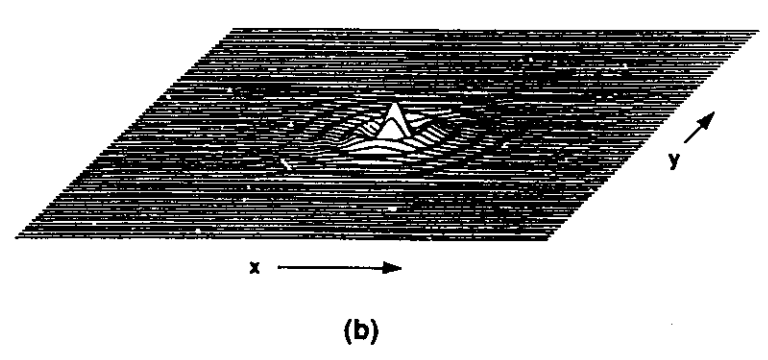
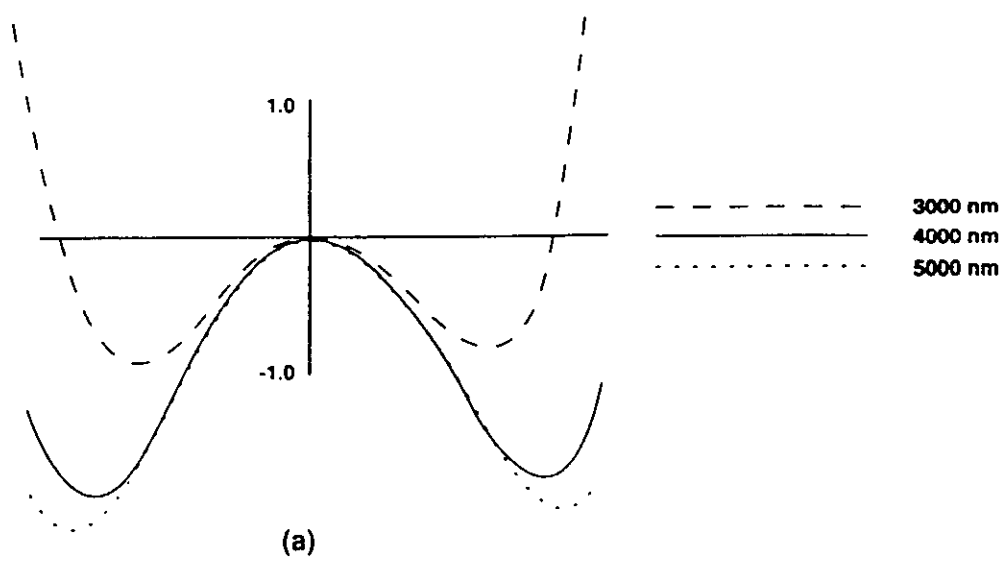


Figure 5-7. (a) The phase aberration and (b) point spread function of a refractive silicon lens.

124931-10

which is representative of, or larger than, the majority of finite bandwidth systems. The phase aberration of the best-designed, spherical surface, refractive element is shown in Figure 5-7(a), and the light distribution at the focal point (i.e., point spread function) is shown in Figure 5-7(b). It is evident from Figure 5-7(a) that the single-element lens has both chromatic and spherical aberration. A diffractive phase profile, described by Equation (5.22) and placed on the back surface of the refractive silicon lens, results in the phase aberration shown in Figure 5-8(a) and the point spread function shown in Figure 5-8(b). The primary chromatic aberration of the refractive lens has been eliminated, as has the spherical aberration at the center wavelength (4  $\mu\text{m}$ ). The residual spherochromatism, that cannot be corrected, has a maximum phase error of 0.2 wave. This is a significant improvement over the maximum phase error of the refractive lens (3 waves).

The silicon lenses described above, with and without the diffractive phase profile, were fabricated and tested. The experimentally measured modulation transfer function (MTF) of both lenses is plotted in Figure 5-9. The resolving capability of the diffractively corrected lens is far superior to that of the completely refractive lens. The discrepancy between the theoretical and experimental performance of the diffractively corrected lens is attributable to the fact that the theoretical prediction assumed a 100-percent efficient diffractive surface for all wavelengths. The experimentally tested lens was an 8-phase level structure with a maximum efficiency, at 4  $\mu\text{m}$ , of only 95 percent. In any case, the diffractively corrected lens far exceeded the completely refractive lens in performance.

The spherochromatism term in Equation (5.23) can be averaged over the operating fractional bandwidth  $\Delta\lambda/\lambda_0$ , resulting in an expression for the average residual spherochromatism

$$\bar{\sigma}_s(r) = C \left( \frac{\Delta\lambda}{2\lambda_0} \right) r^4. \quad (5.24)$$

This equation reveals that the ratio of the residual spherochromatism of a diffractively corrected lens to the spherical aberration of the refractive lens is equal to one-half the fractional bandwidth.

Figure 5-10 illustrates the chromatic and spherical aberration reduction capability of a diffractive profile. Examples are shown for common operating wavelength regions extending from the far-infrared to the ultraviolet. Notice that, in all the examples, the residual rms phase error is less than 0.1 wave.

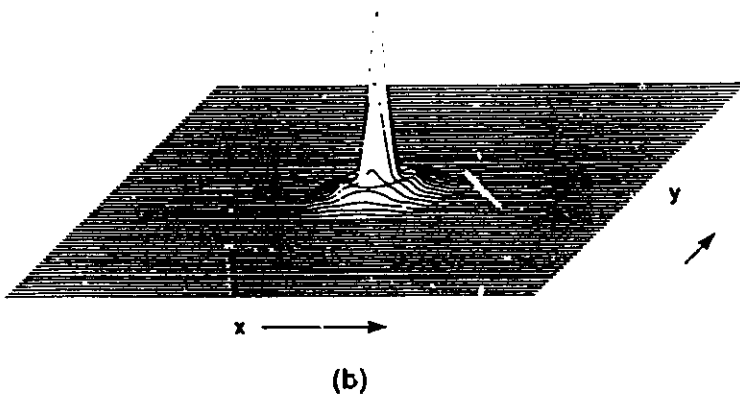
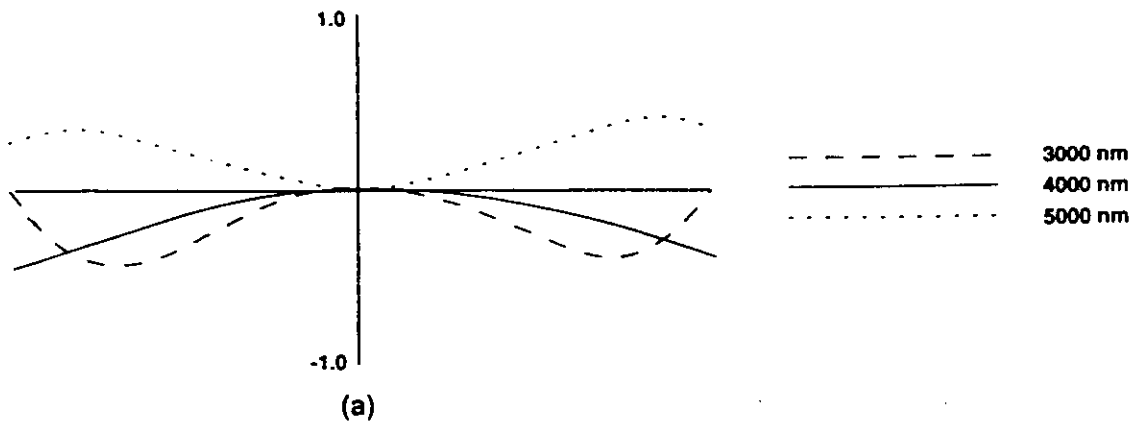
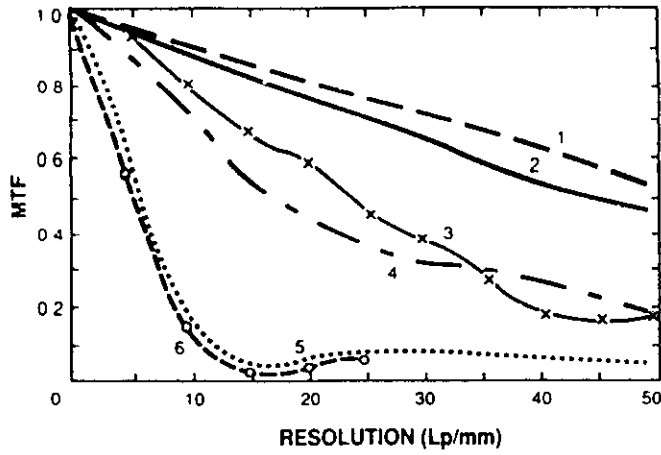


Figure 5-8. (a) The phase aberration and (b) point spread function of a diffractively corrected silicon lens.

124931-17

124931-19



- 1. DIFFRACTION LIMITED OPERATION
- 2. PREDICTED BINARY OPTICS SINGLET
- 3. MEASURED BINARY OPTICS SINGLET
- 4. PREDICTED SPHERICAL TRIPLET
- 5. PREDICTED SPHERICAL SINGLET
- 6. MEASURED SPHERICAL SINGLET

Figure 5-9. MTF curves for the silicon lens with and without diffractive correction.

125929-5

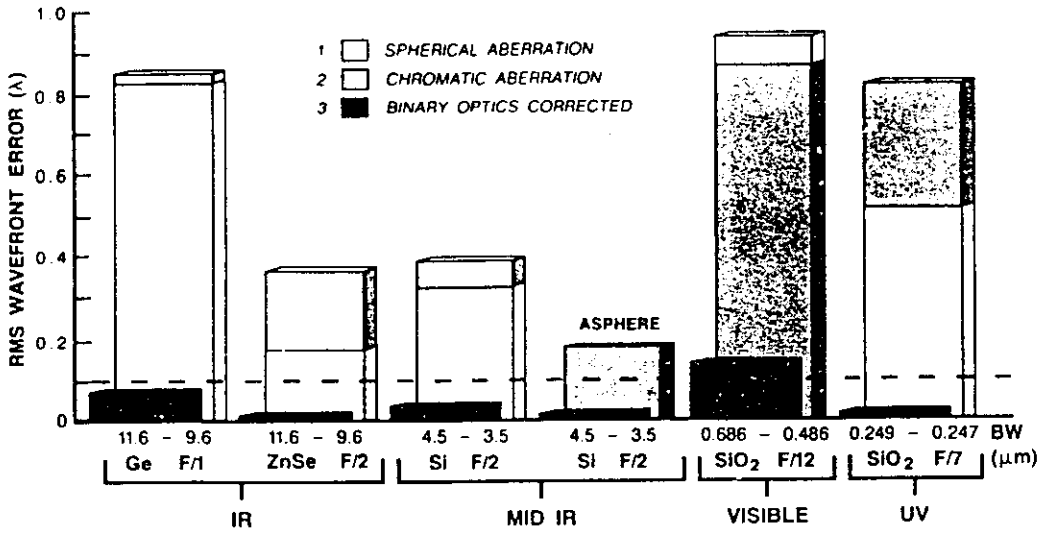


Figure 5-10. Examples of the chromatic and spherical aberration reduction possible by using a diffractive corrector.

## 6. DESIGNING DIFFRACTIVE PHASE PROFILES USING CODE V

In Section 5 we showed analytically the potential usefulness of a diffractive phase profile in reducing aberrations. The analysis was very nonspecific in regard to the exact diffractive phase function needed to optimally reduce the aberrations of a particular refractive lens. The exact determination of the optimum diffractive phase profile for any particular lens requires the assistance of a lens design program which must have the capability to insert diffractive phase profiles into a lens system and optimize the profile.

Inserting and optimizing a diffractive phase profile in a lens system can be accomplished using the commercially available lens design program CODE V: this program is used extensively by lens designers for optimizing and analyzing refractive and reflective systems. Lens designers familiar with the conventional capabilities of CODE V will have little problem learning and using the diffractive surface design capabilities of the program.

In CODE V, as well as other design programs, the lens designer inputs a design that meets the necessary first-order performance specifications of the system. An optimization routine is used that changes the initial first-order design in such a way to achieve maximum optical performance. In the optimization process, the thicknesses, spacings, and radii of curvature of the individual elements are treated as variables. The performance resulting from the optimization routine generally depends on the initial conditions specified by the designer.

CODE V has the ability to insert one or more diffractive surfaces anywhere into a lens system. These diffractive surfaces are specified by parameters that can be optimized to attain the best system performance. The implementation of diffractive surfaces in CODE V was formulated to emulate the recording of optically generated diffractive surfaces (i.e., holographic optical elements). The recording of a holographic surface is specified by the recording wavelength  $\lambda_0$  and the location in space of two point sources, as shown in Figure 6-1. The two point sources, located at  $R_1(x_1, y_1, z_1)$  and  $R_2(x_2, y_2, z_2)$ , produce spherical wavefronts. The interference of these two spherical wavefronts results in a diffractive phase profile at the recording plane given by

$$\phi_H(x, y) = \frac{2\pi}{\lambda_0} \left[ \sqrt{(x - x_1)^2 + (y - y_1)^2 + z_1^2} + \sqrt{(x - x_2)^2 + (y - y_2)^2 + z_2^2} \right]. \quad (6.1)$$

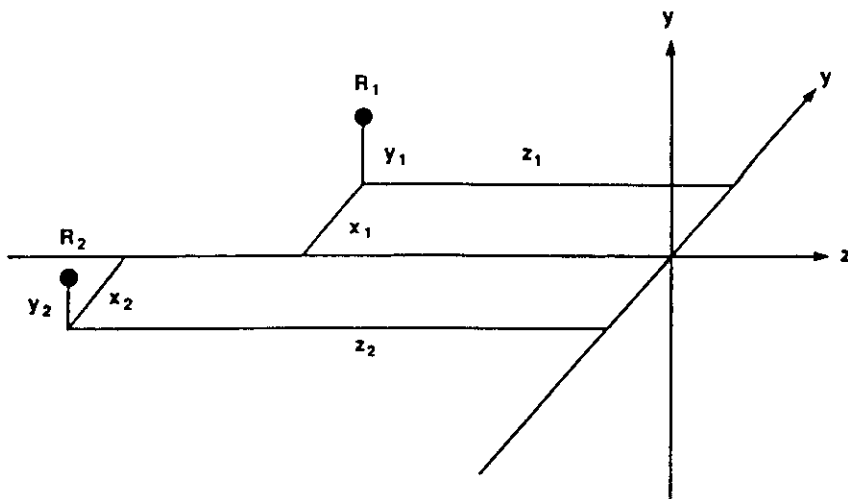
Note that the diffractive phase profiles that can be generated optically are a small subset of the total possible phase profiles. The spherical nature of the two interfering wavefronts restricts the set of optically generated phase profiles.

Fortunately, CODE V has the ability to analyze and optimize a more general set of diffractive phase profiles than that given by Equation (6.1). An additional diffractive phase term

$$\phi'(x, y) = \frac{2\pi}{\lambda_0} \sum_{k=0}^{10} \sum_{l=0}^{10-k} a_k b_l x^k y^l \quad (6.2)$$

can be added, in CODE V, to the optically generated phase profile of Equation (6.1). This additional diffractive phase term makes it possible to optimize and analyze a much larger subset of





124931-15

Figure 6-1. Recording setup for producing an optically generated holographic element.

diffractive phase profiles than those of Equation (6.1). Furthermore, it is possible (and preferable) to let Equation (6.2) completely specify the diffractive phase function. This can be accomplished by setting the two point source locations  $R_1$  and  $R_2$  at the same point in space. The resulting interference pattern from two point sources located at the same point is a constant. The resulting phase, given by Equation (6.1), becomes zero. The total diffractive phase is then given by Equation (6.2).

In the screen mode of CODE V, a diffractive surface can be placed in a lens system by entering the surface data screen (Gold S). Choosing the holographic surface option (Number 7) results in a screen that requires numerous input parameters. The first parameter to be entered is simply the surface number in the lens design on which the diffractive profile is to be placed. The second input parameter is the diffraction order of the diffractive phase profile that will be optimized and analyzed by CODE V. The first diffraction order is almost exclusively the order of interest.

The next parameter to be entered is the holographic recording wavelength. Typically, the wavelength of the laser used to optically record a holographic element would be entered. In our case, the value entered is irrelevant, since the diffractive phase profile is computer generated instead of optically generated. We prefer to set the wavelength equal to the center wavelength of the operating bandwidth only for consistency.

CODE V assumes that the diffraction order chosen will have 100-percent diffraction efficiency unless the next three input parameters are entered. These three parameters, used to model the diffraction efficiency of volume holograms, are: the volume thickness, the volume index of refraction,

and the index of refraction modulation. Since the surface relief profiles described in this report are not volume elements, it is best to leave these three parameters set to their default value of zero. The actual diffraction efficiency of an element will have to be determined outside of CODE V by using the theory developed in previous sections of this report.

The next set of eight input parameters specifies the two point source locations and whether the point sources are real or virtual. The spatial coordinates of both point sources are set to the same location, as mentioned above. It is then irrelevant whether the point sources are real or virtual. We set both point sources to be real for no particular reason other than consistency.

The last entry on the holographic surface screen is the number of aspheric diffractive phase terms, of Equation (6.2), to be entered. This entry is misleading in its wording. It is not the number of terms that should be entered, rather the number corresponding to the maximum term number in the polynomial expansion. The CODE V terminology for the aspheric phase polynomial is

$$O(x, y) = \frac{2\pi}{\lambda_0} \sum_k \sum_l a_{kl} x^k y^l \quad (6.3)$$

The number  $N$ , representing a particular term in the expansion, is determined by the expression

$$N = \frac{1}{2}[(k+l)^2 + 3l + k] \quad (6.4)$$

The polynomial expansion of Equation (6.3) is truncated to values of  $(k+l)$  less than or equal to 10. The total number of possible terms is 65. The term  $N = 65$  for example, as given by Equation (6.4), represents the coefficient  $a_{0,10}$  of the  $y^{10}$  term.

Once the maximum desired term number is entered on the screen, a final input screen consisting of a two-column table will appear. The term numbers  $N$  desired in the expansion are entered in the left-hand column; the values of the corresponding coefficients are entered in the right-hand column, directly opposite the appropriate term number. Any particular coefficient value entered can be set to a variable by pressing the Gold V key after entering the coefficient value. In many cases, there is little *a priori* knowledge as to what the coefficient values should be. For these cases, it is best to enter initial values of zero for all the desired coefficients and let the optimization routine determine their optimum values.

A complete description of entering a diffractive phase profile on a surface in a lens system has been given. The optimum diffractive phase profile is attained by using the CODE V automatic design feature. A deficiency of the CODE V program is that the diffractive aspheric phase terms are neglected in determining the first-order parameters of a lens system. These first-order parameters (i.e., effective focal length,  $F/\#$ , etc.) are often used as constraints in the automatic design routine. If a diffractive element is optimized in CODE V using first-order constraints, the result can be erroneous. Only exact ray trace parameters can be used as constraints when optimizing a diffractive phase profile in CODE V.

The vast majority of optical systems are designed to operate over a field of view that is radially symmetric. If the elements in a lens system are constrained to be radially symmetric, it is only necessary to optimize the performance over a radial slice of the field of view (i.e., y-axis). The lens system is then guaranteed to have the same performance over any radial slice of the field of view. The advantages to optimizing over a radial slice as compared with the full field of view are speed and cost. Each additional field point used in the automatic design routine increases the computation time and, therefore, the expense.

The diffractive phase profile, described in Equation (6.3), is not radially symmetric. If this diffractive profile is to be optimized for use in an optical system that is to operate over a radially symmetric field of view, the field points used in the optimization routine would have to cover the whole field of view. If only field points lying on the y-axis were used in the optimization, the resulting profile would perform well for y-axis field points. Field points lying on the x-axis, or any radial axis other than the y-axis, would not be guaranteed suitable performance.

Within CODE V, a way exists to constrain the diffractive phase profile of Equation (6.3) to be radially symmetric. Constraining the diffractive profile to be radially symmetric allows for the optimization over the complete field of view, using only the y-axis field points. A radially symmetric diffractive phase profile can be expressed in (x,y) coordinates as

$$o_r/(x,y) = \frac{2\pi}{\lambda_0} [a_1(x^2+y^2) + a_2(x^4+2x^2y^2+y^4) + a_3(x^6+3x^4y^2+3x^2y^4+y^6) + \dots]. \quad (6.5)$$

By entering only the (x,y) terms of this equation in the diffractive phase expression [Equation (6.3)] and constraining the coefficient values to conform to the proportions of Equation (6.5), the diffractive phase can be made radially symmetric.

When optimizing a diffractive profile, the coefficients of Equation (6.3) can be constrained to conform to the ratios of Equation (6.5) by introducing a sequence file in the automatic design routine. This sequence file acts as a user-defined constraint in the optimization process. The introduction of the proper sequence file in the automatic design routine allows for the optimization over the total field of view from only field points lying on the y-axis.

The generation of sequence files is explained in the CODE V manual. A sequence file is basically a file type .SEQ in the VMS directory. An example of a user-defined constraint sequence file that forces the diffractive phase profile to be radially symmetric is given below. For this example, the sequence file is given the name HOE2.SEQ:1. It constrains the coefficients of Equation (6.5), on surface Number 2 of the lens system, to be radially symmetric.

Filename: HOE2.SEQ:1

```
@H21:=(HCO S2 C3)-(HCO S2 C5)
@H21=0
@H22:=(HCO S2 C10)-(HCO S2 C14)
@H22=0
@H23:=(HCO S2 C12)-2*(HCO S2 C10)
```

```

@H23=0
@H24:=(HCO S2 C21)-(HCO S2 C27)
@H24=0
@H25:=(HCO S2 C23)-(HCO S2 C25)
@H25=0
@H26:=(HCO S2 C23)-3*(HCO S2 C21)
@H26=0

```

The file HOE2.SEQ:1 will be read into CODE V as a user-defined constraint by entering IN HOE2 while in the command mode version of CODE V's automatic design.

The first line of this sequence file defines a variable, H21, that is the difference between the  $N = 3$  and  $N = 5$  terms of Equation (6.3). The second line of the file constrains H21 to be zero. In other words, the first two lines constrain the coefficients of the  $x^2$  term and  $y^2$  term to be equal. The diffractive phase profile will therefore be radially symmetric in the  $r^2$  term. In a similar fashion, the next four lines constrain the profile to be radially symmetric in  $r^4$ , while the last six lines constrain the profile to be radially symmetric in  $r^6$ . This sequence file could easily be extended to constrain the profile to be radially symmetric up to the  $r^{10}$  term if desired.

The sequence file example given above constrains the diffractive profile on surface 2 of the lens system to be radially symmetric. Similar sequence files can be generated and stored in the user's directory that constrain the diffractive phase profile to be radially symmetric on any surface of the lens system. More elaborate sequence files can also be generated that constrain the diffractive profile in any way desired by the designer.

## 7. SUMMARY

In the past, optical designers have avoided considering diffractive elements as practical alternatives to refractive and reflective elements. The neglect had been justified based on the fact that no reliable and cost-effective fabrication capability existed.

Hopefully, this report has provided the reader some insight into the potential usefulness of multi-level diffractive phase profiles. These profiles can be easily designed and evaluated by using standard lens design programs along with the procedures detailed in this report. The fabrication of these elements has been shown to be reliable and straightforward. The fabrication tools and equipment necessary to produce these elements are not inexpensive. However, it is standard equipment used in the fabrication of integrated circuits and available for use at many places.

Multi-level diffractive elements are in no way the solution to all optical design problems. However, there are many systems where a diffractive element can be used to gain an advantage over a conventional design. The applications section of this report (Section 5) attempted to elucidate some of the distinct capabilities, as well as the limitations, of diffractive elements.

It is our hope that an optical designer, after reading this report, will begin to seriously consider diffractive surfaces as potential solutions to some of his/her lens design problems. The use of these surfaces is in its infancy. The larger the number of designers considering these structures, the faster diffractive elements will begin to appear in real optical systems.

## REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified		1b. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.	
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE			
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report 854		5. MONITORING ORGANIZATION REPORT NUMBER(S) ESD-TR-89-148	
6a. NAME OF PERFORMING ORGANIZATION Lincoln Laboratory, MIT	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Electronic Systems Division	
6c. ADDRESS (City, State, and Zip Code) P.O. Box 73 Lexington, MA 02173-9108		7b. ADDRESS (City, State, and Zip Code) Hanscom AFB, MA 01731	
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Advanced Research Projects Agency	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F19628-85-C-0002 (ARPA Order 6008)	
8c. ADDRESS (City, State, and Zip Code) 1400 Wilson Boulevard Arlington, VA 22209		10. SOURCE OF FUNDING NUMBERS	
		PROGRAM ELEMENT NO. 62702E	PROJECT NO. 305
		TASK NO.	WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) Binary Optics Technology: The Theory and Design of Multi-level Diffractive Optical Elements			
12. PERSONAL AUTHOR(S) Gary J. Swanson			
13a. TYPE OF REPORT Technical Report	13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Year, Month, Day) 1989, August, 14	15. PAGE COUNT 60
16. SUPPLEMENTARY NOTATION None			
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP	
		binary optics                      diffractive optical elements	
19. ABSTRACT (Continue on reverse if necessary and identify by block number)			
Multi-level diffractive phase profiles have the potential to significantly improve the performance of many conventional lens systems. The theory, design, and fabrication of these diffractive profiles are described in detail. Basic examples illustrate the potential usefulness, as well as the limitations, of these elements.			
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT <input type="checkbox"/> UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS		21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt. Col. Hugh L. Southall, USAF		22b. TELEPHONE (Include Area Code) (617) 981-2330	22c. OFFICE SYMBOL ESD/TML

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-90-C-0002 (ARPA Order Number 5328).

This report may be reproduced to satisfy needs of U.S. Government agencies.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

*Hugh L. Southall*

Hugh L. Southall, Lt. Col., USAF  
Chief, ESD Lincoln Laboratory Project Office

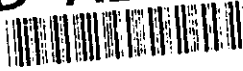
Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document  
when it is no longer needed.

2

AD-A235 404



Technical Report  
914

# Binary Optics Technology: Theoretical Limits on the Diffraction Efficiency of Multilevel Diffractive Optical Elements

G.J. Swanson

1 March 1991

**Lincoln Laboratory**

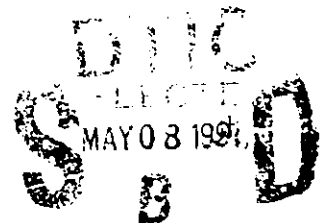
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS



Prepared for the Defense Advanced Research Projects Agency  
under Air Force Contract F19628-90-C-0002.

Approved for public release; distribution is unlimited.





MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
LINCOLN LABORATORY

**BINARY OPTICS TECHNOLOGY:  
THEORETICAL LIMITS ON THE DIFFRACTION EFFICIENCY  
OF MULTILEVEL DIFFRACTIVE OPTICAL ELEMENTS**

*G.J. SWANSON  
Group 52*

TECHNICAL REPORT 914

1 MARCH 1991

Approved for public release; distribution is unlimited.

LEXINGTON

MASSACHUSETTS

91 5 00 105



## TABLE OF CONTENTS

Abstract	iii
List of Illustrations	vii
1. INTRODUCTION	1
2. SCALAR THEORY OF DIFFRACTION EFFICIENCY	5
2.1 Diffraction Efficiency of a Multilevel Phase Grating	5
3. RIGOROUS ELECTROMAGNETIC THEORY OF DIFFRACTION EFFICIENCY	13
4. EXTENDED SCALAR THEORY OF DIFFRACTION EFFICIENCY	15
4.1 Optimum Grating Profile Depth	15
4.2 Extending Scalar Theory Prediction of Diffraction Efficiency	20
5. COMPARISON OF SCALAR, EXTENDED SCALAR, AND ELECTRO-MAGNETIC THEORIES	23
REFERENCES	27

## 1. INTRODUCTION

Diffractive optical elements are being considered as potential solutions to a number of optical design problems that are difficult or impossible to solve with conventional refractive and reflective elements. Two unique characteristics of diffractive elements can be exploited: the first is the dispersion property. Diffractive structures bend light rays of longer wavelengths more than those of shorter wavelengths, which is the reverse of refractive materials; therefore, diffractive structures minimize or eliminate the dispersive effects of refractive materials.

The second unique characteristic is the relative ease with which arbitrary phase profiles can be implemented. Advances in both diamond turning technology and the use of semiconductor fabrication equipment have made possible the construction of a variety of diffractive elements. Diamond turning technology allows fabricating diffractive surfaces over large areas in a relatively short period of time. However, there are limitations: the phase profile has to be circularly symmetric, and the accuracy with which a diffractive profile can be made is dependent on the tip size of the diamond turning tool.

Using semiconductor fabrication equipment to make diffractive elements has become a powerful technique. This particular approach produces a stepped approximation, referred to as a "multilevel structure," to the ideal profile. As the number of levels becomes large, the diffractive structure approaches the continuous profile. Diffractive elements can be made with feature sizes down to  $0.5 \mu\text{m}$ . The diffractive profiles can be very general with no symmetry restrictions, for example, lenslet arrays, which are being used to increase the collection efficiency of detector arrays and as components of wavefront sensing devices. These arrays are composed of individual diffractive lens profiles that are corrected for spherical aberration. Each lens has a rectangular aperture so that 100% of the area is covered. Such lenslet arrays would be difficult to fabricate any other way.

The diffractive optical elements that are fabricated by diamond turning or by using semiconductor fabrication equipment are surface relief elements. Surface relief diffractive elements are a particular class of diffractive elements that impart a phase delay to an incident wavefront in a very thin layer close to the surface of the element. The thickness of this layer is on the order of the incident wavelength. The phase delay is imparted to the incident wavefront by selectively removing material from the surface of the substrate.

Diffractive optical elements are different from reflective or refractive elements in that a light ray incident on a diffractive element is split into many rays, only one of which travels in the desired direction; its magnitude, relative to the sum of the magnitudes of all the split light rays, is called the diffraction efficiency. In most cases, a diffraction efficiency of one is desired, which is equivalent to all the light traveling in the chosen direction.

The diffraction efficiency that can be expected in practice from a particular diffractive element is limited by theory as well as by fabrication tolerances. The ability to fabricate diffractive elements has improved dramatically over the past few years — so much so that the attainable diffraction

## LIST OF ILLUSTRATIONS

Figure No.		Page
1	Surface relief profile of a one-dimensional, multilevel phase grating.	6
2	Light rays traced through two neighboring subperiods.	8
3	The first-order diffraction efficiency as a function of wavelength and the number of phase levels.	11
4	The first-order diffraction efficiency as a function of incident angle.	11
5	Geometrical ray trace through a surface relief grating.	16
6	Extended scalar theory prediction of optimum depth as a function of the period-to-wavelength ratio.	17
7	First-order diffraction efficiency as a function of the wavelength-to-period ratio for an $n = 1.5$ substrate.	19
8	First-order diffraction efficiency as a function of the wavelength-to-period ratio for an $n = 4$ substrate.	19
9	Light shadowing caused by finite depth surface relief profile.	21
10	Predicted first-order diffraction efficiency as a function of the wavelength-to-period ratio for a grating on a substrate with $n = 1.5$ .	24
11	Predicted first-order diffraction efficiency as a function of the wavelength-to-period ratio for a grating on a substrate with $n = 4$ .	24
12	Predicted first-order diffraction efficiency of a diffractive lens as a function of numerical aperture for a substrate with $n = 1.5$ .	25
13	Predicted first-order diffraction efficiency of a diffractive lens as a function of numerical aperture for a substrate with $n = 4$ .	26

efficiency for many elements (particularly those operating in the far infrared) is limited almost exclusively by theory. Performance degradation of diffractive optical elements due to fabrication errors has been investigated by others [1,2]. This report concentrates on the strictly theoretical limitations of achievable diffraction efficiency. It is, therefore, assumed that the surface relief profiles can be fabricated with infinite accuracy. The resulting diffraction efficiency calculations place a theoretical upper limit on attainable performance.

Whether a diffractive element will work for a particular application is ultimately determined by the obtainable diffraction efficiency; for example, consider the case of a lenslet array that is used to increase the light-gathering ability of a detector array. Certain detector arrays are made with a substantial fraction of dead space on the detector plane. A lens, properly placed in front of each detector, would effectively concentrate the light that would have fallen on the dead space onto the detector. For typical detector arrays under consideration, the increase in light-gathering capacity that a lenslet array can achieve is about a factor of 4, assuming that the lenslets have a diffraction efficiency of 100%. If the diffraction efficiency were only 50%, the increase in light-gathering efficiency would be only a factor of 2. If the diffraction efficiency dropped to 25%, the lenslet array would contribute absolutely nothing. Therefore, the diffraction efficiency that can reasonably be expected from a diffractive element is an important parameter.

Conventional lens design programs are now commonly used to model and optimize diffractive phase profiles. These lens design codes assume that the diffraction efficiency of a diffractive element is 100%. These codes are capable of determining phase profiles, but obtainable diffraction efficiency has to be determined separately. Theoretically, diffraction efficiency is a function of a number of parameters: the index of refraction of the substrate, the size of the zones of the diffractive profile relative to the incident wavelength, the polarization and angle of incidence of the incident light, and the depth and shape of the surface profile within a zone.

In theory, Maxwell's equations can determine exactly the diffraction efficiency of any diffractive structure. In practice, it is not possible to obtain exact solutions for the majority of cases. Numerical solutions are possible for certain diffractive structures; however, the necessary algorithms are very computationally intensive.

One of the simplest and most widely used ways to predict diffraction efficiencies is to use a scalar theory. The scalar theory of diffraction from a surface relief structure is based on a simplification of Maxwell's equations and a simplified model of the surface relief structure. The region of validity of the scalar theory is in the limit of the wavelength-to-zone spacing approaching zero. In other words, the size of the diffracting feature has to be very large compared with a wavelength of the incident light. The light is, therefore, deviated from the incident direction by a small angle. Section 2 describes the scalar theory and uses it to predict diffraction efficiency.

When the ratio of the wavelength-to-zone spacing approaches one, the incident light is deviated by large angles approaching 90 deg. It is in this regime that the scalar theory completely breaks down. Reliable estimates of diffraction efficiency can no longer be obtained from the scalar theory; however, numerical solutions to Maxwell's equations can be obtained for periodic diffracting

structures, i.e., gratings. If the grating period becomes much larger than a few wavelengths, the algorithm becomes too computationally intensive. Section 3 describes briefly the electromagnetic theory approach used to solve Maxwell's equations numerically for periodic structures.

In determining the diffraction efficiency of a grating, the scalar theory is valid for large period-to-wavelength ratios while the electromagnetic theory can only be used for very small period-to-wavelength ratios. A large void is left between the two limits where the scalar theory is not very accurate and the electromagnetic theory is numerically prohibitive. An approach to obtaining more reliable results for the diffraction efficiency in this region of period-to-wavelength ratios is to extend the scalar theory. This extended theory, developed in Section 4, combines aspects of geometrical optics with conventional scalar theory.

Section 5 compares the results of the three theories for a few representative examples, and the consequences of the theoretically obtainable diffraction efficiency for various applications are discussed.

## 2. SCALAR THEORY OF DIFFRACTION EFFICIENCY

The scalar theory of diffraction is based on the assumptions that light can be treated as a scalar rather than vector field and that the electric and magnetic field components are uncoupled. Two conditions are commonly stated as necessary for the scalar theory to have any validity: the size of the diffracting features must be large compared to the incident wavelength, and the diffracted field must be observed far from the diffracting structures [3].

A further approximation, referred to as the "Fresnel approximation," allows an integral solution of the propagation of the light field. The Fresnel approximation assumes that spherical waves can be approximated by quadratic waves. Within the realm of Fresnel diffraction, given the light field at some initial plane, the light field can be determined at any plane. Mathematically, the process of Fresnel diffraction is expressed by

$$U(x, y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy U(x_0, y_0) \exp \left\{ \frac{i\pi}{\lambda z} [(x - x_0)^2 + (y - y_0)^2] \right\}, \quad (1)$$

where the initial light field,  $U(x_0, y_0)$ , is propagated a distance  $z$ , resulting in the light field  $U(x, y)$ . Multiplicative factors preceding the integral are generally not important and are omitted.

If the propagation distance is large enough so that the quadratic phase term in the integral of Equation (1) can be ignored, the resulting expression, again neglecting the unimportant multiplicative factors, becomes

$$U(f_x, f_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy U(x_0, y_0) \exp \{-i2\pi[f_x x_0 + f_y y_0]\}, \quad (2)$$

where  $f_x = x/\lambda z$  and  $f_y = y/\lambda z$ . Equation (2) represents the approximation known as the Fraunhofer diffraction and is the foundation for calculating diffraction efficiencies of surface relief diffractive elements in the scalar regime. For a periodic structure, i.e., grating, the amplitudes of the various diffraction orders can be determined by a simple Fourier transformation of the grating transmittance function. This simplification will be used to calculate the theoretical performance of multilevel phase gratings. It should be noted that in the scalar theory, the diffraction efficiency of an arbitrary diffractive optical element can be directly related to the diffraction efficiency of a grating [4]. It is, therefore, only necessary to determine the diffraction efficiency of a grating structure.

### 2.1 Diffraction Efficiency of a Multilevel Phase Grating

The surface relief profile of a one-dimensional, multilevel phase grating is shown in Figure 1. In order to calculate the diffraction efficiency of this grating structure, the far-field of one grating period has to be determined. The transmittance function of one period can be described by the

summation of the transmittances of  $N$  subperiods of width  $T/N$ , where  $N$  is the number of phase levels within one period of dimension  $T$ .

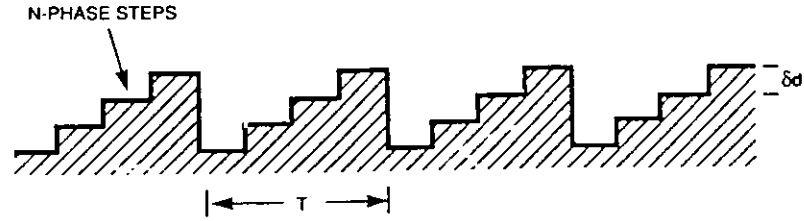


Figure 1. Surface relief profile of a one-dimensional, multilevel phase grating.

Each subperiod is a rect function of width  $T/N$ , centered at  $x = (m + 1/2)T/N$ , where  $m$  is an integer from 0 to  $N - 1$ . In the scalar approximation, the phase delay imparted by each subperiod can be expressed as  $\phi = m\phi_0/N$ , where  $\phi_0$  is the largest phase delay of all subperiods.

The far-field amplitude distribution of a subperiod, centered at a position  $x$ , with a width  $T/N$  and a phase delay of  $\phi$ , can be calculated from Equation (2) with the result

$$U'(f) = \frac{\sin(\pi T f / N)}{\pi T f / N} \exp\{-i2\pi x f\} \exp\{i2\pi \phi\}. \quad (3)$$

The far-field amplitude distribution of a total period can then be expressed as a summation of the far-field amplitude distributions of the  $N$  subperiods within the total period:

$$U(f) = \frac{1}{N} \sum_{m=0}^{N-1} \frac{\sin(\pi T f / N)}{\pi T f / N} \exp\{-i2\pi((m + \frac{1}{2})T/N)f\} \exp\{i2\pi m \phi_0 / N\}. \quad (4)$$

Repeating the period an infinite number of times constrains the far-field to have nonzero values only at positions  $f = l/T$ , where  $l$  is an integer that represents the  $l$ th diffraction order. The far-field amplitude of the  $l$ th diffraction order can be written as

$$A_l = \exp\{-i\pi l / N\} \frac{\sin(\pi l / N)}{\pi l / N} (1/N) \sum_{m=0}^{N-1} \exp\{-i2\pi(l - \phi_0)m / N\}. \quad (5)$$



The diffraction efficiency  $\eta_l$  of the  $l$ th order is  $A_l A_l^*$ ,

$$\eta_l = \frac{\sin^2(\pi l/N)}{(\pi l/N)^2} 1/N^2 \left[ \sum_{m=0}^{N-1} \exp \{-i2\pi(l - \phi_0)m/N\} \right]^2. \quad (6)$$

The summation in Equation (6) can be readily evaluated

$$\left[ \sum_{m=0}^{N-1} \exp \{-i2\pi(l - \phi_0)m/N\} \right]^2 = \frac{\sin^2(\pi(l - \phi_0))}{\sin^2(\pi(l - \phi_0)/N)}. \quad (7)$$

Substituting the result of Equation (7) into (6) gives the expression for the diffraction efficiency of the  $l$ th order as

$$\eta_l^N = \left[ \frac{\sin(\pi(l - \phi_0))}{\pi l} \frac{\sin(\pi l/N)}{\sin(\pi(l - \phi_0)/N)} \right]^2, \quad (8)$$

where  $N$  is the number of phase levels,  $\phi_0 = N\phi$ , and  $\phi$  is the phase depth change in waves of one subperiod.

Equation (8) is the basis for calculating diffraction efficiencies of surface relief diffractive optical elements. Within the scalar theory region of validity, this equation can determine the amount of light in any diffraction order for any number of phase levels. Equation (8) shows that for a given number of phase levels,  $N$ , the diffraction efficiency of the  $l$ th diffraction order is a function of one parameter,  $\phi_0$ . This  $\phi_0$  parameter can be related to the physical step height of a multilevel structure, as well as the incident wavelength and the angle of incidence of light impinging on the diffractive surface.

Figure 2 illustrates the relationship between the parameters necessary to define  $\phi_0$  in terms of physical properties. Two light rays are shown impinging on two neighboring subperiods in a multilevel structure. The index of refraction of the diffractive element is  $n$  and the angle of incidence is  $\theta_1$ . The physical step height between the neighboring subperiods is  $\delta d$ .

The parameter  $\phi$ , previously defined as the phase difference in waves between two neighboring subperiods, is therefore defined in terms of the parameters of Figure 2 as

$$\phi = \frac{1}{\lambda} (ny_1 - y_2), \quad (9)$$

where the distances  $y_1$  and  $y_2$  are geometrically determined to be

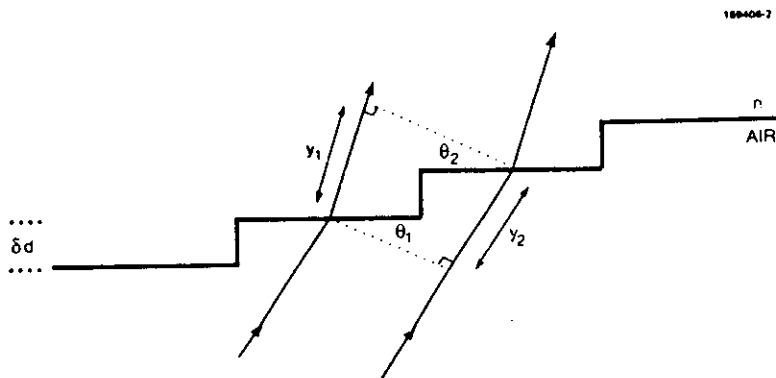


Figure 2. Light rays traced through two neighboring subperiods.

$$y_1 = \frac{\delta d}{\cos \theta_2} + \sin \theta_2 (x - \delta d \tan \theta_2) \quad (10)$$

and

$$y_2 = \frac{\delta d}{\cos \theta_1} + \sin \theta_1 (x - \delta d \tan \theta_1). \quad (11)$$

Inserting Equations (10) and (11) into (9) results in an expression for  $\phi$  that can be trigonometrically reduced to

$$\phi = \frac{\delta d}{\lambda} \{n \cos \theta_2 - \cos \theta_1\}. \quad (12)$$

Relating  $\theta_1$  to  $\theta_2$  in Equation (12) through Snell's law results in the following expression for  $\phi$  as a function of the step height, the index of refraction of the substrate, and the angle of incidence in air:

$$\phi = \frac{\delta d}{\lambda} [\sqrt{n^2 - \sin^2 \theta_1} - \cos \theta_1]. \quad (13)$$

The parameter  $\phi_0$  in Equation (8) is, again,  $\phi_0 = N\phi$ . It should also be noted that for the case of normal incidence  $\theta_1$  becomes zero, and Equation (13) reduces to the simple expression

$$o = \delta d(n - 1) / \lambda. \quad (14)$$

### 2.1.1 Examples

Equation (8) is a general scalar theory expression used to determine the diffraction efficiency of multilevel diffractive elements. An equivalent scalar theory expression for continuous profile diffractive elements, such as those fabricated by diamond turning techniques, can be found by taking the limit of Equation (8) as the number of levels  $N$  approaches infinity. The resulting expression for an infinite number of phase levels becomes

$$\left| \eta_1^\infty = \left[ \frac{\sin(\pi(l - \phi_0))}{\pi(l - \phi_0)} \right]^2. \quad (15)$$

Notice that the diffraction efficiency of the first order,  $\eta_1^\infty$ , becomes 100% when  $\phi_0 = 1$ . This is the result of the scalar theory that claims that 100% diffraction efficiency is possible.

The first diffraction order is usually of most interest and usually requires the highest diffraction efficiency. The diffraction efficiency of the first-order is maximum when  $\phi_0 = 1$ . The first-order diffraction efficiency of an optimized  $N$ -level element can be found by setting  $l$  and  $\phi_0$  both equal to one:

$$\eta_1^N = \left[ \frac{\sin(\pi/N)}{(\pi/N)} \right]^2, \quad (16)$$

expressing the maximum first-order diffraction efficiency one can expect from an  $N$ -level element in the scalar approximation.

The  $\phi_0$  parameter can be expressed as a function of the total depth  $d$  of the diffractive profile rather than the depth  $\delta d$  of a subperiod. The total depth  $d$  is simply related to  $\delta d$ , by  $d = (N - 1)\delta d$ . The  $\phi_0$  parameter, for normal illumination, becomes

$$\phi_0 = \left( \frac{N}{N - 1} \right) \frac{(n - 1)}{\lambda} d. \quad (17)$$

Setting  $\phi_0$  equal to one determines the optimum total depth for an  $N$ -level diffractive profile on a substrate of index  $n$ , to be used at a wavelength  $\lambda$ :

$$d = \frac{(N - 1)}{N} \frac{\lambda}{(n - 1)}. \quad (18)$$

In the limit of the number of levels approaching infinity, the well-known expression for the optimum depth,  $d = \lambda / (n - 1)$ , is obtained.

It is a fact that the diffraction efficiency of a diffractive structure is wavelength dependent. From the previous analysis, it can be deduced that the optimum step height for normal incidence and wavelength  $\lambda_0$  is

$$\delta d = \frac{\lambda_0}{N(n-1)}. \quad (19)$$

Substituting Equation (19) into (13) results in an expression for  $\phi$  from which  $\phi_0$  can be determined to be

$$\phi_0 = \frac{\lambda_0}{\lambda} \left| \frac{\sqrt{n^2 - \sin^2 \theta_1} - \cos \theta_1}{(n-1)} \right|. \quad (20)$$

Equation (8), in conjunction with (20), can be used to determine the diffraction efficiency of an  $N$ -level element as a function of wavelength and incident angle, for which the first-order diffraction efficiency has been maximized for wavelength  $\lambda_0$  and normal incidence.

Figure 3 plots the first-order diffraction efficiency as a function of wavelength for various values of  $N$ . The element was optimized, as described above, to have a maximum diffraction efficiency at wavelength  $\lambda_0$  and normal incidence.

Figure 4 plots the first-order diffraction efficiency at wavelength  $\lambda_0$  as a function of incident angle for various values of  $N$ . The element was optimized to have a maximum diffraction efficiency at wavelength  $\lambda_0$  and normal incidence. The figure reveals that in the scalar approximation the diffraction efficiency of these elements is very insensitive to the angle of incidence. This result reflects positively on the concept of placing diffractive surfaces on refractive optical elements, with the intent that the diffractive surface minimizes the aberrations of the refractive element. In such cases, the period-to-wavelength ratio of the diffractive structure is usually large, lending credibility to the scalar approximations; however, the range of incident angles impinging on the diffractive surface becomes quite large. Figure 4 shows that the diffraction efficiency, in general, will not suffer very much as a consequence of the large range of input angles.

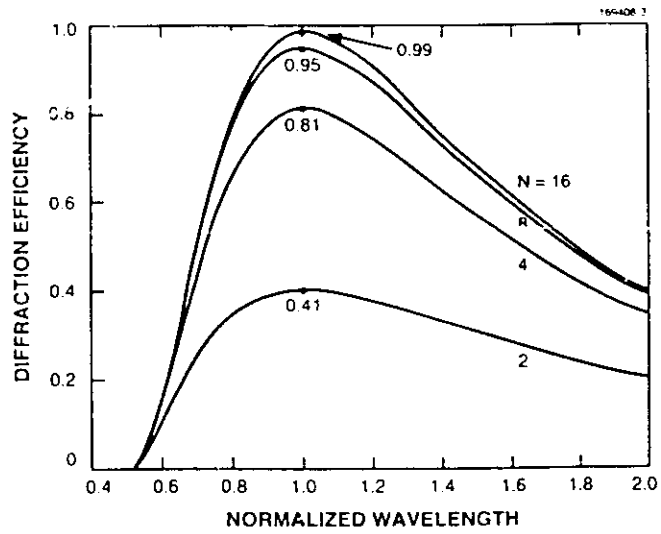


Figure 3. The first-order diffraction efficiency as a function of wavelength and the number of phase levels.

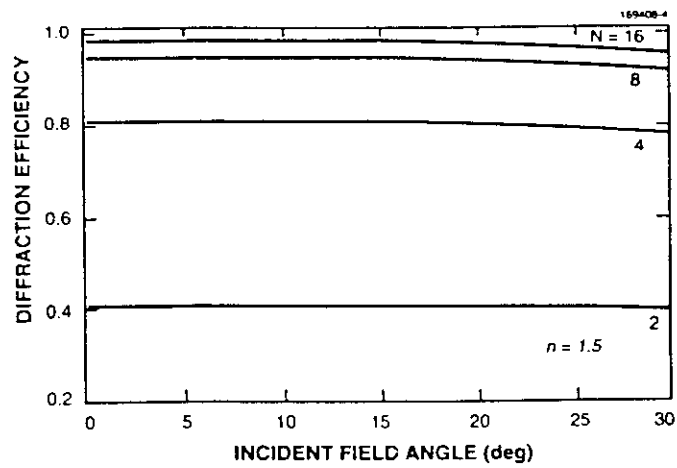


Figure 4. The first-order diffraction efficiency as a function of incident angle.

### 3. RIGOROUS ELECTROMAGNETIC THEORY OF DIFFRACTION EFFICIENCY

In Section 2, analytical expressions for the diffraction efficiency of surface relief phase gratings were developed using a scalar theory. As mentioned earlier, the diffraction efficiency of structures more complex than simple gratings can be directly related to the diffraction efficiency of the gratings through a nonlinear limiter analysis [4]. This allows a closed-form solution of the diffraction efficiency for any surface relief diffractive optical element.

The scalar theory is useful for designing surface relief diffractive elements with periods that are much larger than the wavelength for which the element is to be used. When the periods on the diffractive element become comparable in magnitude to the wavelength, the scalar theory (developed in Section 2) gives unreliable values for diffraction efficiency. The amount of discrepancy between the diffraction efficiency predictions of the scalar theory and reality is a function of the period-to-wavelength ratio and the index of refraction of the substrate.

In order to get a more reliable prediction of expected diffraction efficiencies, a more accurate theory must be used. In principle, Maxwell's equations could be solved for a particular diffractive structure, giving results that would be extremely accurate. In practice the solutions to Maxwell's equations have to be calculated numerically.

Various approaches to solving the electromagnetic equations of grating diffraction exist. Although they are equally valid, this report uses the approach first employed by Moharam and Gaylord [5], which is based on a coupled wave theory approach to solving Maxwell's equations. A brief outline follows. (Because the details are too numerous to discuss in this report, the reader is referred to Reference 5.)

An electromagnetic field incident on a phase grating can be divided into three main regions. The first, described by a homogeneous permittivity  $\epsilon_1$ , is where the incident and reflected fields propagate. The second is the modulation region of the grating profile, with permittivity alternating between  $\epsilon_1$  and  $\epsilon_3$ , the permittivity of the third region. This third region is where the transmitted field propagates and is characterized by the homogeneous permittivity  $\epsilon_3$ . In all three regions, permeability is equal to the permeability of free space.

The electromagnetic fields in the first and third regions can be expanded as sums of plane waves with the wave vectors determined from the Floquet condition. In the second region, the electromagnetic fields are expressed as Fourier expansions of the space harmonic fields. The second region is divided into  $N$  layers of equal thickness, each represented by the characteristics of the grating at the middle of the layer. The permittivity of each layer can be represented by a Fourier expansion. The permittivity in the second region,  $\epsilon_2$ , alternates within a layer between  $\epsilon_1$  and  $\epsilon_3$ .

The solution for the amplitudes of the reflected and transmitted diffraction orders is achieved by applying Maxwell's equations at the boundaries between the  $N$  layers. The electric and magnetic fields must have continuous tangential components.

An extensive computer code, DIFFRACT, has been developed based on the coupled wave theory. The accuracy of the code is dependent on the number of layers used to describe the grating modulation region and the number of orders retained in the Fourier expansion of the electromagnetic fields. The computation time necessary to solve for the diffraction efficiency increases linearly with the number of layers. In other words, the amount of computer time used to solve an  $N$  layer grating structure is twice that of an  $N/2$ .

The computation time necessary to solve for a grating is proportional to the cube of the number of orders retained in the Fourier expansion. In order to obtain an accurate solution, all the propagating orders, as well as a few evanescent orders, should be retained. The number of propagating orders from a grating is determined by the period-to-wavelength ratio; the larger the ratio, the more propagating diffraction orders. The computation time is, therefore, a strong function of the period-to-wavelength ratio. Furthermore, the maximum period-to-wavelength ratio grating that can reasonably be solved is dependent on the available computing power. In general, gratings with period-to-wavelength ratios greater than 10 become unreasonable to try to solve using this algorithm.

As seen above, one of the main constraints of the rigorous coupled wave theory, as well as other rigorous electromagnetic theories, is the limit on the maximum period-to-wavelength ratio grating that can be solved. The scalar theory, on the other hand, is only valid in the very large period-to-wavelength regime. A void remains between the usefulness of the two theories where unfortunately, a large percentage of the diffractive structures are being considered for various applications.

Another property of the rigorous electromagnetic theory is that it lends itself to very little intuitive insight into what to expect for diffraction efficiencies from gratings. Section 4 presents an intermediate theory for multilevel diffractive optical elements that attempts to bridge the gap between the scalar and the rigorous electromagnetic theories. This intermediate theory partially explains, in an intuitive fashion, the falloff of diffraction efficiency as a function of period-to-wavelength ratio.

## 4. EXTENDED SCALAR THEORY OF DIFFRACTION EFFICIENCY

The scalar theory of diffraction, as described in Section 3, is valid only for diffractive structures that have very large period-to-wavelength ratios. The rigorous electromagnetic theories of grating diffraction allow numerical solutions for only small period-to-wavelength ratios due to the computational complexity of the algorithms. A useful theory would function in the region of intermediate values of period-to-wavelength ratios, would be more accurate than the scalar theory, and would be computationally simpler than the rigorous electromagnetic theories.

The intermediate theory presented here, called the extended scalar theory, is like the scalar because it is strictly valid only in the confines of very large period-to-wavelength ratios, but for intermediate values of period to wavelength, agreement with reality is much better.

The major assumption that the extended scalar theory attempts to avoid is that the phase delay of the incident light, caused by the grating, occurs in an infinitely thin layer. The effects of the finite thickness of the grating profile are taken into consideration.

The finite thickness of the grating profile is treated by combining the scalar theory (based on wave propagation) with a geometrical theory (based on ray tracing). The incident light field is assumed to propagate through the thickness of the grating profile according to geometrical optics. Once the light exits the grating profile, the scalar theory based on wave propagation is applied.

### 4.1 Optimum Grating Profile Depth

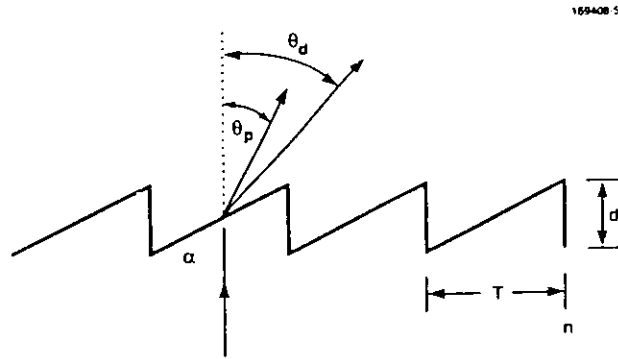
As mentioned above, the most widely used scalar theory assumes that the phase delay associated with a surface relief phase grating occurs in an infinitely thin layer on the surface of the substrate. This phase delay is physically implemented, however, by etching away certain areas of the substrate surface. The phase delay is the result of the optical path length difference due to the variation in surface profile thickness. The conversion of a phase delay into a physical thickness for a diffractive element designed to have a maximum first-order diffraction efficiency was shown in Section 2 to result in a physical depth of  $d$ , where  $d = \lambda/(n - 1)$ . Notice that the optimum depth based on the scalar theory is only a function of the wavelength and index of refraction of the substrate.

The mathematical assumption that the phase delay occurs in an infinitely thin layer is obviously unrealistic. Only for the case of substrates with extremely large refractive indices would the theory begin to agree with reality. Therefore, the scalar value of depth  $d$  is also an approximation. The questions "How bad is the assumption of the scalar theory?" and "What is the actual optimum depth?" need to be answered.

The approach used to determine the optimum depth by extending the scalar theory is shown in Figure 5 for the case of light normally incident on the substrate boundary and traveling from the substrate into air. The angle,  $\theta_d$ , at which the first diffraction order travels from the grating is simply determined by the grating equation



$$\sin \theta_d = \lambda/T. \quad (21)$$



- SNELLS LAW:  $n \sin(\alpha) = \sin(\theta_p + \alpha)$
- GRATING EQUATION:  $\sin \theta_d = \frac{\lambda}{T}$
- SET  $\theta_d = \theta_p$
- SOLVE FOR  $d$

Figure 5. Geometrical ray trace through a surface relief grating.

If one now considers each period of the grating to consist of a miniature refractive prism, light rays can be traced geometrically through each facet. The angle that the light rays exit the prism,  $\theta_p$ , is simply governed by Snell's law

$$n \sin \alpha = \sin(\theta_p + \alpha), \quad (22)$$

where  $\alpha = \arctan d/T$ .

An intuitive argument would suggest that the first diffraction order will have its maximum efficiency when the angle of the light rays traced through the prism  $\theta_p$  is equal to the angle of the first diffraction order  $\theta_d$ . The result of setting  $\theta_p$  equal to  $\theta_d$  and solving for  $d$  is

$$d = \frac{\lambda}{n - \sqrt{1 - (\lambda/T)^2}}. \quad (23)$$

Notice that this value of the grating depth is different from the scalar theory value. The most apparent difference is that the optimum depth given in Equation (23) is a function of the grating period, whereas the scalar theory value is independent of it. This immediately implies that for structures more complicated than gratings, the depth of the diffractive profile should vary as a function of the local period of the structure. Furthermore, it is worth noting that in the limit of the period  $T$  going to infinity, Equation (23) reduces to the scalar theory value.

From this point on, the depth value determined from Equation (23) is referred to as the "optimum depth" and represented by  $d_{opt}$ . The scalar depth value is represented by  $d_{app}$ . In order to see how the optimum varies from the scalar theory depth, it is useful to plot the ratio of the two as a function of the period-to-wavelength ratio, as shown in Figure 6 for two values of the index of refraction of the substrate. As expected, the ratio of  $d_{opt}/d_{app}$  asymptotically approaches a value of one as the period-to-wavelength ratio increases. The depth ratio deviates significantly from a value of one at small period-to-wavelength ratios. The exact period-to-wavelength ratio at which the deviation becomes significant is dependent on the index of refraction of the substrate. For high index of refraction substrates, the deviation occurs at smaller period-to-wavelength ratios than for low index of refraction substrates.

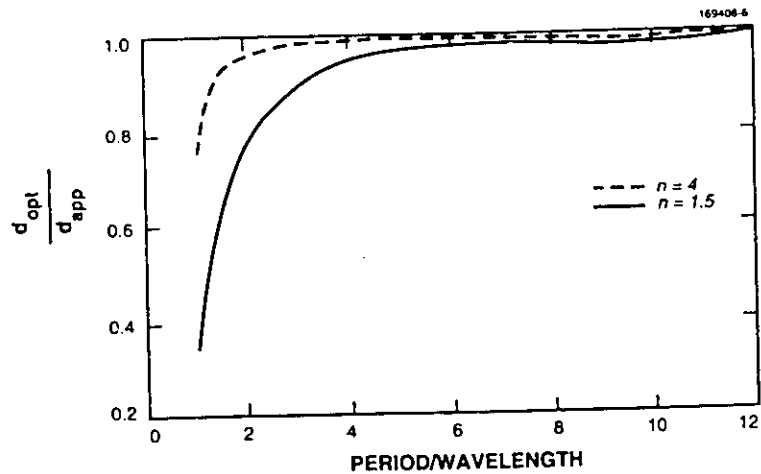


Figure 6. Extended scalar theory prediction of optimum depth as a function of the period-to-wavelength ratio.

Equation (23) was derived for normal incidence on the substrate boundary with the light traveling from the substrate into air. A more general expression for the optimum depth can be

derived using a similar approach to that used to derive Equation (23). Again, the idea is to simply equate the diffraction angle of the grating to the deviation angle of the prism for an arbitrary angle of incidence. The result of such an approach is the expression for the optimum depth as a function of incident angle as well as the wavelength-to-period ratio and the index of refraction:

$$d_{opt} = \frac{\lambda}{n\sqrt{1 - (\sin \theta_i)^2} - \sqrt{1 - (\frac{\lambda}{T} + n \sin \theta_i)^2}} \quad (24)$$

Notice that Equation (24) reduces to (23) when the incident angle  $\theta_i$  is set equal to zero. Equation (24) can also be used to determine the optimum depth for normal illumination when the light is traveling from air into the substrate. In Equation (24),  $\theta_i$  is defined as the incident angle in the substrate material. For the case of normal illumination from air into the substrate,  $\sin \theta_i$  has to be set equal to  $-\frac{\lambda}{nT}$ . The result is the optimum depth for normal incidence traveling from air into the substrate:

$$d_{opt} = \frac{\lambda}{n\sqrt{1 - (\lambda/nT)^2} - 1} \quad (25)$$

For all cases as the wavelength-to-period ratio approaches zero, the depth approaches the scalar theory value of  $d_{app} = \lambda/(n - 1)$ .

The depth values determined above were based on a somewhat intuitive argument. There is no proof that the expressions derived determine the depth that results in a maximum first-order diffraction efficiency. To test these extended scalar theory depth values, the DIFFRACT program (described in Section 3) was used to calculate the theoretical first-order diffraction efficiency for various wavelength-to-period ratio gratings. The minimum ratio tested was 0.5, corresponding to a 30-deg diffraction angle for the first order. Calculations were done for both high- ( $n = 4$ ) and low-index ( $n = 1.5$ ) substrates. The depth of the gratings was varied over a region that included the optimum as well as the scalar theory depth. In all cases, the first-order diffraction efficiency was maximized when the depth was near that predicted by the extended scalar theory.

Figures 7 and 8 plot the first-order diffraction efficiency as a function of the wavelength-to-period ratio. Curves are plotted for gratings having depth values equal to both the scalar theory and the optimum. Figure 7 plots a substrate with a low index of refraction ( $n = 1.5$ ), and Figure 8 plots a substrate with a high index ( $n = 4$ ). The calculations for the high-index substrate include a single layer antireflection coating; the low-index substrate had none. In all cases, the optimum depth value, as predicted using the extended scalar theory, results in a higher diffraction efficiency than that predicted using the scalar theory.

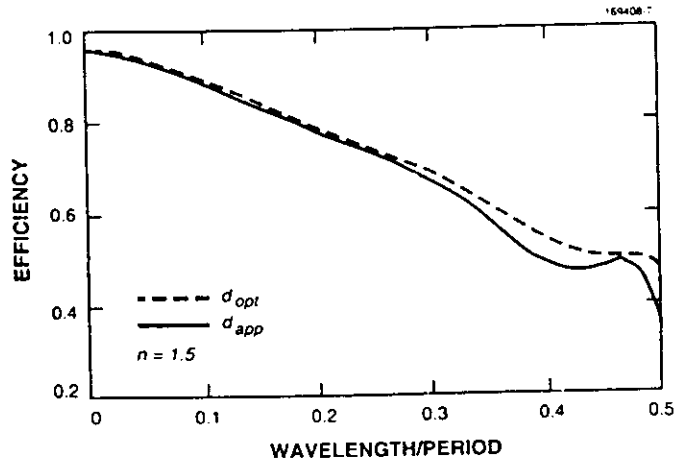


Figure 7. First-order diffraction efficiency as a function of the wavelength-to-period ratio for an  $n = 1.5$  substrate.

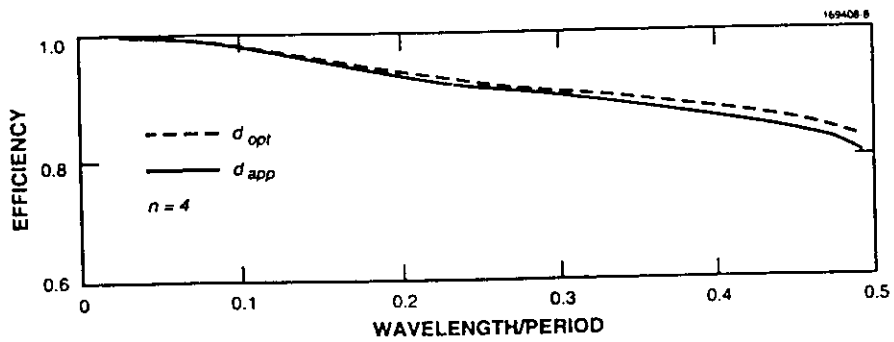


Figure 8. First-order diffraction efficiency as a function of the wavelength-to-period ratio for an  $n = 4$  substrate.

## 4.2 Extending Scalar Theory Prediction of Diffraction Efficiency

Diffraction efficiency predictions based on the scalar theory are completely independent of the wavelength-to-period ratio. Figures 7 and 8 clearly show, however, that the diffraction efficiency is a function of the wavelength-to-period ratio. One of the major reasons that the scalar theory fails to predict this falloff is, again, largely due to the assumption that the phase delay occurs in an infinitely thin boundary of the substrate.

The concept of geometrically tracing rays through the finite depth of the diffractive structure and subsequently applying the scalar theory can be used to extend the prediction of diffraction efficiency. This approach, though obviously not an exact solution to the diffraction problem, is more consistent with the electromagnetic theory calculations.

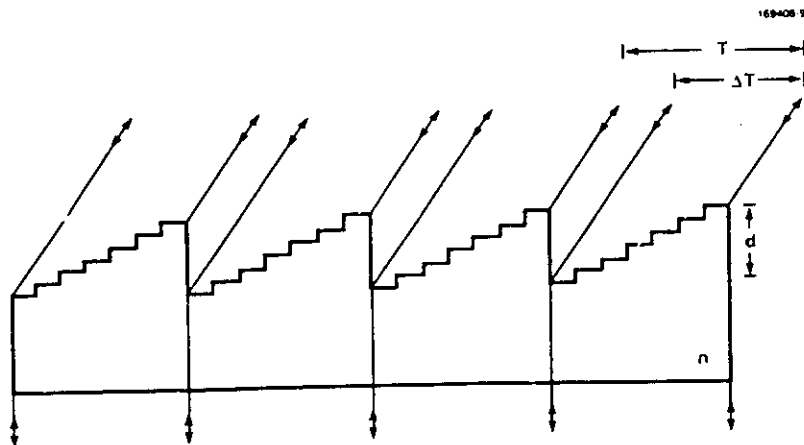
The most apparent feature that emerges from geometrically tracing rays through the depth of the diffractive structure is an effect referred to as "light shadowing." Figure 9 illustrates the geometrical ray trace and shows the light shadowing resulting from a finite thickness structure. Light rays traveling in a direction normal to the substrate boundary are refracted at the substrate/air interface. The angle that the light rays deviate is determined from Snell's law. The depth  $d$  is assumed to be the value determined in Section 4.1 that optimizes the first-order diffraction efficiency. The period of the grating is  $T$ , and the index of refraction of the substrate is  $n$ .

The light rays that exit the grating structure in the first diffracted order no longer fill the entire grating area. Immediately after the grating, the ratio of the area filled with light to the total area is called the duty cycle ( $DC$ ) and is equal to  $\Delta T/T$ . From a geometrical construction, the  $DC$  for the case illustrated in Figure 9 can be expressed as

$$DC = 1 - \frac{d\lambda}{T^2 \sqrt{1 - (\lambda/T)^2}} \quad (26)$$

Once the light rays are traced through the grating profile and the  $DC$  of the first diffraction order is determined, the scalar theory is applied to the exiting field. The light in the first diffraction order immediately after the grating resembles an unfilled aperture. It is a well-known result of the scalar theory that the amount of light that travels undiffracted through an unfilled aperture is equal to the  $DC$  of the unfilled aperture.

The light that is traced through one period of the grating encounters a stepped profile if the grating is made in a multilevel fabrication process. For this case, a fraction of the incident light equal to the  $DC$  given by Equation (26) is lost. Therefore, the fraction of light that resides in the first diffraction order can be approximately expressed by the product of the  $DC$  squared and the efficiency predicted from the scalar theory. Note from Equation (26) that the  $DC$  and, therefore, the first-order diffraction efficiency, is a function of the wavelength-to-period ratio; going to zero, the  $DC$  approaches one, and the first-order diffraction efficiency approaches the scalar theory value.



- $\frac{\Delta T}{T}$  = DUTY CYCLE = DC
- FIRST-ORDER EFFICIENCY =  $(DC)^2 \cdot \eta_{\text{SCALAR}}$
- $d$  ASSUMED TO BE OPTIMIZED

Figure 9. Light shadowing caused by finite depth surface relief profile.

A further extension could be approximated by including polarization effects. The scalar theory and its extension are polarization independent. These effects could be added to the extended scalar theory by including losses at the grating facet boundaries due to Fresnel reflection losses.

The extended theory is designed to be strictly valid only in the large period-to-wavelength ratio limit, as is the scalar theory, and more accurate for moderate wavelength-to-period ratios. As the period-to-wavelength ratio decreases, the extended scalar theory breaks down. The theory completely breaks down for a given index of refraction at the point where the slope of the individual facets within one period become large enough so that a light ray traced at the boundary will suffer from total internal reflection. Combining the equations for total internal reflection and the optimum grating depth results in an upper limit on the wavelength-to-period ratio for which extended scalar theory has any validity. This upper limit is expressed as

$$\left(\frac{\lambda}{T}\right)_{\text{max}} = \sqrt{1 - 1/n^2}; \quad (27)$$

---

for example, the extended scalar theory for a substrate with an index of refraction equal to 4 will totally break down when the wavelength-to-period ratio is equal to 0.97. For a substrate with a 1.5 index of refraction, the breakdown occurs at a wavelength-to-period ratio of 0.74. Section 5 compares the extended scalar theory with rigorous electromagnetic calculations. The maximum value of the wavelength-to-period ratio used in these comparisons is 0.5.

## 5. COMPARISON OF SCALAR, EXTENDED SCALAR, AND ELECTROMAGNETIC THEORIES

Three theories have been presented that can predict the diffraction efficiency from diffractive optical elements; each has strong points and weaknesses, and each complements the other in terms of information.

Obviously, the electromagnetic theory results in an exact solution to the problem of diffraction from a grating. Solutions to the electromagnetic theory can only be calculated numerically and computation time increases rapidly as the period-to-wavelength ratio increases; thus, there are two limitations. The first is the upper bound on the period-to-wavelength ratio for which a solution can be calculated, which is a function of the computer speed and how long one is willing to wait for the solution. The second limitation is the lack of any real insight into trying to optimize the diffraction efficiency of a diffractive structure.

The scalar theory is the least accurate yet easiest to use of the three; it allows for analytical expressions for the diffraction efficiency as a function of physical parameters. The analytical expressions give an insight into the design and/or feasibility of diffractive optical elements for a particular application. The diffraction efficiency calculated using the scalar theory is completely independent of the period-to-wavelength ratio. The value calculated can be used, however, as an upper bound on the obtainable diffraction efficiency. Scalar theory accuracy increases as the period-to-wavelength ratio increases. Thus, the theory becomes valid when the electromagnetic theory cannot be used due to computation time.

The extended scalar theory fills the void between the scalar and the electromagnetic. It retains the closed-form solution of the scalar theory and has a functional dependence on the period-to-wavelength ratio. Using the basic concepts of the extended scalar theory allows for a degree of insight into the optimum design of grating structures.

A graphical comparison of the results from the three theories is useful to visualize the differences in predicting diffraction efficiencies. Figures 10 and 11 plot the predicted first-order diffraction efficiencies as a function of the wavelength-to-period ratio for substrates with refractive indices of 1.5 and 4, respectively. The grating profiles are 16 phase level approximations to the optimum continuous profiles. The gratings on the  $n = 4$  substrate are assumed to have an optimum quarter-wave antireflection coating; the  $n = 1.5$  substrate is uncoated.

The most important feature of Figures 10 and 11 is the significant deviation between the scalar and the other two theories for moderate wavelength-to-period ratios. The curves confirm that the scalar theory is only valid for very small wavelength-to-period ratios. Another feature illustrated in the figures is the effect of the index of refraction of the substrate. Higher-index substrates suffer a smaller diffraction efficiency falloff than do low-index substrates. This effect is readily explained from the light shadowing concept presented in Section 4.



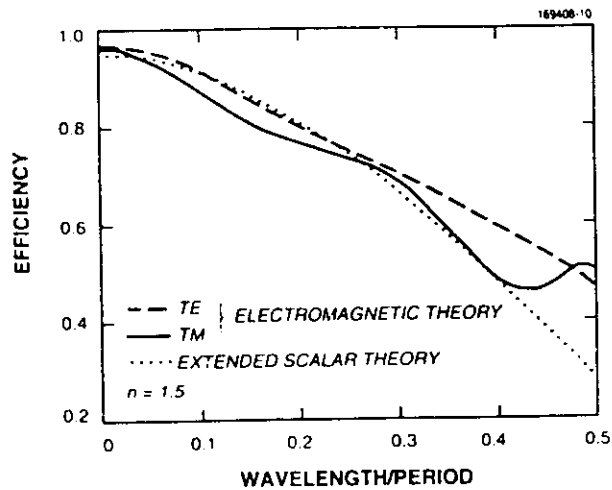


Figure 10. Predicted first-order diffraction efficiency as a function of the wavelength-to-period ratio for a grating on a substrate with  $n = 1.5$ .

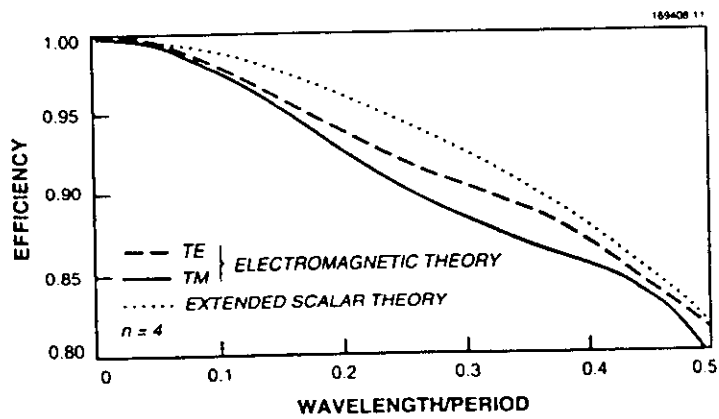


Figure 11. Predicted first-order diffraction efficiency as a function of the wavelength-to-period ratio for a grating on a substrate with  $n = 4$ .

It has been noted that the diffraction efficiency results of the extended scalar and the electromagnetic theories are dependent on the period-to-wavelength ratio; therefore, diffractive structures more complicated than simple periodic gratings have diffraction efficiencies that are a function of position on the element. Assigning a single diffraction efficiency value to an element requires sampling the aperture.

The diffraction efficiency of a diffractive lens, for example, can be approximately determined by assigning a periodicity to the lens that is a function of radial position. The lens can then be divided into annular regions of equal area. Each annular region is assigned a period equal to the period at its center. The extended scalar or the electromagnetic theory can then be used to determine the approximate diffraction efficiency of the annular regions. Since each region is of equal area, the lens can be assigned a diffraction efficiency that is simply the average of all the efficiencies of the annular regions. The accuracy of this approach is determined mainly by the number of annular regions into which the lens is segmented.

Using the approach described above, a first-order diffraction efficiency can be assigned to a diffractive lens as a function of its numerical aperture. Figures 12 and 13 plot the theoretical diffraction efficiencies as a function of numerical aperture for substrates with indices of refraction of 1.5 and 4, respectively. The substrate with an index of refraction of 4 is, as in the previous calculations, assumed to have an antireflection coating. The substrate with an index of refraction of 1.5 is uncoated. Curves are plotted from calculations of the electromagnetic and the extended scalar theories.

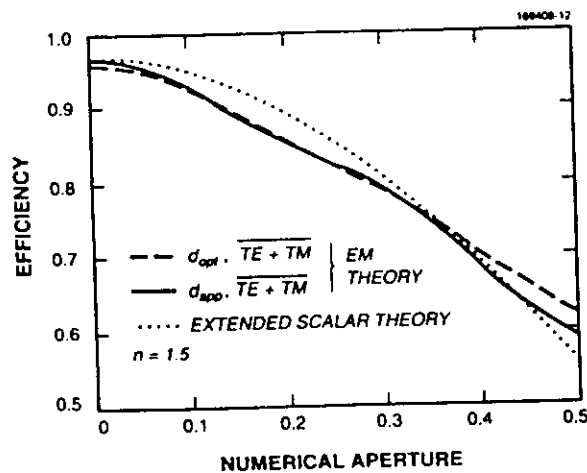


Figure 12. Predicted first-order diffraction efficiency of a diffractive lens as a function of numerical aperture for a substrate with  $n = 1.5$ .

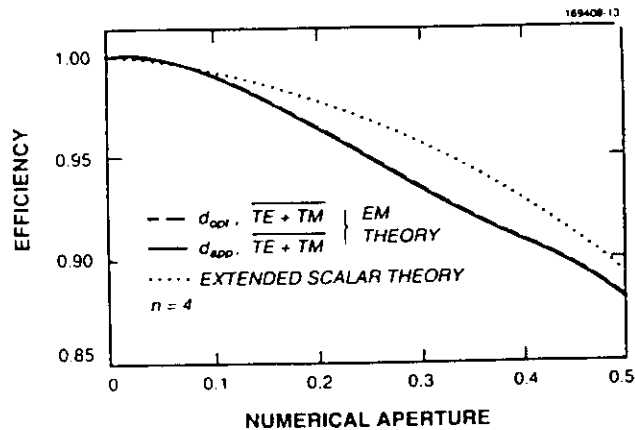


Figure 13. Predicted first-order diffraction efficiency of a diffractive lens as a function of numerical aperture for a substrate with  $n = 4$ .

The extended scalar theory calculations in Figures 12 and 13 were done for diffractive lenses with an optimum depth, while the electromagnetic theory calculations were done for diffractive lenses that had optimum depth profiles, as well as the approximate depth, determined from the scalar theory. Since optimum depth is a function of period, it varies for a lens as a function of radial position. Diffractive lenses with radially varying depths cannot realistically be fabricated using lithographic techniques; however, they can be produced using diamond turning methods.

Another difference is that the extended scalar theory is polarization independent, while the electromagnetic theory is dependent on the polarization of the incident light. On a radially symmetric diffractive lens, different angular positions are illuminated with different polarizations. The net effect over the entire aperture is simply an average of the diffraction efficiencies of the transverse electric (TE) and transverse magnetic (TM) polarization states.

The main point elucidated in Figures 12 and 13 is that the diffraction efficiency from a diffractive lens is theoretically limited. The difference in efficiency between that predicted from the scalar theory and that predicted from a more accurate theory is dependent on the numerical aperture of the lens, and the difference becomes quite large as the numerical aperture increases. Diffraction efficiency is also a function of the index of refraction of the substrate. Diffractive lenses of a given numerical aperture have a higher theoretical efficiency on high-index substrates than on low-index substrates.

## REFERENCES

1. M.W. Farn and J.W. Goodman, "Effect of VLSI fabrication errors on kinoform efficiency," *SPIE Proc.* Vol. 1211, 125-132 (1990).
2. J.A. Cox, T.R. Werner, J.C. Lee, S.A. Nelson, B.S. Fritz, and J.W. Bergstrom, "Diffraction efficiency of binary optical elements," *SPIE Proc.* Vol. 1211, 116-124 (1990).
3. J.W. Goodman, "Introduction to Fourier Optics." New York: McGraw-Hill (1968).
4. W.H. Lee, "Computer-generated holograms: Techniques and applications," in *Progress in Optics* 16, E. Wolf (ed.), 119-232 (1978).
5. M.G. Moharam and T.K. Gaylord. "Diffraction analysis of dielectric surface-relief grating." *J. Opt. Soc. Am.* 72, 1383-1392 (1982).

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.</small>				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 1 March 1991	3. REPORT TYPE AND DATES COVERED Technical Report		
4. TITLE AND SUBTITLE Binary Optics Technology: Theoretical Limits on the Diffraction Efficiency of Multilevel Diffractive Optical Elements		5. FUNDING NUMBERS  C — F19628-90-C-0002 PE — 62702E PR — 305		
6. AUTHOR(S)  Gary J. Swanson		7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)  Lincoln Laboratory, MIT P.O. Box 73 Lexington, MA 02173-9108		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)  Defense Advanced Research Projects Agency 1100 Wilson Boulevard Arlington, VA 22209		8. PERFORMING ORGANIZATION REPORT NUMBER  TR-914		
11. SUPPLEMENTARY NOTES  None		10. SPONSORING/MONITORING AGENCY REPORT NUMBER  ESD-TR-90-188		
12a. DISTRIBUTION/AVAILABILITY STATEMENT  Approved for public release; distribution is unlimited.		12b. DISTRIBUTION CODE		
13. ABSTRACT (Maximum 200 words)  Theoretical constraints limit the diffraction efficiency obtainable from multilevel diffractive optical elements. The scalar theory, commonly used to predict diffraction efficiencies, is overly optimistic. An extension to this theory is presented and compared with rigorous electromagnetic theory calculations. The extended scalar theory adds a degree of intuition in understanding why the diffraction efficiency of these elements is limited.				
14. SUBJECT TERMS binary optics diffractive optical elements		15. NUMBER OF PAGES 38		16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT SAR	

16  
pages / 40  
46?

6

Technical Report  
854

# Binary Optics Technology: The Theory and Design of Multi-level Diffractive Optical Elements

G.J. Swanson

14 August 1989

**Lincoln Laboratory**  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
*LEXINGTON, MASSACHUSETTS*



Prepared for the Defense Advanced Research Projects Agency  
under Air Force Contract F19628-85-C-0002.

Approved for public release; distribution is unlimited.

89 10 130 19

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F1962E-85-C-0002 (ARPA Order 6008).

This report may be reproduced to satisfy needs of U.S. Government agencies.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

*Hugh L. Southall*

Hugh L. Southall, Lt. Col., USAF  
Chief, ESD Lincoln Laboratory Project Office

Non-Lincoln Recipients  
**PLEASE DO NOT RETURN**

Permission is given to destroy this document  
when it is no longer needed.

# Algorithm for the rigorous coupled-wave analysis of grating diffraction

Nicolas Chateau\* and Jean-Paul Hugonin

*Institute d'Optique Théorique et Appliquée—Unité Associée au Centre Nationale de la Recherche Scientifique No. 14, Université de Paris-Sud, B.P. 147, 91403 Orsay Cedex, France*

Received December 29, 1992; revised manuscript received November 15, 1993; accepted November 16, 1993

Diffraction of light by periodic gratings is analyzed with a characteristic-matrix formalism based on a rigorous coupled-wave approach. This formalism is particularly convenient for modeling the diffraction by nonuniform periodic structures. In order to overcome numerical difficulties that are due to inhomogeneous eigenmodes, we propose a new algorithm that remains stable for gratings of any thickness. We obtain the stability by distinguishing in the computation the growing and the decaying inhomogeneous modes. Numerical examples and comparisons with previous results are given.

## 1. INTRODUCTION

Volume gratings have found applications<sup>1</sup> in various areas such as integrated optics, optical data processing and computing, holography, and spectroscopy. Their diffraction characteristics have stimulated many investigations<sup>2-20</sup> over more than two decades (Refs. 2 and 3 provide an excellent review of grating modeling). Since the analysis of Kogelnik,<sup>4</sup> the coupled-wave approach has been extensively studied. This theory has engendered wide interest because of its good physical insight and the simplicity of its mathematical resolution. The Kogelnik model<sup>4</sup> has the advantage of an analytic formulation, but its accuracy is limited by several approximations. Further research on the coupled-wave model resulted in more rigorous formulations,<sup>5-7</sup> in new solving methods and algorithms,<sup>5-8</sup> and in a generalization to numerous physical cases. The coupled-wave theory was extended to a variety of periodically modulated structures: planar transmission and transmission volume gratings<sup>4-11</sup> (possibly slanted and absorbing), surface relief gratings,<sup>12-15</sup> gratings with multiple coating layers,<sup>16</sup> nonuniform (or attenuated) gratings,<sup>17</sup> multiple superimposed gratings,<sup>18</sup> and anisotropic gratings.<sup>19</sup> The model was applied to structures of arbitrary profile and thickness, illuminated at any incidence angle and with any polarization. Yet some of the solution algorithms are unstable for relatively thick modulated layers, as noted in earlier papers.<sup>12,15,20</sup> Recently Pai and Awada<sup>20</sup> proposed a stable method for gratings of any thickness, for which solutions were found in the form of iterative one-way wave multiple reflection series; however, the calculation of the series coefficients seems time consuming, especially for gratings with narrow resonance.

In this paper we propose a rigorous and efficient method for calculating the coupled-wave diffraction of periodic gratings of arbitrary thickness without numerical problems. In Section 2 we derive a characteristic-matrix formalism of grating diffraction, well adapted to handle periodic structures with nonuniform modulation. In our model such a structure is represented by a stack of uniform subgratings of equal spatial period; the diffraction

matrix of the whole structure is simply obtained as the product of all the subgrating matrices. As in the analysis of Moharam and Gaylord,<sup>6</sup> the most straightforward solution method of the model involves two main steps:

- (1) Calculation of the eigenvalues and the eigenvectors of a constant coefficient matrix that characterizes the diffracted wave propagation and coupling (the eigenvectors represent the characteristic modes of the grating) and
- (2) Resolution of a linear system deduced from the boundary matching conditions. The system coefficients contain exponential functions of the product eigenvalue  $\times$  thickness.

With such a method some numerical difficulties are predictable if the grating thickness is relatively large: the linear system coefficients that correspond to the eigenvalues with a negative real part then become too large to be handled correctly by a computer. For the modeling of very deep modulated structures, it is useful to overcome these numerical problems through a convenient computer implementation. In Section 3 we propose an alternative and stable algorithm based on our characteristic-matrix formalism. The new algorithm takes the inhomogeneous eigenmodes into account in the resolution of the boundary field matching equations but avoids the calculation of very large exponentials by suitably reordering the eigenvalues in each characteristic matrix and recurrently defining a new sequence of well-behaved matrices. The algorithm is easy to compute, and its execution time is quite small compared with that of eigenvalue and eigenvector searching routines. In Section 4 we present numerical results for surface relief gratings and nonuniform slanted gratings.

## 2. CHARACTERISTIC-MATRIX FORMALISM OF WAVE DIFFRACTION INSIDE A GRATING

We consider a planar volume grating with a periodic index profile. For simplicity, we assume that the incident light



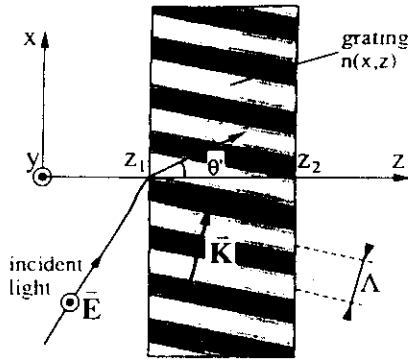


Fig. 1. Grating and incident wave geometry.

has transverse electric (TE) polarization and a known internal incidence angle, and we assume that the fringe planes are perpendicular to the plane of incidence. In addition, we assume that the grating has finite conductivity, and thus we neglect all the surface currents. At the moment we make no hypothesis about the media that surround the grating.

**A. Notation**

See Fig. 1 for the grating and incident wave geometry, with the following notation:

$$j = \sqrt{-1};$$

$\mathbb{Z}$ ,  $\mathbb{Z}^*$ , and  $\mathbb{R}$  represent the sets of integers, nonzero integers, and real numbers, respectively;

The  $z$  axis is normal to the grating surface;

The  $x$  axis is the intersection of the grating surface and the plane of incidence;

The  $y$  axis is perpendicular to the  $x$  and  $z$  axes;

$\Lambda$  is the grating fringe spacing;

$\mathbf{K}$  is the grating vector ( $\|\mathbf{K}\| = 2\pi/\Lambda$ ) perpendicular to the fringe plane and thus lies in the plane of incidence;

$\theta'$  is the internal incidence angle;

$n_0$  is the average refractive index;

$\lambda$  is the wavelength in free space;

$k_0 = 2\pi/\lambda = \omega/c$  is the corresponding wave number [time dependence  $\exp(-j\omega t)$ ];

$\mathbf{h} = \mu c \mathbf{H}$  defines the modified magnetic field, introduced to simplify the notation.

The periodic modulation is represented by the Fourier expansion

$$n^2(x, z) = \sum_{i=-\infty}^{+\infty} \bar{n}_i \exp[ji(K_x x + K_z z)]. \tag{2.1}$$

**B. Derivation of the Coupled-Wave Equations**

We analyze the propagation of waves inside the grating, using the tangential components  $E_y(x, z)$  and  $h_x(x, z)$ . These components of the electromagnetic field are continuous on the boundaries. We introduce the fundamental coupled-wave expansions:

$$\begin{aligned} E_y(x, z) &= \sum_{i=-\infty}^{+\infty} E_y^{(i)}(z) \exp(jk_x^{(i)} x), \\ h_x(x, z) &= \sum_{i=-\infty}^{+\infty} h_x^{(i)}(z) \exp(jk_x^{(i)} x). \end{aligned} \tag{2.2}$$

The  $x$  components of the wave vectors are obtained in the following way: phase matching with the incident wave yields

$$k_x^{(0)} = k_0 n_0 \sin(\theta'), \tag{2.3}$$

and  $k_x^{(i)} (i \in \mathbb{Z}^*)$  is given by the Floquet condition

$$k_x^{(i)} = k_x^{(0)} + iK_x, \quad i \in \mathbb{Z}. \tag{2.4}$$

Phase matching along boundaries implies that the components  $k_x^{(i)} (i \in \mathbb{Z})$  are continuous; thus the field subcomponents  $E_y^{(i)}(z)$  and  $h_x^{(i)}(z)$  are also continuous on the grating boundaries.

The Maxwell equation  $\nabla \mathbf{s} \wedge \mathbf{E} = j\omega\mu\mathbf{H}$  yields

$$h_x(x, z) = \frac{j}{k_0} \frac{\partial E_y(x, z)}{\partial z}. \tag{2.5}$$

Substitution of field expansions (2.1) and (2.2) into Eq. (2.5) and the projection of the resulting relation on the basis of functions of the variable  $x$  ( $x \rightarrow \exp[jk_x^{(i)} x]$ ) ( $i \in \mathbb{Z}$ ) give

$$\frac{dE_y^{(i)}(z)}{dz} = -jk_0 h_x^{(i)}(z), \quad i \in \mathbb{Z}. \tag{2.6}$$

In the equation of Helmholtz,

$$\nabla^2 E_y(x, z) + k_0^2 n^2(x, z) E_y(x, z) = 0; \tag{2.7}$$

after we represent the index modulation and the electric field by expansions (2.1) and (2.2), respectively, we introduce expression (2.6) to eliminate the derivatives of  $E_y(x, z)$  and project the resulting equation on  $x \rightarrow \exp[jk_x^{(i)} x]$  ( $i \in \mathbb{Z}$ ). We obtain

$$\begin{aligned} \frac{dh_x^{(i)}(z)}{dz} &= -j \left\{ \frac{[k_x^{(i)}]^2}{k_0} E_y^{(i)}(z) + k_0 \sum_{l \neq i} \bar{n}_{i-l} \right. \\ &\quad \left. \times \exp[j(j-l)K_z z] E_y^{(l)}(z) \right\}, \quad i \in \mathbb{Z}, \end{aligned} \tag{2.8}$$

where

$$[k_z^{(i)}]^2 = k_0^2 \bar{n}_0 - [k_x^{(i)}]^2, \quad i \in \mathbb{Z}. \tag{2.9}$$

In Eq. (2.8) we introduce the factors  $\exp(jiK_z z)$  ( $i \in \mathbb{Z}$ ) to obtain a differential system with constant coefficients where the unknowns are the functions of variable  $z$  [ $z \rightarrow E_y^{(i)}(z) \exp(-jiK_z z)$  and  $z \rightarrow h_x^{(i)}(z) \exp(-jiK_z z)$ ]:

$$\begin{aligned} &\frac{d[h_x^{(i)}(z) \exp(-jiK_z z)]}{dz} \\ &= -j \left[ iK_z h_x^{(i)}(z) \exp(-jiK_z z) + \frac{[k_z^{(i)}]^2}{k_0} E_y^{(i)}(z) \exp(-jiK_z z) \right. \\ &\quad \left. + k_0 \sum_{l \neq i} \bar{n}_{i-l} E_y^{(l)}(z) \exp(-jlK_z z) \right], \quad i \in \mathbb{Z}; \end{aligned} \tag{2.10}$$

the same operation performed in Eq. (2.6) gives

$$\begin{aligned} \frac{d[E_y^{(i)}(z) \exp(-jiK_z z)]}{dz} &= -j [iK_z E_y^{(i)}(z) \exp(-jiK_z z) \\ &\quad + k_0 h_x^{(i)}(z) \exp(-jiK_z z)], \quad i \in \mathbb{Z}. \end{aligned} \tag{2.11}$$

**C. Algebraic Resolution**

Equations (2.10) and (2.11) define an infinite system of first-order differential equations. For the numerical resolution we must retain a finite number  $N$  of diffracted orders. The method is rigorous at the limit  $N = +\infty$ ; practically,  $N$  can be chosen sufficiently large for obtaining a good precision.

The truncated resulting system may then be written in the matrix form

$$\frac{d\mathbf{U}(z)}{dz} = [\mathbf{M}]\mathbf{U}(z), \quad (2.12)$$

where the  $2N$  vector  $\mathbf{U}(z)$  is given by

$$\mathbf{U}(z) = \begin{bmatrix} \vdots \\ E_y^{(i-\nu)}(z)\exp[-j(i-\nu)K_z z] \\ \vdots \\ h_x^{(i-\nu)}(z)\exp[-j(i-\nu)K_z z] \\ \vdots \end{bmatrix}. \quad (2.13)$$

The first component of each  $N$  subvector corresponds to  $i = 0$ , the second one corresponds to  $i = 1$ , etc. . . .  $\nu$  is the number of negative orders retained; if we choose a set of diffracted orders centered on the zero order,  $\nu$  is equal to the integer part of  $N/2$ .

$[\mathbf{M}]$  is a  $2N \times 2N$  matrix with constant coefficients that may be expressed as

$$[\mathbf{M}] = -j \begin{bmatrix} K_z[\Delta] & | & k_0[\mathbf{I}_N] \\ \hline k_0[\Omega] & | & K_z[\Delta] \end{bmatrix}. \quad (2.14)$$

In Eq. (2.14) the  $N \times N$  submatrices are defined as follows:

$[\mathbf{I}_N]$  is the  $N \times N$  identity matrix;  
 $[\Delta]$  is a diagonal matrix with elements given by

$$\Delta_{i,i} = i - \nu, \quad i \in \{0, \dots, N - 1\}; \quad (2.15)$$

$$\begin{bmatrix} \vdots \\ E_y^{(i-\nu)}(z_1) \\ \vdots \\ h_x^{(i-\nu)}(z_1) \\ \vdots \end{bmatrix} = [\mathbf{P}(z_1)] \begin{bmatrix} \exp[-e_0(z_2 - z_1)] & & & 0 \\ & \exp[-e_1(z_2 - z_1)] & & \\ & & \ddots & \\ 0 & & & \exp[-e_{2N-1}(z_2 - z_1)] \end{bmatrix} [\mathbf{P}(z_2)]^{-1} \begin{bmatrix} \vdots \\ E_y^{(i-\nu)}(z_2) \\ \vdots \\ h_x^{(i-\nu)}(z_2) \\ \vdots \end{bmatrix}. \quad (2.22)$$

$[\Omega]$  is defined by

$$\Omega_{i,i} = \frac{k_z^{(i-\nu)^2}}{k_0^2}, \quad i \neq l: \quad \Omega_{i,l} = \tilde{n}_{i-l}, \quad (i, l) \in \{0, \dots, N - 1\}^2. \quad (2.16)$$

The solution of the shift-invariant system (2.12) between two arbitrary coordinates  $z_2$  and  $z_1$  ( $z_2 > z_1$ ) involves a matrix exponential function:

$$\mathbf{U}(z_1) = \exp\{-(z_2 - z_1)[\mathbf{M}]\}\mathbf{U}(z_2). \quad (2.17)$$

We shall express the matrix exponential in terms of eigenvectors and eigenvalues of  $[\mathbf{M}]$ . Diagonalizing matrix  $[\mathbf{M}]$ , we obtain

$$[\mathbf{M}] = [\tilde{\mathbf{P}}][\mathbf{D}][\tilde{\mathbf{P}}]^{-1}, \quad (2.18)$$

where the columns of matrix  $[\tilde{\mathbf{P}}]$  are the eigenvectors of  $[\mathbf{M}]$  and  $[\mathbf{D}]$  is the diagonal matrix of the eigenvalues of  $[\mathbf{M}]$ :

$$[\mathbf{D}] = \begin{bmatrix} e_0 & & 0 \\ & e_1 & \\ & & \ddots \\ 0 & & & e_{2N-1} \end{bmatrix}. \quad (2.19)$$

Using the definition of the matrix exponential and the associativity of the matrix product, we change relation (2.17) into

$$\mathbf{U}(z_1) = [\tilde{\mathbf{P}}]\exp\{-(z_2 - z_1)[\mathbf{D}]\}[\tilde{\mathbf{P}}]^{-1}\mathbf{U}(z_2). \quad (2.20)$$

**D. Characteristic Matrix**

From the coefficients of the eigenvector matrix  $[\tilde{\mathbf{P}}]$  we define a new  $2N \times 2N$  matrix  $[\tilde{\mathbf{P}}(z)]$ :

$$\begin{aligned} i \in \{0, \dots, N - 1\}: & \quad P_{i,l}(z) = \exp[j(i - \nu)K_z z]\tilde{P}_{i,l}, \\ i \in \{N, \dots, 2N - 1\}: & \quad P_{i,l}(z) = \exp[j(i - N - \nu)K_z z]\tilde{P}_{i,l} \\ & \quad l \in \{0, \dots, 2N - 1\}. \end{aligned} \quad (2.21)$$

Introducing definitions (2.19) and (2.21) at coordinates  $z_1$  and  $z_2$  into relation (2.20), we obtain

We define a  $2N \times 2N$  characteristic matrix of wave propagation inside the grating by the product  $[\mathbf{P}(z)]\exp\{-(z_2 - z_1)[\mathbf{D}]\}[\mathbf{P}(z_2)]^{-1}$ . The characteristic matrix relates the initial and final values of electromagnetic field subcomponents that are continuous on the boundaries.

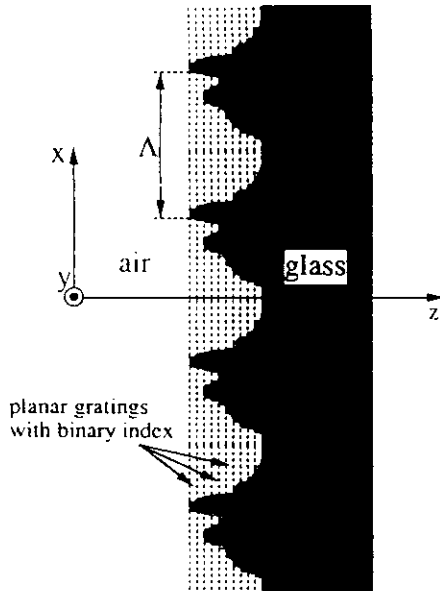


Fig. 2. Representation of a surface relief grating as a stack of planar volume elementary gratings with binary index.

**E. Generalization of the Formalism**

We have derived our analysis in the case of the plane diffraction of a TE polarized wave. The rigorous multi-wave coupled-wave theory was also used to describe the plane diffraction of a transverse magnetic (TM) incident field,<sup>10</sup> the more general conical diffraction of an arbitrarily polarized wave,<sup>11</sup> and the diffraction by anisotropic gratings.<sup>19</sup> These cases were shown to require the resolution of differential systems similar to Eq. (2.12), and thus they may be represented by a characteristic-matrix formulation such as Eq. (2.22).

The characteristic-matrix formalism is particularly convenient for the modeling of cascaded gratings with equal spatial period (i.e., same  $K_x$ ); since the field components in relation (2.22) are continuous on the boundaries, the matrix of the whole stack is obtained by multiplication of the elementary grating matrices. The case of a grating with nonuniform modulation versus depth is derived in the same manner; as was initially proposed by Kermisch,<sup>17</sup> such a grating can be represented by several uniformly modulated slices with the same fringe period.

A surface relief dielectric grating may be considered a particular stack of planar volume gratings,<sup>14,15</sup> where each slice has a binary periodic index. Such a grating is depicted in Fig. 2. The spatial frequency  $K_x$  is common to every slice, but the grating duty cycle may vary through the stack. Thus the representation of a surface relief grating also leads to the case of multiple cascaded gratings with equal period.

Our formalism also applies to the propagation in uniform layer coatings surrounding the grating. In this case, there is no coupling between the diffracted orders. If we adapt the Abeles<sup>21</sup> formalism of wave propagation in stratified media to the case of  $N$  propagating waves, it is straightforward to derive a multiwave characteristic matrix similar to Eq. (2.22) (in a uniform layer the eigenvalues are directly found in the form  $\pm j[k_0^2 n_0^2 - k_x^{(i)2}]^{1/2}$ ).

A generalized characteristic matrix of a spatially periodic structure including  $m$  substructures (surface grat-

ings, volume gratings, and uniform layer coatings, as illustrated in Fig. 3) is thus expressed in the form

$$\prod_{l=0}^{m-1} ([P_l(z_l)] \exp\{-(z_{l+1} - z_l)[D_l]\} [P_l(z_{l+1})]^{-1}), \quad (2.23)$$

where  $(z_k)_{k \in \{0, 1, \dots, m\}}$  are the coordinates of the interfaces ( $z_0 < z_1 < \dots < z_m$ ).

**F. Boundary Conditions**

We assume that both external media are homogeneous. When applying the boundary conditions, we need to distinguish between forward- and backward-propagating waves. We thus designate by  $f_F^{(i)}$  and  $b_F^{(i)}$ , respectively, the electric-field complex amplitudes of the incident and reflected waves in the first half-space, and we use the Rayleigh field expansions

$$E_y(x, z) = \sum_{i=-\infty}^{+\infty} f_F^{(i)} \exp\{j[k_x^{(i)}x + k_{Fz}^{(i)}z]\} + \sum_{i=-\infty}^{+\infty} b_F^{(i)} \exp\{j[k_x^{(i)}x - k_{Fz}^{(i)}z]\},$$

$$h_x(x, z) = -\frac{1}{k_0} \sum_{i=-\infty}^{+\infty} k_{Fz}^{(i)} f_F^{(i)} \exp\{j[k_x^{(i)}x + k_{Fz}^{(i)}z]\} + \frac{1}{k_0} \sum_{i=-\infty}^{+\infty} k_{Fz}^{(i)} b_F^{(i)} \exp\{j[k_x^{(i)}x - k_{Fz}^{(i)}z]\}, \quad (2.24)$$

where the  $x$  components of the wave vectors are still defined by relations (2.3) and (2.4) (since these components are constant on the boundaries) and the  $z$  components are given by  $k_{Fz}^{(i)} = [k_0^2 n_F^2 - k_x^{(i)2}]^{1/2}$ , with  $n_F$  being the complex refractive index of the first medium.

We see that the fields in external half-spaces are described either by components  $E_y^{(i)}$  and  $h_x^{(i)}$ , which are more convenient to describe wave propagation inside the modulated structure, or by parameters  $f_F^{(i)}$  and  $b_F^{(i)}$ , which are more intuitive and, as we shall see below, permit a simple writing of the boundary conditions. We derive a  $2N \times 2N$  matrix that acts as an interface between both representations at coordinate  $z_0$  (first boundary),

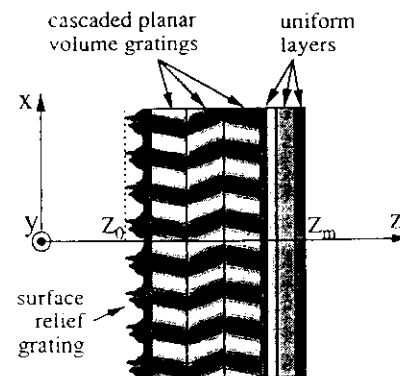


Fig. 3. Example of a compound periodic structure that can be represented by a characteristic-matrix formalism, including surface relief gratings, volume gratings, and uniform optical layers.

$$\begin{aligned}
 [C(z_0)] = & \begin{bmatrix} \ddots & & & 0 \\ & \exp[jk_{Fz}^{(i-\nu)} z_0] & & \\ & 0 & \ddots & \\ 0 & & & 0 \end{bmatrix} \\
 & \times \begin{bmatrix} \ddots & & & 0 \\ & \exp[-jk_{Fz}^{(i-\nu)} z_0] & & \\ & 0 & \ddots & \\ 0 & & & 0 \end{bmatrix} \\
 & \times \begin{bmatrix} \ddots & & & 0 \\ & \frac{k_{Fz}^{(i-\nu)}}{k_0} \exp[jk_{Fz}^{(i-\nu)} z_0] & & \\ & 0 & \ddots & \\ 0 & & & 0 \end{bmatrix}, \quad (2.25)
 \end{aligned}$$

and we obtain the relation

$$\begin{bmatrix} E_y^{(i-\nu)}(z_0) \\ \vdots \\ h_x^{(i-\nu)}(z_0) \end{bmatrix} = [C(z_0)] \begin{bmatrix} f_F^{(i-\nu)} \\ \vdots \\ b_F^{(i-\nu)} \end{bmatrix}, \quad (2.26)$$

Similar relations can be derived in the last half-space; we denote by  $f_L^{(i)}$  and  $b_L^{(i)}$  the complex amplitudes of the forward- and backward-propagating waves, respectively, of the Rayleigh field expansion in the last medium, and  $[C(z_m)]$  is the interface matrix at coordinate  $z_m$  (last boundary).

Introducing the Rayleigh coefficients and the interface matrices in the characteristic-matrix relation, we obtain the  $2N \times 2N$  matrix relation

$$\begin{bmatrix} \vdots \\ f_F^{(i-\nu)} \\ \vdots \\ b_F^{(i-\nu)} \\ \vdots \end{bmatrix} = [C(z_0)]^{-1} \prod_{l=0}^{m-1} ([P_l(z_l)] \exp\{-(z_{l+1} - z_l)[D]\}) \times [P_l(z_{l+1})]^{-1} [C(z_m)] \begin{bmatrix} \vdots \\ f_L^{(i-\nu)} \\ \vdots \\ b_L^{(i-\nu)} \\ \vdots \end{bmatrix}. \quad (2.27)$$

It is now straightforward to apply the usual boundary conditions:

(1) One incident wave is incident in the first half-space with amplitude equal to 1:

$$\begin{bmatrix} \vdots \\ f_F^{(i-\nu)} \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \quad (2.28)$$

(2) there are no backward-propagating waves in the last half-space:

$$\begin{bmatrix} \vdots \\ b_L^{(i-\nu)} \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \vdots \\ 0 \\ \vdots \\ \vdots \end{bmatrix}. \quad (2.29)$$

Substituting these vectors into Eq. (2.27), we obtain a linear system of  $2N$  equations with the  $2N$  unknowns  $b_F^{(i-\nu)}$  and  $f_L^{(i-\nu)}$ .

When the characteristic matrix is numerically well behaved (without huge exponential coefficients), system (2.27) is easily solved by classical inversion methods. Generally, several eigenvalues exhibit a negative real part, revealing the existence of growing inhomogeneous modes in the structure; if the grating layer is thick enough, the corresponding exponential terms in Eq. (2.22) become too large, yielding numerical instabilities or overflows. One solution is to neglect the most inhomogeneous modes of propagation by reducing the number of diffracted orders involved in the calculus; this solution is not satisfactory, except in some limiting cases, because it may induce a significant loss of accuracy. In Section 3 we propose a rigorous method for solving the problem without numerical difficulties, even with extremely large thickness values.

### G. Diffraction Efficiencies

After the determination of the reflected and transmitted amplitudes  $b_F^{(i-\nu)}$  and  $f_L^{(i-\nu)}$ , respectively, the reflection and transmission diffraction efficiencies, denoted  $\eta_F^{(i-\nu)}$  and  $\eta_B^{(i-\nu)}$ , respectively, are obtained by the formulas

$$\eta_B^{(i-\nu)} = \frac{k_{Fz}^{(i-\nu)}}{k_{Fz}^{(0)}} |b_F^{(i-\nu)}|^2, \quad \eta_F^{(i-\nu)} = \frac{k_{Lz}^{(i-\nu)}}{k_{Fz}^{(0)}} |f_L^{(i-\nu)}|^2. \quad (2.30)$$

### 3. ALGORITHM

In this section we use the following vector notation: incident waves  $\mathbf{I} = [1 \ 0 \ \dots \ 0]^T$ , reflected waves  $\mathbf{R} = [\dots \ b_F^{(i-\nu)} \ \dots]^T$ , transmitted waves  $\mathbf{T} = [\dots \ f_L^{(i-\nu)} \ \dots]^T$ , and backward-propagating waves in the last half-space  $\mathbf{O} = [\dots \ 0 \ \dots]^T$ . The  $2N$  unknowns of the problem are thus represented by  $\mathbf{R}$  and  $\mathbf{T}$ . We recall that the dimension of all these vectors is  $N$ . An eigenvalue has a critical negative real part if the number  $\exp(-\text{eigenvalue} \times \text{thickness})$  is too large to be correctly handled and a critical positive real part if  $\exp(+\text{eigenvalue} \times \text{thickness})$  is too large.

Our algorithm makes use of the following property: in a given characteristic matrix the number of eigenvalues with a critical negative real part is smaller than  $N$ , and the maximum number of eigenvalues with a critical positive real part is also restricted to  $N$ . A demonstration is given in Appendix A for the case of a non-absorbing grating. In our numerical investigations this property was always verified, even in absorbing gratings (see Subsection 4.B.2). By rearranging the position of the eigenvector matrix columns in relation (2.18), we can put the eigenvalues in growing order on the diagonal of matrix  $[D]$ . We now assume that such permutations are performed in each elementary characteristic matrix of relation (2.27).

Relation (2.27) may be rewritten in the form

$$\begin{bmatrix} \mathbf{I} \\ \mathbf{R} \end{bmatrix} = \prod_{k=0}^{m'-1} [\mathbf{A}_k] \begin{bmatrix} \mathbf{T} \\ \mathbf{O} \end{bmatrix}, \quad (3.1)$$

where  $m' = 3m + 2$  is the total number of matrices in Eq. (2.27); each matrix  $[\mathbf{A}_k]$  is either a well-behaved matrix, such as  $[\mathbf{P}_l(z_{l+1})]$ ,  $[\mathbf{P}_l(z_l)]^{-1}$  ( $l \in \{0, 1, \dots, m-1\}$ ),  $[\mathbf{C}(z_0)]$ , and  $[\mathbf{C}(z_m)]^{-1}$ , or a diagonal matrix of exponential terms. Among these diagonal elements only the first  $N$  ones may be critically positive and only the last  $N$  ones may be critically negative.

We divide each matrix  $[\mathbf{A}_k]$  into four submatrices:

$$[\mathbf{A}]_k = \begin{bmatrix} [\mathbf{A}_k^{00}] & | & [\mathbf{A}_k^{01}] \\ \hline \hline [\mathbf{A}_k^{10}] & | & [\mathbf{A}_k^{11}] \end{bmatrix}, \quad k \in \{0, 1, \dots, m'-1\}; \quad (3.2)$$

and we define two sets of vectors,  $\mathbf{X}_k$  and  $\mathbf{Y}_k$ , of dimension  $N$ :

$$k \in \{0, 1, \dots, m'-1\}: \begin{bmatrix} \mathbf{X}_k \\ \mathbf{Y}_k \end{bmatrix} = \left( \prod_{l=k}^{m'-1} [\mathbf{A}_l] \right) \begin{bmatrix} \mathbf{T} \\ \mathbf{O} \end{bmatrix},$$

$$\begin{bmatrix} \mathbf{X}_{m'} \\ \mathbf{Y}_{m'} \end{bmatrix} = \begin{bmatrix} \mathbf{T} \\ \mathbf{O} \end{bmatrix}. \quad (3.3)$$

From Eqs. (3.2) and (3.3) we immediately derive the recurrence relations

$$\begin{aligned} \mathbf{X}_{k-1} &= [\mathbf{A}_{k-1}^{00}]\mathbf{X}_k + [\mathbf{A}_{k-1}^{10}]\mathbf{Y}_k, \\ \mathbf{Y}_{k-1} &= [\mathbf{A}_{k-1}^{01}]\mathbf{X}_k + [\mathbf{A}_{k-1}^{11}]\mathbf{Y}_k, \end{aligned} \quad k \in \{1, 2, \dots, m'\}. \quad (3.4)$$

We now seek two families of  $N \times N$  matrices  $[\mathbf{P}_k]$  and  $[\mathbf{Q}_k]$  that verify

$$\begin{aligned} [\mathbf{P}_k]\mathbf{X}_k &= \mathbf{Y}_k, \\ [\mathbf{Q}_k]\mathbf{X}_k &= \mathbf{T}, \end{aligned} \quad k \in \{0, 1, \dots, m'\}. \quad (3.5)$$

We introduce  $[\mathbf{P}_k]$  and  $[\mathbf{Q}_k]$  into relations (3.4) and eliminate  $\mathbf{X}_k$  and  $\mathbf{Y}_k$ ; we obtain two sets of relations that are together sufficient for equalities (3.4) to be verified:

$$\begin{aligned} [\mathbf{P}_{k-1}] &= ([\mathbf{A}_{k-1}^{10}] + [\mathbf{A}_{k-1}^{11}][\mathbf{P}_k])([\mathbf{A}_{k-1}^{00}] + [\mathbf{A}_{k-1}^{01}][\mathbf{P}_k])^{-1}, \\ [\mathbf{Q}_{k-1}] &= [\mathbf{Q}_k]([\mathbf{A}_{k-1}^{00}] + [\mathbf{A}_{k-1}^{01}][\mathbf{P}_k])^{-1}, \end{aligned} \quad k \in \{1, 2, \dots, m'\}. \quad (3.6)$$

The preceding relations are a descending recurrence definition for  $[\mathbf{P}_k]$  and  $[\mathbf{Q}_k]$ . The initial terms of the recurrence are obtained from Eqs. (3.3) and (3.5):

$$\begin{aligned} [\mathbf{P}_{m'}] &= N \times N \text{ null matrix,} \\ [\mathbf{Q}_{m'}] &= N \times N \text{ identity matrix.} \end{aligned} \quad (3.7)$$

If  $[\mathbf{A}_{k-1}]$  is a well-behaved matrix, the calculation of  $[\mathbf{P}_{k-1}]$  and  $[\mathbf{Q}_{k-1}]$  by recurrence equations (3.6) yields no problem. If  $[\mathbf{A}_{k-1}]$  is a diagonal matrix of exponential terms, relations (3.6) simplify to

$$\begin{aligned} [\mathbf{P}_{k-1}] &= [\mathbf{A}_{k-1}^{11}][\mathbf{P}_k][\mathbf{A}_{k-1}^{00}]^{-1}, \\ [\mathbf{Q}_{k-1}] &= [\mathbf{Q}_k][\mathbf{A}_{k-1}^{00}]^{-1}, \end{aligned} \quad k \in \{1, 2, \dots, m'\}. \quad (3.8)$$

The calculation of matrix  $[\mathbf{A}_{k-1}^{11}]$  (the lower-right-hand part of  $[\mathbf{A}_{k-1}]$ ) is not problematic, since the large elements belong to the upper-left-hand submatrix  $[\mathbf{A}_{k-1}^{00}]$ .  $[\mathbf{A}_{k-1}^{00}]$  is a diagonal matrix, and its elements are exponentials of possibly large positive real numbers; thus matrix  $[\mathbf{A}_{k-1}^{00}]^{-1}$  is diagonal, and its elements are the exponentials of the opposites of the same numbers. Because of the convenient eigenvalue redistribution, the elements of both matrices  $[\mathbf{A}_{k-1}^{11}]$  and  $[\mathbf{A}_{k-1}^{00}]^{-1}$  remain of tractable magnitude.

For any value of  $k$ , recurrence relations (3.6) are calculated without numerical problems. Starting from the initial values at  $k = m'$ , their repetition leads to the determination of matrices  $[\mathbf{P}_0]$  and  $[\mathbf{Q}_0]$ . Then relations (3.3) and (3.5) applied to  $k = 0$  give

$$\mathbf{R} = [\mathbf{P}_0]\mathbf{I}, \quad \mathbf{T} = [\mathbf{Q}_0]\mathbf{I}. \quad (3.9)$$

Equation (3.1) was thus changed into relations (3.9). Equation (3.1) relates the input and output wave amplitudes in the first half-space to their values in the last half-space. Relations (3.9) express the output wave amplitudes (in both half-spaces) as functions of the input amplitude. These relations directly represent the physical transformation of light by the grating; consequently, matrices  $[\mathbf{P}_0]$  and  $[\mathbf{Q}_0]$  are well behaved.

To summarize, the algorithm proceeds as follows: first,  $[\mathbf{P}_k]$  and  $[\mathbf{Q}_k]$  are initialized at  $k = m'$  with relations (3.7). Then recurrence relations (3.6) are applied  $m'$  times downward through the stack. When matrix  $[\mathbf{A}_{k-1}]$  is well behaved, Eqs. (3.6) are simply calculated with the help of standard mathematical routines. When  $[\mathbf{A}_{k-1}]$  is a diagonal matrix of exponential terms, a special matrix inversion routine is required; we must calculate the coefficients of  $[\mathbf{A}_{k-1}^{00}]^{-1}$  directly by using the opposites of the critical eigenvalues without calculating  $[\mathbf{A}_{k-1}^{00}]$ . After  $[\mathbf{P}_0]$  and  $[\mathbf{Q}_0]$  are determined, the values of the reflected and transmitted amplitudes are directly obtained by relations (3.9), and the diffraction efficiencies in all the diffracted orders are given by formulas (2.30).

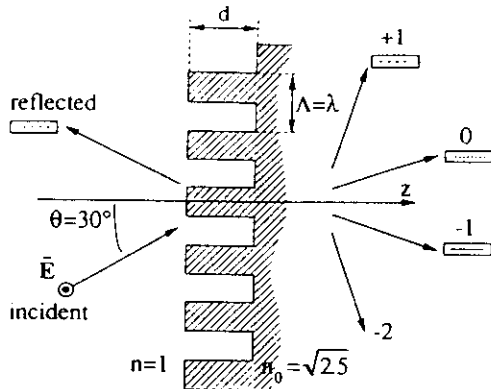


Fig. 4. Surface relief grating with a rectangular profile; the spatial period is equal to the incident wavelength, and the angle of incidence is  $\theta = 30^\circ$ .

#### 4. NUMERICAL RESULTS

##### A Surface Relief Gratings

Our algorithm is well adapted to the coupled-wave modeling of surface relief gratings. In the analysis we saw that such a grating might be simply considered a stack of cascaded planar volume gratings with a binary index. Because of the nonsinusoidal modulation of these elementary index gratings and because of the possibly large difference between the internal and external indices, several higher-order harmonics of the index profile are nonnegligible and induce a significant coupling involving several higher diffraction orders. Thus an accurate calculation of the diffraction efficiencies must account for these orders, including the evanescent ones. Beyond a critical groove depth the direct calculation and resolution of linear system (2.27) become unstable (we shall refer to this method as direct resolution), while our new algorithm still gives stable results with high accuracy.

For comparison, we treat two examples that were presented earlier by Moharam and Gaylord.<sup>15</sup> In both cases the first medium is air, the grating substrate extends to infinity on the positive  $z$  side and has a real index  $n_0 = (2.5)^{1/2}$ , and the wavelength in air is assumed to be equal to the groove spacing ( $\lambda = \Lambda$ );  $d$  represents the groove depth. A plane wave with TE polarization is incident on the grating at the first Bragg angle ( $\theta_B = 30^\circ$ ).

##### 1. Rectangular Grating (Square Wave)

In the first example the grating has a rectangular groove shape as shown in Fig. 4; thus we treat it as a single planar binary index grating with thickness  $d$ . We plot in Fig. 5 the diffracted intensity in order  $-1$  as a function of the groove depth expressed in wavelength units. We compare the results of our algorithm using  $N = 8$  diffracted orders with three-wave and five-wave calculations using direct resolution. The three-wave results exhibit poor agreement with the two other curves and become erratic for  $d/\lambda \geq 7.4$  because of too large exponential values. The beginning of the five-wave curve obtained by direct resolution is almost exactly superimposed upon the results of our algorithm, but numerical problems occur for  $d/\lambda \geq 3$ . Figure 6 represents the efficiency variations of transmitted, reflected, and diffracted orders calculated with our new algorithm retaining 8 diffracted

orders. These curves are similar to those presented in Ref. 15 for groove depths  $d/\lambda \leq 4$ .

In order to demonstrate good stability with very deep grooves, we extended the curves of Fig. 6 up to  $d/\lambda = 10$ . Our algorithm remains stable for much larger values that may not correspond to common physical situations: we increased the groove depth beyond 1000 wavelengths without any numerical instability. We also tested the behavior of the computer program when the number  $N$  of diffracted orders is increased: for  $N \geq 8$  the changes in the efficiencies of Fig. 6 are less than  $10^{-5}$ . Our computer program was able to handle up to 18 diffracted orders; beyond this value of  $N$  the iterative eigenvector searching routine diverged.

In the analysis we explained that our matrix formalism and our algorithm could also be applied to TM polarization; this is illustrated in Fig. 7, where we have plotted the diffraction efficiencies as functions of the groove depth for the TM case.

##### 2. Stairstep Grating

The second example is a grating with a stairstep profile that we represent as two cascaded binary gratings with thickness  $d/2$ . In Fig. 8 we distinguish two geometries,

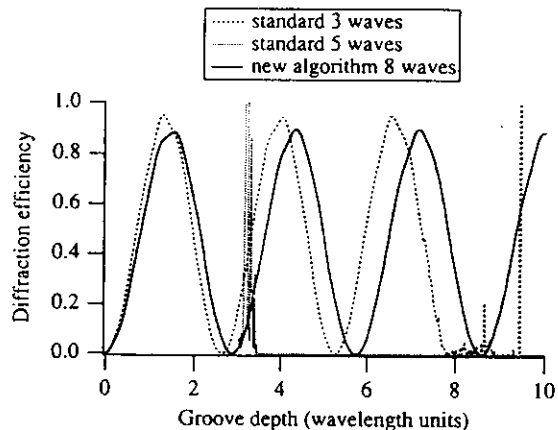


Fig. 5. Comparison between the standard and new algorithms for rectangular surface profile: variations of the  $-1$ -order diffraction efficiency as a function of the groove depth in wavelengths with TE polarization.

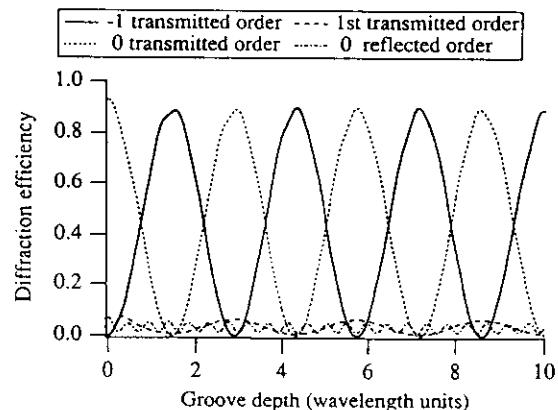


Fig. 6. Variations of the main diffracted orders as functions of the groove depth in wavelengths for the rectangular surface profile with TE polarization.

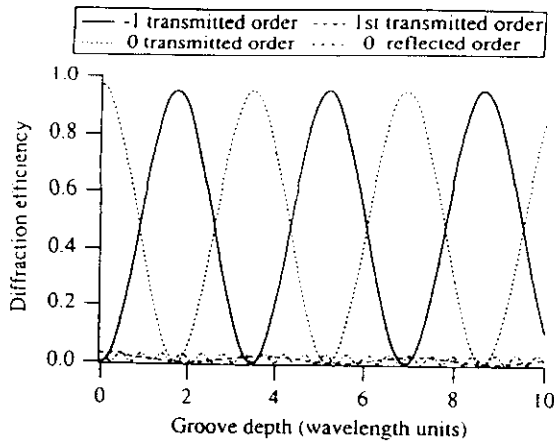


Fig. 7. Variations of the main diffracted orders as functions of the groove depth in wavelengths for the rectangular surface profile with TM polarization.

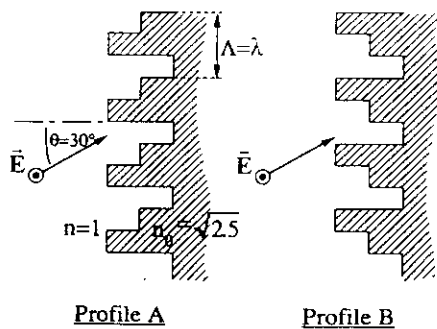


Fig. 8. Two possible geometries for a staircase surface profile; the spatial period is equal to the incident wavelength, and the angle of incidence is  $\theta = 30^\circ$ .

denoted A and B, corresponding to opposite profiles or incidence angles. The variations of diffraction efficiencies versus the groove depth with TE polarization are represented for the two opposite profiles in Fig. 9; the maximum values of diffraction efficiencies in order -1 are equal to those given in Ref. 15 within 0.1% (A, 67.7%; B, 71.8%). We extended the calculation again to very large values of  $d/\Lambda$  ( $>1000$ ) without any numerical difficulty.

**B. Holographic Volume Gratings**

In this subsection we present numerical results for slanted volume gratings with attenuated index modulation. Both index and absorption gratings are considered. The recording medium is a holographic film with thickness  $d = 13.5 \mu\text{m}$  and refractive index  $n_0 = 1.53$  deposited upon a glass substrate with index  $n_s = 1.5$ . The incidence medium is air. At recording, the photosensitive layer is illuminated by two uniform and coherent plane waves with wavelength  $\lambda = 500 \text{ nm}$  and respective angles of incidence in air of  $\theta_1 = -45^\circ$  and  $\theta_2 = +30^\circ$  (see Fig. 10). The corresponding angles inside the film,  $\theta_1' = -27.5^\circ$  and  $\theta_2' = +19.1^\circ$ , define a slanted interference pattern. For greater similarity with a real grating we account in our model for the absorption of the recording light by the holographic material: the exposure intensity exponentially decays from the surface to the second interface, and we assume that the residual amplitude on the film-substrate interface is half the am-

plitude of that on the air-film face. Thus we write the expression of the recording light intensity inside the film as follows:

$$E(x, z) = 2^{-z/d} (1 + \cos\{ (2\pi n_0/\lambda) [x(\sin \theta_1' + \sin \theta_2') + z(\cos \theta_1' + \cos \theta_2')] \} ) \quad (4.1)$$

Depending on the material that is used, either an index or an absorption grating is obtained. In both cases we assume that the energy response of the material is linear; thus the modulation amplitude is exponentially attenuated in the film depth. In our computation we treat the attenuated grating as a stack of several cascaded elementary gratings with uniform modulation, and we retain  $N = 5$  diffracted orders.

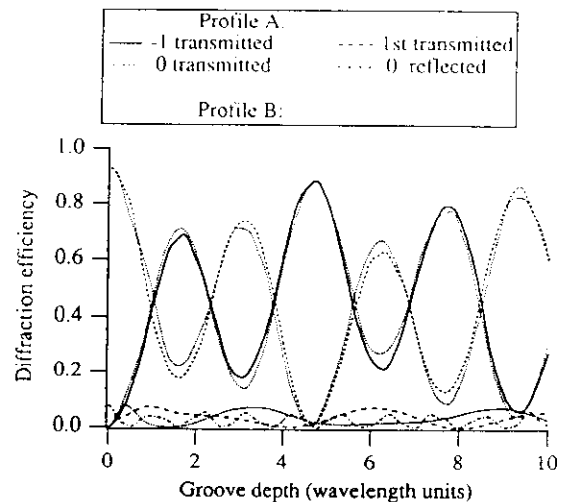


Fig. 9. Variations of the main diffracted orders as functions of the groove depth in wavelength units for two staircase surface profiles with TE polarization. The zero reflected order of case B is superimposed upon that of case A; the other orders of case B are recognizable by their similarity with case A.

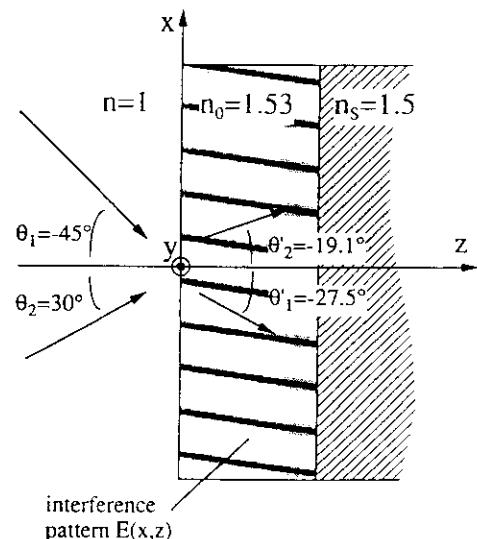


Fig. 10. Holographic recording geometry of a planar volume grating.

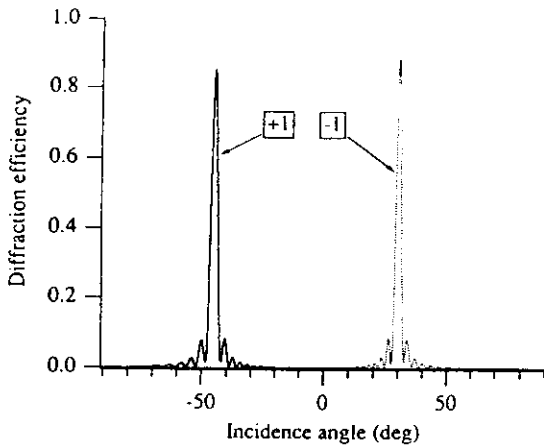


Fig. 11. Angular variations of diffraction efficiencies in +1 and -1 orders for the phase volume grating with TE polarization.

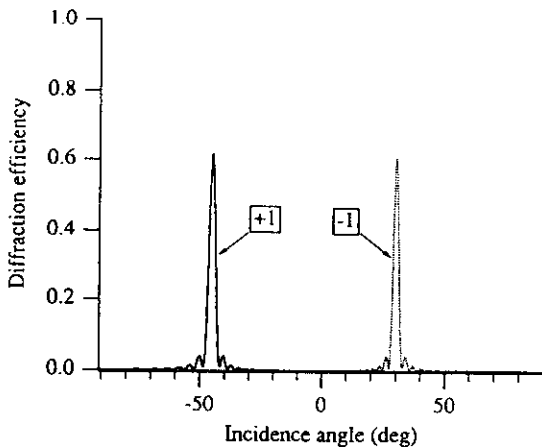


Fig. 12. Angular variations of diffraction efficiencies in +1 and -1 orders for the phase volume grating with TM polarization.

**Table 1. Example of Eigenvalue Distribution, Ordered by Growing Real Part, for Phase and Absorption Volume Gratings, Retaining Five Orders, with Wavelength  $\lambda = 500$  nm and Incidence Angle  $\theta = -30^\circ$**

	Phase Grating		Absorption Grating	
	Real	Imaginary	Real	Imaginary
1	-2.4921071	-0.2187540	-2.4802980	-0.2093668
2	-1.1757074	0.1436834	-1.1504568	0.1543284
3	-0.7951013	-0.1281361	-0.7574917	-0.1065848
4	0.0000000	-1.4637870	-0.0158833	1.4090277
5	0.0000000	1.3775048	-0.0052835	1.4090145
6	0.0000000	-1.2813433	0.0102463	-1.4827499
7	0.0000000	1.3987121	0.0109204	-1.3046831
8	0.7951013	-0.1281361	0.7574917	-0.1457060
9	1.1757074	0.1436834	1.1504568	0.1285696
10	2.4921071	-0.2187540	2.4802980	-0.2231450

1. Phase Grating

In a holographic material such as photopolymers or dichromated gelatin, the exposition by an interference pattern induces variations of the real part of the refrac-

tive index. In our simulation we discretize the modulation over the elementary slices, assuming the following index response:

$$n(x, z) = n_0 - 0.02E(x, z). \tag{4.2}$$

Figure 11 represents the angular response of the attenuated index grating, replayed at  $\lambda = 500$  nm with TE polarization, in the -1 and +1 diffracted orders. Figure 12 shows the corresponding results obtained with TM polarization. In these examples the grating was divided into  $m = 10$  elementary gratings. For a fixed incidence angle of  $\theta = -30^\circ$  we printed the eigenvalues of the first characteristic matrix, which corresponds to the closest grating to the air surface. Table 1 contains these values ordered by growing real part. As predicted in Appendix A, the opposite of the complex conjugate of each eigenvalue in Table 1 is also an eigenvalue. We observed the evolution of the results when the number of elementary gratings was increased to  $m = 100$ : the algorithm was perfectly stable, and the efficiency variations remained within  $10^{-5}$  for  $m < 24$ .

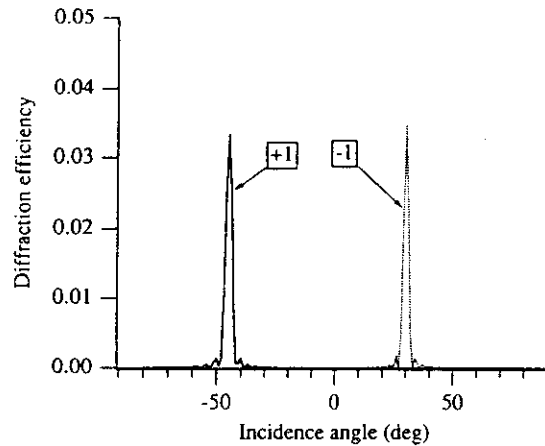


Fig. 13. Angular variations of diffraction efficiencies in +1 and -1 orders for the absorption volume grating with TE polarization.

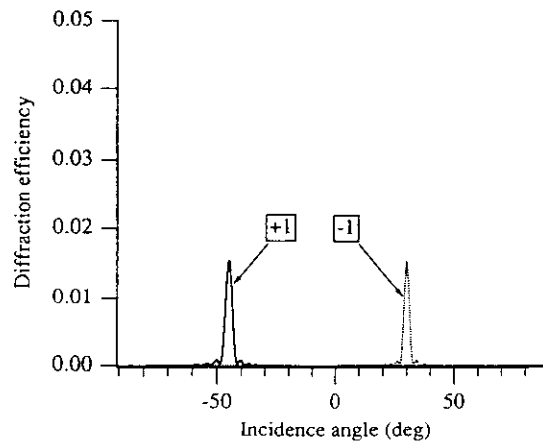


Fig. 14. Angular variations of diffraction efficiencies in +1 and -1 orders for the absorption volume grating with TM polarization.



## 2. Absorption Grating

In this case, for instance in a silver halide photographic plate, variations of the imaginary part of the refractive index are obtained. The final index is assumed to be given by

$$n(x, z) = n_0 + 0.01jE(x, z). \quad (4.3)$$

Figure 13 shows the angular variations of diffraction efficiency of the attenuated absorption grating, reconstructed at  $\lambda = 500$  nm with TE polarization, in the  $-1$  and  $+1$  diffracted orders. The results for TM polarization are represented in Fig. 14. As above, we give an example of the eigenvalue distribution in Table 1. Confirming the initial assumption of Section 3, we observe five eigenvalues with a negative real part and five eigenvalues with a positive real part. Only three pairs of eigenvalues corresponding to nonpropagating modes have opposite real parts, and the imaginary parts are all different.

## 5. CONCLUSION

We have derived a characteristic formalism for the rigorous coupled-wave theory of grating diffraction, which applies to planar volume gratings, surface relief gratings, and stacks of planar gratings with equal period. With the help of this formalism we proposed a new algorithm that overcomes numerical instabilities encountered in the coupled-wave modeling of very deep modulated structures. We presented numerical results for surface relief gratings and for phase and amplitude nonuniform volume holographic gratings. These results revealed the very good stability of our algorithm. In conclusion, the method permits accurate calculation of light diffraction by gratings of arbitrary profile and thickness; we believe that it contributes to enlargement of the field of application of the coupled-wave theory.

## APPENDIX A: JUSTIFICATION OF THE EIGENVALUE REDISTRIBUTION IN THE CASE OF A PURE DIELECTRIC GRATING

In our description of the algorithm (Section 3) we made the following hypothesis about the matrix  $[\mathbf{M}]$  given by relation (2.14): the numbers of eigenvalues with a critical negative real part and of those with a critical positive real part are both smaller than  $N$ . We now demonstrate this property in the case of a nonabsorbing grating. In this appendix complex conjugation is represented by an asterisk, and matrix transposition is denoted by the superscript  $T$ .

We consider an arbitrary eigenvalue  $x$  of matrix  $[\mathbf{M}]$  and its complex conjugate  $x^*$ . We know that  $x$  is a root of the characteristic polynomial  $p(X)$  of matrix  $[\mathbf{M}]$ :

$$p(x) = \det([\mathbf{M}] - x[\mathbf{I}_{2N}]) = 0. \quad (A1)$$

We now calculate the value of  $p(X)$  at  $X = -x^*$ :

$$\begin{aligned} p(-x^*) &= \det([\mathbf{M}] + x^*[\mathbf{I}_{2N}]) \\ &= \det \left( \left[ \begin{array}{c|c} -jK_z[\Delta] + x^*[\mathbf{I}_N] & -jk_0[\mathbf{I}_N] \\ \hline -jk_0[\Omega] & -jK_z[\Delta] + x^*[\mathbf{I}_N] \end{array} \right] \right). \end{aligned} \quad (A2)$$

If we transpose the matrix in Eq. (A2), the diagonal submatrices  $-jK_z[\Delta] + x^*[\mathbf{I}_N]$  are not affected, and the determinant remains unchanged:

$$\begin{aligned} p(-x^*) &= \det \left( \left[ \begin{array}{c|c} -K_z j[\Delta] + x^*[\mathbf{I}_N] & -jk_0[\Omega]^T \\ \hline -jk_0[\mathbf{I}_N]^T & -jK_z[\Delta] + x^*[\mathbf{I}_N] \end{array} \right] \right). \end{aligned} \quad (A3)$$

Since the modulated index profile is real, the coefficients of its Fourier expansion (2.1) verify that

$$\bar{n}_{-i} = \bar{n}_i^*, \quad i \in \mathbb{Z}. \quad (A4)$$

This implies that the  $N \times N$  matrix  $[\Omega]$  defined by Eqs. (2.16) is Hermitian:

$$[\Omega]^T = [\Omega]. \quad (A5)$$

Introducing Eq. (A5) into Eq. (A3) and using the fact that both matrices  $[\Delta]$  and  $[\mathbf{I}_N]$  are real, we obtain

$$\begin{aligned} p(-x^*) &= \det \left( \left[ \begin{array}{c|c} (jK_z[\Delta])^* + x^*[\mathbf{I}_N]^* & (jk_0[\Omega])^* \\ \hline (jk_0[\mathbf{I}_N])^* & (jK_z[\Delta])^* + x^*[\mathbf{I}_N]^* \end{array} \right] \right). \end{aligned} \quad (A6)$$

The permutations between the first and last  $N$  lines and between the first and last  $N$  columns yield

$$\begin{aligned} p(-x^*) &= (-1)^{2N} \\ &\times \det \left( \left[ \begin{array}{c|c} (jK_z[\Delta])^* + x^*[\mathbf{I}_N]^* & (jk_0[\mathbf{I}_N])^* \\ \hline (jk_0[\Omega])^* & (jK_z[\Delta])^* + x^*[\mathbf{I}_N]^* \end{array} \right] \right). \end{aligned} \quad (A7)$$

Thus we recognize the opposite of the complex conjugate

of  $p(x)$ :

$$p(-x^*) = -[p(x)]^* = 0. \quad (\text{A8})$$

We have demonstrated that if  $x$  is an eigenvalue of matrix  $[M]$ , then the opposite of its complex conjugate is also an eigenvalue of  $[M]$ . In other words, either the eigenvalue  $x$  is purely imaginary or another eigenvalue exists with the same imaginary part and the opposite real part. Thus the numbers of eigenvalues with a critical negative real part and of those with a critical positive real part are equal.

\*Present address, ESSILOR International, 81 Boulevard Oudry, 94 000 Créteil, France.

## REFERENCES

1. T. K. Gaylord and M. G. Moharam, "Analysis and applications of optical diffraction by gratings," *Proc. IEEE* **73**, 894-937 (1985).
2. R. Petit, ed., *Electromagnetic Theory of Gratings* (Springer-Verlag, Berlin, 1980).
3. L. Solymar and D. J. Cooke, *Volume Holography and Volume Gratings* (Academic, London, 1981).
4. H. Kogelnik, "Coupled wave theory for thick hologram gratings," *Bell Syst. Tech. J.* **48**, 2909-2947 (1969).
5. J. A. Kong, "Second-order coupled-mode equations for spatially periodic media," *J. Opt. Soc. Am.* **67**, 825-829 (1977).
6. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," *J. Opt. Soc. Am.* **71**, 811-818 (1981).
7. M. V. Vasnetsov, M. S. Soskin, and V. B. Taranenko, "Grazing diffraction by volume phase gratings," *Opt. Acta* **32**, 891-899 (1985).
8. X. Y. Chen, "Using the finite element method to solve coupled wave equations in volume holograms," *J. Mod. Opt.* **35**, 1383-1391 (1988).
9. T. K. Gaylord and M. G. Moharam, "Coupled-wave analysis of reflection gratings," *Appl. Opt.* **20**, 240-244 (1981).
10. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of grating diffraction—E-mode polarization and losses," *J. Opt. Soc. Am.* **72**, 1385-1392 (1982).
11. M. G. Moharam and T. K. Gaylord, "Three-dimensional vector coupled-wave analysis of planar grating diffraction," *J. Opt. Soc. Am.* **73**, 1105-1112 (1983).
12. M. Nevière, "Sur un formalisme différentiel pour les problèmes de diffraction dans le domaine de résonance: application à l'étude des réseaux optiques et de diverses structures périodiques," thèse de doctorat (Université d'Aix-Marseille, Aix-Marseille, France, 1975).
13. D. E. Treman and K. K. Mei, "Application of the unimoment method to scattering from periodic dielectric structures," *J. Opt. Soc. Am.* **68**, 775-783 (1978).
14. K. C. Chang, V. Shah, and T. Tamir, "Scattering and guiding of waves by dielectric gratings with arbitrary profiles," *J. Opt. Soc. Am.* **70**, 804-813 (1980).
15. M. G. Moharam and T. K. Gaylord, "Diffraction analysis of dielectric surface-relief gratings," *J. Opt. Soc. Am.* **72**, 1385-1392 (1982).
16. A. K. Cousins and S. C. Gottschalk, "Application of the impedance formalism to diffraction gratings with multiple coating layers," *Appl. Opt.* **29**, 4268-4271 (1990).
17. D. Kermisch, "Non-uniform sinusoidally modulated dielectric gratings," *J. Opt. Soc. Am.* **59**, 1409-1413 (1969).
18. I. R. Redmond, "Holographic optical elements in dichromated gelatin," Ph.D. dissertation (Heriot-Watt University, Edinburgh, UK, 1989).
19. E. N. Glytsis and T. K. Gaylord, "Three-dimensional (vector) rigorous coupled-wave analysis of anisotropic grating diffraction," *J. Opt. Soc. Am. A* **7**, 1399-1420 (1990).
20. D. M. Pai and K. A. Awada, "Analysis of dielectric gratings of arbitrary profiles and thicknesses," *J. Opt. Soc. Am. A* **8**, 755-762 (1991).
21. M. Born and E. Wolf, *Principles of Optics*, 6th ed. (Pergamon, New York, 1987), pp. 36-70.

# Differential theory of gratings: extension to deep gratings of arbitrary profile and permittivity through the $R$ -matrix propagation algorithm

F. Montiel and M. Nevière

Laboratoire d'Optique Electromagnétique, Unité de Recherche Associée au Centre de la Recherche Scientifique 843, Faculté des Sciences et Techniques, Centre de St-Jérôme, Case 262, 13397 Marseille Cedex 20, France

Received February 23, 1994; revised manuscript received July 27, 1994; accepted July 28, 1994

The analysis of gratings of arbitrary depth, profile, and permittivity is conducted by cutting the modulated region into different slices for which the differential theory of gratings is able to compute the diffracted field for both TE and TM polarization without numerical instabilities. The use of a suitable transition matrix ( $R$  matrix) then allows one to analyze the entire stack without encountering the numerical instabilities that generally occur with use of the  $T$ -transmission matrix, which is well known in stratified media theory. The use of the  $R$ -matrix propagation algorithm provides a breakthrough for grating theoreticians in the sense that it not only permits the study of grating of arbitrary depth but also eliminates the numerical instabilities that have plagued the differential theory in TM polarization during the past 20 years.

## 1. INTRODUCTION

The past 20 years have seen the spread of grating use from the restricted domain of spectroscopy to various domains of physics, including acoustics, solid-state physics, nonlinear optics, x-ray instrumentation, optical communications, and optical computing; and gratings have begun to appear in common life as safety features on credit cards, bank notes, and stamps as well as in a wide variety of display and advertising applications.<sup>1</sup> The result is that grating theoreticians are confronted with gratings whose groove-depth-to-groove-spacing ratio,  $h/d$ , is no longer limited to the classical (0.05–0.5) range used in spectroscopy<sup>2</sup> but can reach several units. As the groove depth increases, the boundary value problem that is related to Maxwell's equations and the associated boundary conditions on the grating surface become increasingly difficult to resolve numerically. Both the computation time (i.e., the cost of the calculation) and the number of terms used to describe the electromagnetic field increase. The presence of ever larger undesired exponential functions (those terms of Rayleigh expansions that one wants to eliminate because of their divergent behavior at infinity) may produce overflows and contaminate the desired solutions, leading to a loss of accuracy. Thus all existing grating formalisms<sup>3</sup> have limitations with respect to the modulations that they can tackle; these limitations depend strongly on the spectral domain, the refractive index of the grating material, and the polarization of the incident light.

The first idea that can be tried in the attempt to overcome this limitation is to cut the modulated region into different slices that are thin enough that the usual formalisms can be used. It is then possible to determine the  $T$  transmission matrix of each slice.<sup>4,5</sup> Then, as for plane stratified media,<sup>6</sup> the  $T$  matrix of the entire grating is simply the product of the matrices of the different slices. Such a process has turned out to be quite effi-

cient for studying x-ray gratings etched inside a multi-dielectric layer.<sup>5</sup> However, the  $T$ -matrix propagation algorithm is known to be unstable. Thus it does not improve the range of validity of any formalism with respect to the groove depth.

Two recent contributions have brought some new ideas to the problem of increasing the range of validity. The first one is the multiple-reflection series of Pai and Awada,<sup>7</sup> which sews together the reflected and the transmitted orders at each interface in a way similar to that of the series used to study the Fabry–Perot interferometer. The second is the so called  $R$ -matrix propagation algorithm<sup>8,9</sup> developed in 1976 to study chemical reactions and recently introduced in grating theory by DeSandre and Elson<sup>10</sup> and Li.<sup>11</sup> Unlike the research in Ref. 7, which was presented only for dielectric gratings used in TE polarization, the research in Ref. 11 was developed for both dielectric and metallic gratings used in conical diffraction. This research extends the modal method of Botten *et al.*<sup>12,13</sup> to conical mountings<sup>14</sup> with the  $R$ -matrix propagation algorithm and produces interesting convergent results. We have tried the methods of both Ref. 7 and Ref. 11 and found the latter to be much more powerful. Thus we decided to use the  $R$ -matrix propagation algorithm to try to improve the range of validity of the differential formalism.<sup>15</sup> The choice of this method is related to the method's wide applicability to any groove shape, spectral domain, stack of gratings, grating material, phase and amplitude gratings, and so on. However, for highly reflecting metallic gratings, because of numerical instabilities the previous method was strictly limited for TM polarization and visible or near-infrared regions to shallow gratings.<sup>16</sup> It was thus tempting to see whether the new method could get rid of the old problem. Our research differs from the research in Ref. 11 in that we do not use the multilayer modal method. In addition, we implement the  $R$ -matrix propagation algorithm in two different ways,

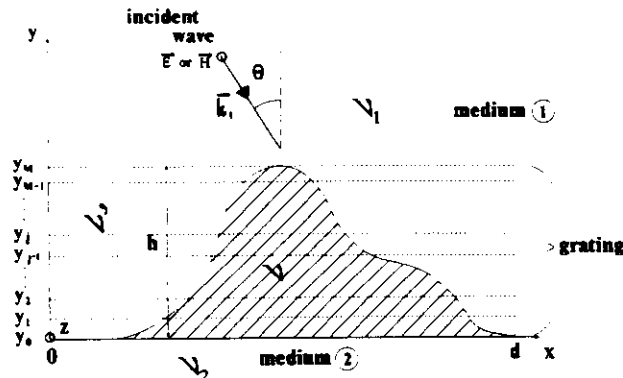


Fig. 1. Decomposition of the modulated region into  $M$  different slices.

and we compare the corresponding numerical results from the two methods.

## 2. THEORY

Figure 1 shows a period of a surface periodically modulated with period  $d$  with respect to  $x$  and describes the notation used in what follows. The groove shape is arbitrary. The refractive index of the bumps,  $\nu$ , can be different from the refractive index of the substrate,  $\nu_2$ , and the refractive index inside the grooves,  $\nu'$ , can be different from that of the superstrate,  $\nu_1$ . An incident linearly polarized plane wave illuminates the grating under incidence  $\theta$ , and the incident wave vector lies in the cross-section plane of the grating. The general vectorial problem is thus reduced to the study of the two fundamental cases of polarization, and the unknown function  $u(x, y)$  is the  $z$  component of the electric or the magnetic field for TE and TM polarizations, respectively.

Outside the modulated region defined by  $0 \leq y \leq h$ , the field (in media 1 and 2) can be represented by Rayleigh expansions:

$$u_1 = \sum_n \{A_n^{(1)} \exp[-i\beta_n^{(1)}y] + B_n^{(1)} \exp[i\beta_n^{(1)}y]\} \exp(i\alpha_n x), \quad (1)$$

$$u_2 = \sum_n \{A_n^{(2)} \exp[-i\beta_n^{(2)}y] + B_n^{(2)} \exp[i\beta_n^{(2)}y]\} \exp(i\alpha_n x), \quad (1')$$

with

$$\begin{aligned} \alpha_n &= \alpha + n \frac{2\pi}{d}, & \alpha &= k_0 \nu_1 \sin \theta, \\ k_0 &= \frac{2\pi}{\lambda}, & \beta_m^{(i)} &= \sqrt{k_0^2 \nu_i^2 - \alpha_n^2}, \\ i &\in [1, 2], & \text{Re } \beta_n^{(i)} + \text{Im } \beta_n^{(i)} &> 0, \end{aligned}$$

and  $\lambda$  is the wavelength in vacuum. Of course, when applied to the entire grating, Eqs. (1) and (1') have to be simplified through the use of the outgoing wave condition, which implies that  $A_n^{(1)} = \delta_{n0}$ ,  $B_n^{(2)} = 0$ , and  $\forall n$ ; but when we consider an arbitrary slice as illustrated in Fig. 1, lying between ordinates  $y_{j-1}$  and  $y_j$ , with  $j \in [2, M-1]$ , all  $A_n^{(j)}$  and  $B_n^{(j)}$  Rayleigh coefficients have to be kept. However, at the computational level only  $N$  values of  $n$  will be retained.  $N$  will be called the truncation order, and the number  $M$  of slices will be called the stratification order.

### A. Definition of the $t$ Transmission Matrix

The definition of the  $t$  matrix used here is a little different from the one previously used in grating theory<sup>4,5</sup> and denoted  $T$ . Expanding the field inside the modulated region by a modified Fourier series,

$$u = \sum_n U_n(y) \exp(i\alpha_n x), \quad (2)$$

we first note that the continuity of the tangential components of  $\mathbf{E}$  and  $\mathbf{H}$  at  $y = y_j$  leads to the continuity of  $u(x, y)$  and  $(\partial u / \partial y)(x, y)$  for the TE case and of  $u(x, y)$  and  $1/[k^2(x, y)](\partial u / \partial y)(x, y)$  for the TM case, where  $k(x, y)$  is the modulus of the wave vector in the various regions. Let us call  $V_n(y)$  the components of the  $\exp(i\alpha_n x)$  basis of functions  $\partial u / \partial y$  in the TE case and  $(1/k^2)(\partial u / \partial y)$  in the TM case. We introduce the  $t$  matrix of the  $j$ th layer as the matrix linking the components of continuous quantities through Eq. (3):

$$\begin{bmatrix} U_n(y_j) \\ V_n(y_j) \end{bmatrix} = t^{(j)} \begin{bmatrix} U_n(y_{j-1}) \\ V_n(y_{j-1}) \end{bmatrix}. \quad (3)$$

The reader interested in the method of computing the  $t^{(j)}$  matrices can consult Ref. 5, which gives all the details of obtaining the  $T$  matrix, which is similar to the  $t$  matrix. The use of Eq. (3) from  $j = 1$  to  $j = M$  shows that

$$\begin{bmatrix} U_n(h) \\ V_n(h) \end{bmatrix} = t^{(M)} t^{(M-1)} \dots t^{(j)} \dots t^{(2)} t^{(1)} \begin{bmatrix} U_n(0) \\ V_n(0) \end{bmatrix}. \quad (4)$$

Since there exist linear relations<sup>5</sup> between  $U_n$  and  $V_n$  on one side and the Rayleigh coefficients  $[A_n^{(1)}, B_n^{(1)}]$  or  $[A_n^{(2)}, B_n^{(2)}]$  on the other side, Eq. (4) enables us, through the method described in Refs. 5 and 17, to compute the Rayleigh coefficients everywhere and thus to find the efficiencies diffracted in the various spectral orders. However, when the groove depth is high enough, such a method is known to be unstable. This is the reason that instead of using the  $t$  matrix we introduce the  $r$  reflection matrix.

### B. Definition of the $r$ Matrix

Let us define a new matrix  $r^{(j)}$ , for the  $j$ th slice, by Eq. (5):

$$\begin{bmatrix} U_n(y_{j-1}) \\ U_n(y_j) \end{bmatrix} = r^{(j)} \begin{bmatrix} V_n(y_{j-1}) \\ V_n(y_j) \end{bmatrix} \quad (5)$$

and divide the  $2N \times 2N$   $r^{(j)}$  matrix into four  $N \times N$  matrices, as shown in Eq. (6):

$$r^{(j)} = \begin{bmatrix} r_{11}^{(j)} & r_{12}^{(j)} \\ r_{21}^{(j)} & r_{22}^{(j)} \end{bmatrix}. \quad (6)$$

Let us apply the same decomposition to matrix  $t^{(j)}$ . Elementary algebra, with use of Eqs. (3) and (5), allows us to express the new  $N \times N$   $r$  matrices as functions of the  $t$  matrices by

$$\begin{aligned} r_{11}^{(j)} &= -[t_{21}^{(j)}]^{-1} t_{22}^{(j)}, \\ r_{12}^{(j)} &= [t_{21}^{(j)}]^{-1}, \\ r_{21}^{(j)} &= t_{12}^{(j)} - t_{11}^{(j)} [t_{21}^{(j)}]^{-1} t_{22}^{(j)}, \\ r_{22}^{(j)} &= t_{11}^{(j)} [t_{21}^{(j)}]^{-1}. \end{aligned} \quad (7)$$

**C. R-Matrix Propagation Algorithm**

Let us now consider the global  $R$  matrix defined by

$$\begin{bmatrix} U_n(0) \\ U_n(y_j) \end{bmatrix} = R^{(j)} \begin{bmatrix} V_n(0) \\ V_n(y_j) \end{bmatrix} \quad (8)$$

and divide it into four submatrices as stated by Eq. (6).

Previous research<sup>8,9</sup> has established that the block elements of the global  $R$  matrix obey a set of recursion formulas which are recalled in Ref. 11:

$$\begin{aligned} R_{11}^{(j)} &= R_{11}^{(j-1)} + R_{12}^{(j-1)} Z^{(j)} R_{21}^{(j-1)}, \\ R_{12}^{(j)} &= -R_{12}^{(j-1)} Z^{(j)} r_{12}^{(j)}, \\ R_{21}^{(j)} &= r_{21}^{(j)} Z^{(j)} R_{21}^{(j-1)}, \\ R_{22}^{(j)} &= r_{22}^{(j)} - r_{21}^{(j)} Z^{(j)} r_{12}^{(j)}, \end{aligned} \quad (9)$$

where

$$Z^{(j)} = [r_{11}^{(j)} - R_{22}^{(j-1)}]^{-1}.$$

The  $R$ -matrix propagation algorithm starts from the value of  $R^{(1)}$ , which of course is equal to  $r^{(1)}$ ; the block elements of  $R^{(1)}$  are given by Eqs. (7), in which  $j = 1$ . The  $R$ -matrix propagation algorithm computes the  $r^{(j)}$  matrix for each slice from Eqs. (7) and deduces the  $R^{(j)}$  matrices from Eqs. (9). The process ends with the production of matrix  $R^{(M)}$ . We get

$$\begin{bmatrix} U_n(0) \\ U_n(y_M) \end{bmatrix} = R^{(M)} \begin{bmatrix} V_n(0) \\ V_n(y_M) \end{bmatrix}. \quad (10)$$

**D. Determination of the Rayleigh Coefficients**

The matching of the numerical solution at the frontiers of the modulated area ( $y = 0$  and  $y = h$ ) with the corresponding Rayleigh expansions leads to

$$\begin{aligned} U_n(h) &= A_n^{(1)} \exp(-i\beta_{1,n}h) + B_n^{(1)} \exp(i\beta_{1,n}h), \\ V_n(h) &= q_1(-i\beta_{1,n})[A_n^{(1)} \exp(-i\beta_{1,n}h) - B_n^{(1)} \exp(i\beta_{1,n}h)], \\ U_n(0) &= A_n^{(2)} + B_n^{(2)}, \\ V_n(0) &= q_2(-i\beta_{2,n})[A_n^{(2)} - B_n^{(2)}], \end{aligned}$$

where

$$q_i = \begin{cases} 1 & \forall i \text{ for TE polarization} \\ \frac{1}{k_0^2 \nu_i^2} & \text{for TM polarization, } i = 1, 2 \end{cases}$$

Then Eq. (10) leads to

$$\begin{bmatrix} A_n^{(2)} + B_n^{(2)} \\ A_n^{(1)} \exp(-i\beta_{1,n}h) + B_n^{(1)} \exp(i\beta_{1,n}h) \end{bmatrix} = \begin{bmatrix} R_{11}^{(M)} & R_{12}^{(M)} \\ R_{21}^{(M)} & R_{22}^{(M)} \end{bmatrix} \times \begin{bmatrix} q_2(-i\beta_{2,n})[A_n^{(2)} - B_n^{(2)}] \\ q_1(-i\beta_{1,n})[A_n^{(1)} \exp(-i\beta_{1,n}h) - B_n^{(1)} \exp(i\beta_{1,n}h)] \end{bmatrix}. \quad (11)$$

Overflow problems linked with exponentially growing functions are avoided by the introduction of new Rayleigh coefficients, given by

$$\begin{aligned} \tilde{A}_n^{(1)} &= A_n^{(1)} \exp(-i\beta_{1,n}h), \\ \tilde{B}_n^{(1)} &= B_n^{(1)} \exp(i\beta_{1,n}h). \end{aligned} \quad (12)$$

Equation (11) leads to the linear algebraic set

$$\begin{aligned} \{A_n^{(2)}\} + \{B_n^{(2)}\} &= -iq_2 R_{11}^{(M)}[\beta_{2,n}]\{\{A_n^{(2)}\} - \{B_n^{(2)}\}\} \\ &\quad - iq_1 R_{12}^{(M)}[\beta_{1,n}]\{\{\tilde{A}_n^{(1)}\} - \{\tilde{B}_n^{(1)}\}\}, \\ \{\tilde{A}_n^{(1)}\} + \{\tilde{B}_n^{(1)}\} &= -iq_2 R_{21}^{(M)}[\beta_{2,n}]\{\{A_n^{(2)}\} - \{B_n^{(2)}\}\} \\ &\quad - iq_1 R_{22}^{(M)}[\beta_{1,n}]\{\{\tilde{A}_n^{(1)}\} - \{\tilde{B}_n^{(1)}\}\}, \end{aligned} \quad (12)$$

where  $\{\}$  represents a column vector with  $N$  elements and  $[\ ]$  represents  $N \times N$  diagonal matrices with elements  $\beta_{i,n}$  ( $i = 1, 2$ ).

The use of the outgoing wave condition implies that  $A_n^{(1)} = \delta_{n,0}$  and  $B_n^{(2)} = 0$ , and we call  $T_n$  the Rayleigh coefficients  $[A_n^{(2)}]$  of the downgoing waves in the substrate. If we introduce four new  $N \times N$  matrices  $P_{ij}$  by

$$\begin{aligned} P_{11} &= iq_2 R_{11}^{(M)}[\beta_{2,n}], \\ P_{12} &= iq_1 R_{12}^{(M)}[\beta_{1,n}], \\ P_{21} &= iq_2 R_{21}^{(M)}[\beta_{2,n}], \\ P_{22} &= iq_1 R_{22}^{(M)}[\beta_{1,n}], \end{aligned}$$

and with

$$Q = \{\exp(-i\beta_{1,n}h)\delta_{n,0}\},$$

the unknown Rayleigh coefficients  $\tilde{B}_n^{(1)}$  and  $T_n$  are solutions of the linear algebraic system

$$\begin{bmatrix} \mathbf{I} + P_{11} & -P_{12} \\ -P_{21} & -\mathbf{I} + P_{22} \end{bmatrix} \begin{bmatrix} \{T_n\} \\ \{\tilde{B}_n^{(1)}\} \end{bmatrix} = \begin{bmatrix} -P_{12}Q \\ (\mathbf{I} + P_{22})Q \end{bmatrix}, \quad (13)$$

where  $\mathbf{I}$  is the unit matrix.

Its resolution on a computer gives  $T_n$  and  $\tilde{B}_n^{(1)}$ , from which are deduced the efficiencies in the various spectral orders.

**E. Variant of the Method: The  $R'$ -Matrix Propagation Algorithm**

The aim of this subsection is to propose a different implementation of the  $R$ -matrix propagation algorithm, which will be called the  $R'$ -matrix propagation algorithm in what follows. In order to point out the similarities and differences between the two methods, we give the equations that are at the basis of the  $R'$ -matrix propagation algorithm the same numbers as those used for the  $R$ -matrix one but with primes added.

Let us define the  $r'^{(j)}$  matrix for the  $j$ th slice by

$$\begin{bmatrix} U_n(y_{j-1}) \\ V_n(y_j) \end{bmatrix} = r'^{(j)} \begin{bmatrix} V_n(y_{j-1}) \\ U_n(y_j) \end{bmatrix} \quad (5')$$

and divide it into four blocks:

$$r'^{(j)} = \begin{bmatrix} r'_{11}{}^{(j)} & r'_{12}{}^{(j)} \\ r'_{21}{}^{(j)} & r'_{22}{}^{(j)} \end{bmatrix}. \quad (6')$$

The use of Eqs. (3) and (5') allows us to derive the four blocks in terms of the  $t'^{(j)}$  blocks by

$$\begin{aligned} r'_{11}{}^{(j)} &= -(t'_{11}{}^{(j)})^{-1} t'_{12}{}^{(j)}, \\ r'_{12}{}^{(j)} &= [t'_{11}{}^{(j)}]^{-1}, \\ r'_{21}{}^{(j)} &= t'_{22}{}^{(j)} - t'_{21}{}^{(j)} (t'_{11}{}^{(j)})^{-1} t'_{12}{}^{(j)}, \\ r'_{22}{}^{(j)} &= t'_{21}{}^{(j)} [t'_{11}{}^{(j)}]^{-1}. \end{aligned} \quad (7')$$

We then introduce the global  $R'$  matrix defined by

$$\begin{bmatrix} U_n(0) \\ V_n(y_j) \end{bmatrix} = R^{(j)} \begin{bmatrix} V_n(0) \\ U_n(y_j) \end{bmatrix}. \tag{8'}$$

Appendix A establishes among its four blocks the following recurrence relations:

$$\begin{aligned} R_{11}^{(j)} &= R_{11}^{(j-1)} + R_{12}^{(j-1)} Z^{(j)} r_{11}^{(j)} R_{21}^{(j-1)}, \\ R_{12}^{(j)} &= R_{12}^{(j-1)} Z^{(j)} r_{12}^{(j)}, \\ R_{21}^{(j)} &= r_{21}^{(j)} Y^{(j)} R_{21}^{(j-1)}, \\ R_{22}^{(j)} &= r_{22}^{(j)} + r_{21}^{(j)} Y^{(j)} R_{22}^{(j-1)} r_{12}^{(j)}, \end{aligned} \tag{9'}$$

where

$$\begin{aligned} Z^{(j)} &= [\mathbf{I} - r_{11}^{(j)} R_{22}^{(j-1)}]^{-1}, \\ Y^{(j)} &= [\mathbf{I} - R_{22}^{(j-1)} r_{11}^{(j)}]^{-1}, \end{aligned}$$

and  $\mathbf{I}$  is a unit matrix.

Similarly to what was done in Subsection 2.C, the  $R'$ -matrix propagation algorithm may start from  $R^{(1)}$ , which is equal to  $r'(1)$ , whose block elements are given by Eqs. (7'), in which we take  $j$  equal to 1; we then use Eqs. (9')  $\forall j \in [2, M]$ . But in contrast to what happens for the first formulation of the algorithm (the  $R$ -matrix one), one may also initiate the algorithm with

$$R^{(0)} = \begin{bmatrix} 0 & \mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix}$$

and use Eqs. (9')  $\forall j \in [1, M]$ . Our main interest in this second formulation is to use it to provide an independent way of checking the numerical results.

From the numerical point of view, the two algorithms turn out to be equivalent. They give convergent results on the eighth or ninth digit when computations are conducted in double precision (i.e., with 16 digits) as soon as the integration step  $\delta y$  is approximately  $10^{-5}$  to  $10^{-4} \lambda$ , and they remain stable when  $\delta y \rightarrow 0$ , as long as  $\delta y \approx 10^{-10} \lambda$ .

### 3. NUMERICAL IMPLEMENTATION

Maxwell's equations used in the sense of distributions allow us to write the propagation equations in the entire space as

$$\begin{aligned} \frac{\partial \tilde{H}_x}{\partial y} &= -\frac{\partial^2 E_z}{\partial x^2} - k^2(x, y) E_z, \\ \frac{\partial E_z}{\partial y} &= \tilde{H}_x, \end{aligned} \tag{14}$$

where  $\tilde{H}_x = i\omega\mu_0 H_x$  and  $k(x, y)$  is the product of  $k_0$  by the refractive index at point  $(x, y)$ . This set holds for TE polarization and must be replaced by

$$\begin{aligned} \frac{\partial H_z}{\partial y} &= k^2 \tilde{E}_x, \\ \frac{\partial \tilde{E}_x}{\partial y} &= -\frac{\partial}{\partial x} \left( \frac{1}{k^2(x, y)} \frac{\partial H_z}{\partial x} \right) - H_z, \end{aligned} \tag{15}$$

where  $\tilde{E}_x = E_x / (i\omega\mu_0)$  for TM polarization.

Expanding  $k(x, y)$  on the Fourier basis with respect to the  $x$  coordinate and expanding the field on the  $\exp(i\alpha_n x)$  basis, we obtain

$$\begin{aligned} \frac{d\tilde{H}_{x,n}}{dy} &= \alpha_n^2 E_{z,n} = \sum_{m=-\infty}^{+\infty} (k^2)_{n-m} E_{z,m}, \\ \frac{dE_{z,n}}{dy} &= \tilde{H}_{x,n} \end{aligned} \tag{16}$$

for TE polarization and

$$\begin{aligned} \frac{dH_{z,n}}{dy} &= \sum_{m=-\infty}^{+\infty} (k^2)_{n-m} \tilde{E}_{x,m}, \\ \frac{d\tilde{E}_{x,n}}{dy} &= \alpha_n \sum_{m=-\infty}^{+\infty} \alpha_m \left( \frac{1}{k^2} \right)_{n-m} H_{z,m} - H_{z,n} \end{aligned} \tag{17}$$

for TM polarization.

Thus for both polarizations the propagation equation is transformed into a set of first-order coupled differential equations with nonconstant coefficients, for which no analytical solution exists. Only a numerical solution can be found with the help of a computer. This numerical integration is done with use of the classical fourth-order Runge-Kutta algorithm<sup>18</sup> after truncation of the set of equations to order  $N$  (i.e., after limitation of the field series to  $N$  components). Thus the computed absolute efficiencies  $e_p$  in the  $p$ th order, which are derived directly from the Rayleigh coefficients  $B_n^{(1)}$  and  $T_n$ , are indeed functions of  $M$  and  $N$ , where  $M$  is the stratification order. In order to check the convergence of the method we define two criteria  $\Delta_M$  and  $\Delta_N$  by

$$\begin{aligned} \Delta_M &= \log_{10} \left| \frac{e_p(M + M_0, N) - e_p(M, N)}{e_p(M, N)} \right|, \\ \Delta_N &= \log_{10} \left| \frac{e_p(M, N + N_0) - e_p(M, N)}{e_p(M, N)} \right|, \end{aligned}$$

where  $M_0$  and  $N_0$  are integer increments of integers  $M$  and  $N$ .

### 4. EXTENSION OF THE METHOD TO A STACK OF GRATINGS

It is worth noting that the present method can be easily extended to more-complicated periodic diffracting struc-

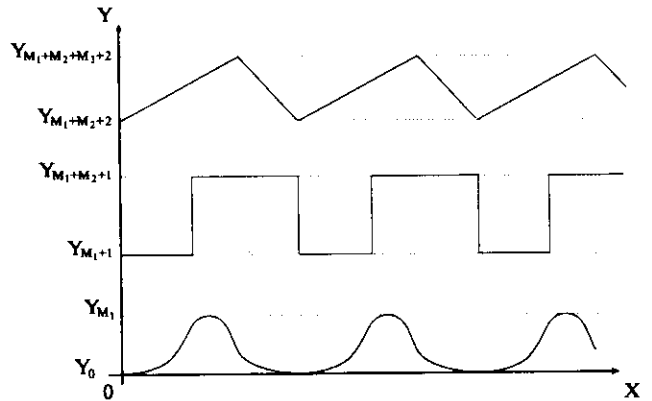


Fig. 2. Illustration of a stack of superimposed dielectric gratings.

**Table 1. Evolution of the Transmitted Efficiencies and Total Diffracted Energy as Functions of  $M$  for Grating Configuration 1 and TE and TM Polarizations**

$M$	$Q$	$e_0^t$	$e_1^t$	$e_2^t$	$\sum e_p$
<b>TE</b>					
2	150	—	—	—	3.699429712776
3	100	0.200358593616	$0.851696069801 \times 10^{-1}$	$0.915842843856 \times 10^{-1}$	1.000007146496
4	75	0.200358606774	$0.851696196233 \times 10^{-1}$	$0.915842843415 \times 10^{-1}$	1.000007200927
5	60	0.200358606773	$0.851696196228 \times 10^{-1}$	$0.915842843426 \times 10^{-1}$	1.000007200930
6	50	0.200358606773	$0.851696196228 \times 10^{-1}$	$0.915842843426 \times 10^{-1}$	1.000007200930
<b>TM</b>					
2	150	—	—	—	10.665929249490
3	100	—	—	—	1.025207828867
4	75	0.134345559535	$0.783486335950 \times 10^{-1}$	0.161406945312	1.000008158920
5	60	0.134345559610	$0.783486335581 \times 10^{-1}$	0.161406945355	1.000008158836
6	50	0.134345559609	$0.783486335582 \times 10^{-1}$	0.161406945355	1.000008158836
10	30	0.134345559610	$0.783486335581 \times 10^{-1}$	0.161406945355	1.000008158836
12	25	0.134345559610	$0.783486335579 \times 10^{-1}$	0.161406945355	1.000008158836
15	20	0.134345559610	$0.783486335581 \times 10^{-1}$	0.161406945355	1.000008158835
20	15	0.134345559610	$0.783486335581 \times 10^{-1}$	0.161406945355	1.000008158836
25	12	0.134345559610	$0.783486335581 \times 10^{-1}$	0.161406945355	1.000008158836

**Table 2. Same as Table 1 for Grating Configuration 2 Made by Rods of Chromium Embedded in a Dielectric Grating**

$M$	$Q$	$e_0^t$	$e_1^t$	$e_2^t$	$\sum e_n$
<b>TE</b>					
2	150	—	—	—	77.004115131057
3	100	—	—	—	17.149623137215
4	75	0.214987159632	$0.925631896432 \times 10^{-1}$	$0.640430990970 \times 10^{-3}$	0.685151606333
5	60	0.214986891210	$0.925617500762 \times 10^{-1}$	$0.640267975488 \times 10^{-3}$	0.685150462982
6	50	0.214986891178	$0.925617497050 \times 10^{-1}$	$0.640268013452 \times 10^{-3}$	0.685150463269
10	30	0.214986891247	$0.925617496724 \times 10^{-1}$	$0.640268007591 \times 10^{-3}$	0.685150463318
12	25	0.214986891247	$0.925617496724 \times 10^{-1}$	$0.640268007590 \times 10^{-3}$	0.685150463318
<b>TM</b>					
2	150	—	—	—	21.13900846
3	100	—	—	—	23.50368409
4	75	$0.866398874452 \times 10^{-1}$	$0.613453740205 \times 10^{-1}$	$0.200134798700 \times 10^{-1}$	0.528234288135
5	60	$0.866397062083 \times 10^{-1}$	$0.613453565735 \times 10^{-1}$	$0.200135862619 \times 10^{-1}$	0.528233956629
6	50	$0.866397074296 \times 10^{-1}$	$0.613453574941 \times 10^{-1}$	$0.200135865498 \times 10^{-1}$	0.528233959644
10	30	$0.866397074313 \times 10^{-1}$	$0.613453574958 \times 10^{-1}$	$0.200135865500 \times 10^{-1}$	0.528233959651
12	25	$0.866397074313 \times 10^{-1}$	$0.613453574958 \times 10^{-1}$	$0.200135865500 \times 10^{-1}$	0.528233959651

**Table 3. Evolution of the Reflected Efficiencies and Total Diffracted Energy as Functions of  $M$  for Grating Configuration 3 and TE and TM Polarizations**

$M$	$Q$	$e_0^r$	$e_1^r$	$e_2^r$	$\sum e_p$
<b>TE</b>					
2	150	—	—	—	1.563422751880
3	100	—	—	—	18.274452695896
4	75	0.448690625723	$0.431613129733 \times 10^{-1}$	$0.984271226609 \times 10^{-2}$	0.554698844748
5	60	0.448690737488	$0.431614911395 \times 10^{-1}$	$0.984264519365 \times 10^{-2}$	0.554699010109
6	50	0.448690737336	$0.431614911333 \times 10^{-1}$	$0.984264517625 \times 10^{-2}$	0.554699009914
10	30	0.448690737306	$0.431614911298 \times 10^{-1}$	$0.984264518693 \times 10^{-2}$	0.554699009939
12	25	0.448690737305	$0.431614911296 \times 10^{-1}$	$0.984264518692 \times 10^{-2}$	0.554699009938
<b>TM</b>					
2	150	—	—	—	2.997285452595
3	100	—	—	—	2.321194426985
4	75	0.162562615351	$0.719406380992 \times 10^{-1}$	$0.337719255117 \times 10^{-2}$	0.313198245250
5	60	0.162562665636	$0.719406028677 \times 10^{-1}$	$0.337718029851 \times 10^{-2}$	0.313198232007
6	50	0.162562665702	$0.719406028518 \times 10^{-1}$	$0.337718029506 \times 10^{-2}$	0.313198231998
10	30	0.162562665698	$0.719406028519 \times 10^{-1}$	$0.337718029457 \times 10^{-2}$	0.313198231991
12	25	0.162562665698	$0.719406028519 \times 10^{-1}$	$0.337718029457 \times 10^{-2}$	0.313198231991

tures such as a stack of two or more dielectric gratings with the same period but different groove shapes (Fig. 2). Such a device consists of several modulated regions separated by homogeneous ones. We divide the modulated regions into  $M_1, M_2, M_3 \dots$  slices in order to apply the  $R$ -matrix propagation method. Concerning the homogeneous slices (e.g., the ones defined by  $y_{M_i} < y < y_{M_{i+1}}$ ), there is no need to divide them, and Appendix B shows how the  $r$  or the  $r'$  matrices can be simply derived from the Rayleigh expansions of the field and of its normal derivative. Thus in the  $R$ -matrix propagation algorithm each homogeneous region, regardless of its thickness, plays the role of a simple slice, and the algorithm can be started from the bottom of the stack and continued as far as the top.

Of course, in the above-mentioned stack each modulated region could be a grating covered by a thin layer of a lossless or lossy dielectric grating with a thickness smaller than the groove depth. Such a dielectric-coated grating can be treated in a straightforward manner by the differential method, at the cost of a slight change in the Fourier coefficients  $k_n^2$  of function  $k^2(x, y)$ . The trivial case of a single dielectric-coated grating can then be studied with the  $R$ -matrix algorithm along the same lines as the study of a bare grating.

The method can be extended to the case of a stack of gratings with different periods, with the restrictions that the coarser grating have a period that is a multiple of the other periods and that its period define the periodicity of the entire stack. Such stacks of two superimposed dielectric gratings have previously been used<sup>19,20</sup> as grating interferometers in high-precision measurements. The possibility of producing fine-pitch gratings by means of photolithography opens new potential applications for such devices. They are at present the subject of large numerical studies in the framework of a Basic Research in Industrial Technologies-European Research on Advanced Materials (BRITE-EURAM) European project, "Flat Optical Antennas," the conclusions from which will be published in a future paper.

## 5. NUMERICAL RESULTS

### A. Checking the $R$ -Matrix Propagation Algorithm

In order to check the validity of the new method, we choose three different grating configurations. The first one consists of a deep lamellar grating with  $d = 1 \mu\text{m} = h$ ,  $\lambda = 0.365 \mu\text{m}$ ,  $\nu_1 = 1.536 = \nu'$ ,  $\nu_2 = 1$ , and  $\nu = 2.3$ , with a  $0.5\text{-}\mu\text{m}$  groove width illuminated under normal incidence. This particular groove shape leads to fast computations, because all the slices are identical, and thus the  $t^{(j)}$  matrices must be computed for one slice only. Table 1 shows the transmitted efficiencies  $e_p^t$  in the zero, first, and second orders and the total diffracted energy ( $\sum e_p$ ) for different values of  $M$  and different numbers of integration steps  $Q$ , varied in such a way that  $MQ = 300$ . It can be shown that the value of  $M$  does not matter too much, as long as it is high enough to avoid divergence of the results. Accuracy is indeed linked with the value of  $MQ$ ; the results are the same as high as the tenth digit, when  $M$  is varied, with  $MQ$  kept constant. The same conclusions apply to both TE and TM polarization.

Similar conclusions are reached from Table 2 for configuration 2, which we derived from configuration 1 by changing  $\nu$  from 2.3 to the complex refractive index of chromium ( $1.53 + i3.21$ ). No loss of accuracy was introduced by the metal losses, which absorb between 30% and 50% of the energy. Configuration 3, related to a full chromium lamellar grating ( $\nu_1 = 1 = \nu'$ ;  $\nu_2 = \nu = 1.53 + i3.21$ ), led to the results shown in Table 3, which again show an excellent convergence of the numerical results of the truncated differential set of equations. Let us point out that we carried out all calculations for the three different configurations while keeping 21 Fourier coefficients for representing the field (from  $-10$  to  $+10$ ). The question of whether the truncated Fourier series approaches the real field sufficiently will be addressed below, but Tables 1-3 already show that for the deep gratings considered here ( $h/d = 1$ ), supporting several diffracted orders ( $\lambda/d = 0.365$ ), the overflow linked with increasing exponential functions as well as other numerical instabilities have been removed by the  $R$ -matrix propagation algorithm.

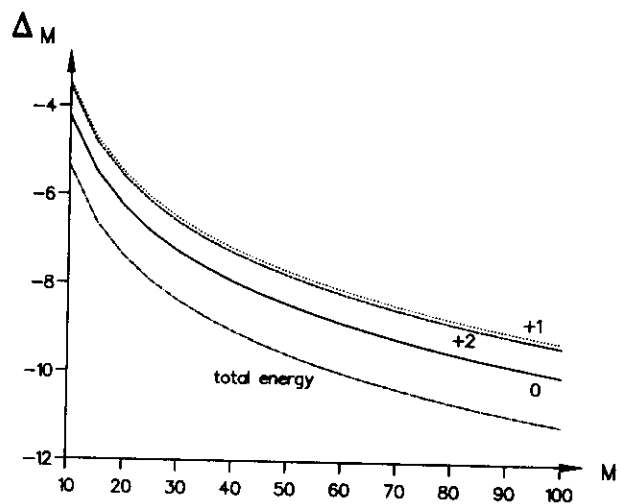


Fig. 3. Evolution of the accuracy  $\Delta_M$  on 0, +1, and +2 reflected efficiencies and on the total diffracted energy as functions of number of slices, for grating configuration 3 and TE polarization.

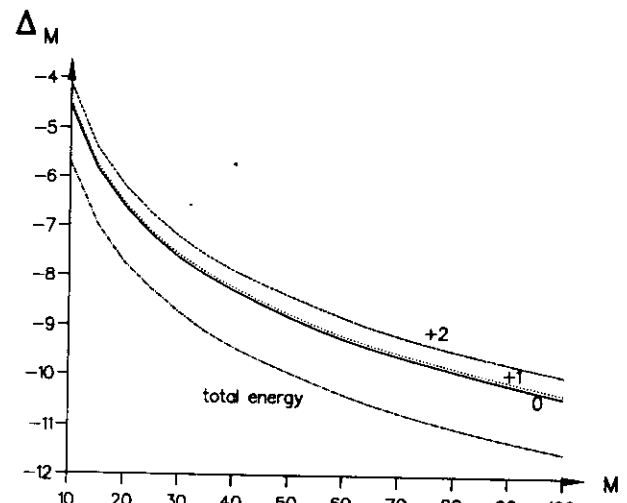


Fig. 4. Same as for Fig. 3 but for TM polarization.



**Table 4. Convergence of the 0- and -1-Order Efficiencies and Total Diffracted Energy of a Symmetrical Lamellar Chromium Grating as  $N$  Increases, with TM Polarization and Littrow Mount<sup>a</sup>**

$N$	$e_0^t$	$e_{-1}^r$	$\sum e_n^r$
11	0.10261	0.17314	0.27575
21	0.11594	0.21421	0.33015
31	0.12048	0.23379	0.35427
41	0.11349	0.24987	0.36336
51	0.11762	0.25646	0.37409
61	0.11259	0.26200	0.37460
71	0.11701	0.26528	0.38230
81	0.11298	0.26802	0.38101
91	0.11706	0.27007	0.38713
101	0.11367	0.27173	0.38540
111	0.11727	0.27316	0.39044
121	0.11436	0.27429	0.38865
141	0.11497	0.27618	0.39116
151	0.11772	0.27702	0.39475
161	0.11550	0.27766	0.39316
171	0.11790	0.27833	0.39624
181	0.11594	0.27883	0.39478

<sup>a</sup> $d = 0.4 \mu\text{m}$ ,  $h = 0.1 \mu\text{m}$ ,  $\lambda = 0.365 \mu\text{m}$ ,  $M = 10$ ,  $Q = 50$ .

In a second step we studied the evolution of the accuracy of the computation as function of product  $MQ$  by plotting criterion  $\Delta_M$  for different spectral orders, as well as for the total diffracted energy, as a function of  $M$  (with  $Q$  kept constant and equal only to 20). Figure 3 shows the results for configuration 3. An accuracy of  $10^{-8}$  is quickly obtained on the total energy, as soon as  $M = 20$ , whereas it is necessary to double the value of  $M$  to obtain similar accuracy on each diffracted efficiency. No divergence of the results occurs even when a stratification number  $M$  as high as 100 is used. These conclusions apply to both TE (Fig. 3) and TM (Fig. 4) polarization.

When the stability of the method has been established, the problem that remains is the convergence of the numerical results when the number  $N$  of Fourier coefficients is increased. For TE polarization numerical tests on the deep gratings of configurations 1–3 show that criterion  $\Delta_N$  applied to the 0, +1, and +2 diffracted efficiencies becomes less than  $-3$  when  $N$  becomes close to 61. For TM polarization the convergence is slower as  $N$  is increased; and for deep metallic gratings, oscillations can be observed in the values of  $\Delta_N$ . However, Table 4

shows that, for current groove depths produced by grating manufacturers, a reasonable accuracy of  $10^{-2}$  in diffracted efficiencies is obtained for  $N \approx 81$ . An even faster convergence is obtained for the rod grating in Table 5.

The last step in checking the new method was to compare its predictions with those obtained with other methods. In the range of validity of the previously developed differential formalism,<sup>15</sup> we first checked that the new method leads to the same results as the previous one; the discrepancies occurred on the fourth or fifth digit. Outside this range of validity, we compared the new results with those obtained with the integral method. For the example studied in Table 4, the integral method gave a difference of a unit on the second digit compared with the results obtained through our method when  $N$  reached 91.

### B. Examples of Applications

This subsection gives examples of grating computations that cannot be performed without the  $R$ -matrix propagation algorithm. We choose a 2000-groove/mm gold grating in a Littrow mount. The grooves have a symmetrical triangular shape, and the grating is used with TE polarization at  $0.6\text{-}\mu\text{m}$  wavelength for which the refractive index of gold is  $0.2 + i2.897$ . Figure 5 shows the evolutions of the -1- and 0-order reflected efficiencies as functions of groove depth  $h$ , as high as  $h/d = 10$ . For the highest values of the groove depth, one performs the computations by taking  $M = 20$ ,  $Q = 25$ , and  $N = 11$ , and no instability occurs. Such a curve could never have been produced by use of the classical differential method only. The second example deals with a stack of two sinusoidal gratings, as illustrated in Fig. 2, i.e., a corrugated waveguide with corrugations on both sides. The groove spacing is  $1 \mu\text{m}$ , the groove depth is  $0.3 \mu\text{m}$ , the wavelength is  $0.8 \mu\text{m}$ , and incidence is  $30^\circ$ . The substrate is silver ( $\nu_2 = 0.09 + i5.45$ ), the superstrate is air, and the dielectric layer has a refractive index of 1.5. Figure 6 shows the evolution of the 0-order efficiency as function of the thickness  $e$  of the homogeneous region between the two profiles (i.e., when  $e = 0$ , the silver grating is already coated with a  $0.3\text{-}\mu\text{m}$ -thick layer of dielectric). After a very narrow region ( $e \approx 0$ ) in which no guided wave can propagate and thus in which the 0-order reflectivity remains close to the reflectivity of silver, many strong and thin anomalies can be found. The curve in

**Table 5. Evolution of the Transmitted and Reflected Efficiencies of a Grating Made with Rectangular Chromium Rods as Functions of  $N$ <sup>a</sup>**

$N$	$e_0^t$	$e_{-1}^t$	$e_0^r$	$e_{-1}^r$	$\sum e_n$
81	0.24313605	0.13103090	0.16930986	0.11910124	0.66257807
91	0.24034564	0.13237731	0.17370030	0.11934494	0.66576826
101	0.24543773	0.13326091	0.17086301	0.11949403	0.66905568
111	0.24341668	0.13417700	0.17361233	0.11966288	0.67086890
121	0.24703351	0.13479741	0.17184844	0.11978479	0.67346417
131	0.24562508	0.13546833	0.17360525	0.11990952	0.67460819
141	0.24823707	0.13593543	0.17252051	0.12000920	0.67670222
151	0.24728619	0.13645036	0.17364343	0.12010446	0.67748446
161	0.24918965	0.13681777	0.17300384	0.12018610	0.67919739
171	0.24857610	0.13722602	0.17370573	0.12026052	0.67976839
181	0.24996719	0.13752376	0.17336586	0.12032774	0.68118456

<sup>a</sup>Same as Table 4, but here the substrate is vacuum.

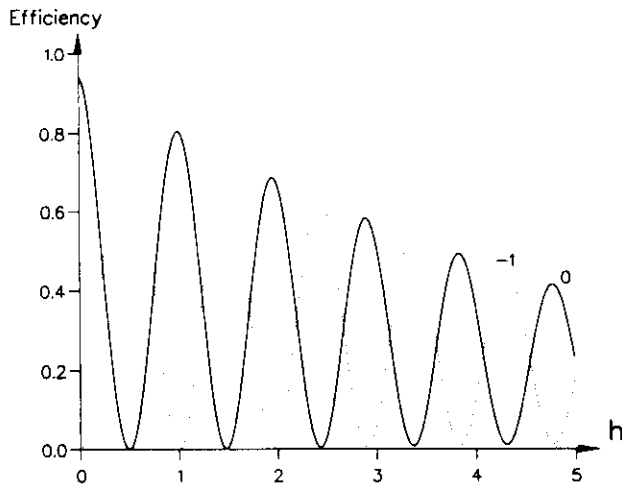


Fig. 5. Evolutions of the  $-1$ - and  $0$ -order efficiencies as functions of the groove depth for a symmetric triangular-profile gold grating with Littrow mount, for TE polarization.

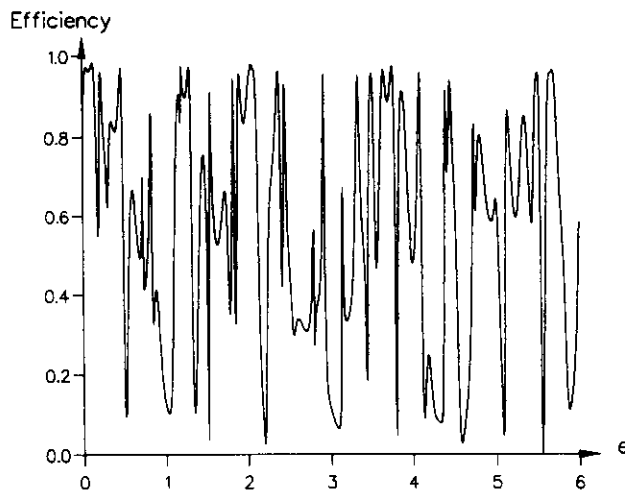


Fig. 6.  $0$ -Order efficiency of a dielectric coated sinusoidal silver grating as a function of dielectric thickness ( $M = 10$ ,  $N = 6$ ,  $Q = 20$ ), for TE polarization.

Fig. 6, which may cast doubt into the minds of many physicists about the validity of the numerical results, has indeed been fully confirmed for small values of  $e$  by the previously developed differential method.<sup>15</sup> It will not surprise people in the field of optics who are acquainted with multilayer gratings, who know that thick dielectric coatings introduce many sharp anomalies that are related to the resonant excitation of guided modes. The multiplicity of modes, along with the multiplicity of ways of exciting each of them through grating periodicity, produces the spectacular behavior shown in Fig. 6. With the  $R$ -matrix propagation algorithm, one could continue the curve far above  $e = 6 \mu\text{m}$  without encountering numerical instabilities.

## 6. CONCLUSION

This introduction to the  $R$ -matrix propagation algorithm considerably enlarges the range of application of existing theories with respect to the groove depth and the total

thickness of the diffracting device. The  $R$ -matrix algorithm allows us to study not only very deep gratings, i.e., with groove depth equal to several times the groove spacing, but also stacks of superimposed dielectric gratings without encountering numerical instabilities and overflows. Its property of removing undesired increasing exponential functions makes it useful for studying the problem of échelle gratings in visible or infrared regions as well as the problem of x-ray multilayer gratings used in orders as high as the 60th one. The latter topic is the subject of study for a future paper.

## APPENDIX A: EXPRESSIONS OF THE $R^{(j)}$ THE $r^{(j)}$ MATRICES

We start from Eq. (3) above:

$$\begin{bmatrix} U_m(y_j) \\ V_m(y_j) \end{bmatrix} = \begin{bmatrix} t_{11}^{(j)} & t_{12}^{(j)} \\ t_{21}^{(j)} & t_{22}^{(j)} \end{bmatrix} \begin{bmatrix} U_m(y_{j-1}) \\ V_m(y_{j-1}) \end{bmatrix}, \quad (\text{A1})$$

and we want to compute matrix  $r^{(j)}$  given by

$$\begin{bmatrix} U_m(y_{j-1}) \\ V_m(y_j) \end{bmatrix} = \begin{bmatrix} r_{11}^{(j)} & r_{12}^{(j)} \\ r_{21}^{(j)} & r_{22}^{(j)} \end{bmatrix} \begin{bmatrix} V_m(y_{j-1}) \\ U_m(y_j) \end{bmatrix}. \quad (\text{A2})$$

Equation (A1) gives

$$\begin{aligned} U_m(y_j) &= t_{11}^{(j)} U_m(y_{j-1}) + t_{12}^{(j)} V_m(y_{j-1}), \\ V_m(y_j) &= t_{21}^{(j)} U_m(y_{j-1}) + t_{22}^{(j)} V_m(y_{j-1}), \end{aligned} \quad (\text{A3})$$

from which we deduce

$$\begin{aligned} U_m(y_{j-1}) &= [t_{11}^{(j)}]^{-1} U_m(y_j) - [t_{11}^{(j)}]^{-1} t_{12}^{(j)} V_m(y_{j-1}), \\ V_m(y_j) &= t_{21}^{(j)} [t_{11}^{(j)}]^{-1} U_m(y_j) \\ &\quad + \{t_{22}^{(j)} - t_{21}^{(j)} [t_{11}^{(j)}]^{-1} t_{12}^{(j)}\} V_m(y_{j-1}). \end{aligned} \quad (\text{A4})$$

Comparison with Eq. (A2) gives

$$\begin{aligned} r_{11}^{(j)} &= -[t_{11}^{(j)}]^{-1} t_{12}^{(j)}, \\ r_{12}^{(j)} &= [t_{11}^{(j)}]^{-1}, \\ r_{21}^{(j)} &= t_{22}^{(j)} - t_{21}^{(j)} [t_{11}^{(j)}]^{-1} t_{12}^{(j)}, \\ r_{22}^{(j)} &= t_{21}^{(j)} [t_{11}^{(j)}]^{-1}, \end{aligned} \quad (\text{A5})$$

which is identical to Eqs. (7').

In order to establish the  $R$ -matrix propagation algorithm, we start from Eqs. (A2) and (8') written for the  $j - 1$  slice:

$$\begin{bmatrix} U_m(0) \\ V_m(y_{j-1}) \end{bmatrix} = \begin{bmatrix} R_{11}^{(j-1)} & R_{12}^{(j-1)} \\ R_{21}^{(j-1)} & R_{22}^{(j-1)} \end{bmatrix} \begin{bmatrix} V_m(0) \\ U_m(y_{j-1}) \end{bmatrix}, \quad (\text{A6})$$

and we want to compute matrix  $R^{(j)}$  defined by

$$\begin{bmatrix} U_m(0) \\ V_m(y_j) \end{bmatrix} = \begin{bmatrix} R_{11}^{(j)} & R_{12}^{(j)} \\ R_{21}^{(j)} & R_{22}^{(j)} \end{bmatrix} \begin{bmatrix} V_m(0) \\ U_m(y_j) \end{bmatrix}. \quad (\text{A7})$$

From Eqs. (A2) and (A6) we get

$$\begin{aligned} U_m(y_{j-1}) &= r_{11}^{(j)} V_m(y_{j-1}) + r_{12}^{(j)} U_m(y_j), \\ V_m(y_{j-1}) &= R_{21}^{(j-1)} V_m(0) + R_{22}^{(j-1)} U_m(y_{j-1}), \end{aligned} \quad (\text{A8})$$

from which we obtain

$$\begin{aligned} [\mathbf{I} - r_{11}^{(j)} R_{22}^{(j-1)}] U_m(y_{j-1}) &= r_{12}^{(j)} U_m(y_j) \\ &\quad + r_{11}^{(j)} R_{21}^{(j-1)} V_m(0), \\ [\mathbf{I} - R_{22}^{(j-1)} r_{11}^{(j)}] V_m(y_{j-1}) &= R_{21}^{(j-1)} V_m(0) \\ &\quad + R_{22}^{(j-1)} r_{12}^{(j)} U_m(y_j). \end{aligned} \quad (\text{A9})$$

Equations (A9) will be simplified by introduction of  $Z'$  and  $Y'$  given by

$$\begin{aligned} Z^{(j)} &= [\mathbf{I} - r_{11}^{(j)} R_{22}^{(j-1)}]^{-1}, \\ Y^{(j)} &= [\mathbf{I} - R_{22}^{(j-1)} r_{11}^{(j)}]^{-1}. \end{aligned} \quad (\text{A10})$$

Similarly, the other two terms from Eqs. (A2) and (A6) give

$$\begin{aligned} U_m(0) &= R_{11}^{(j-1)} V_m(0) + R_{12}^{(j-1)} U_m(y_{j-1}), \\ V_m(y_j) &= r_{21}^{(j)} V_m(y_{j-1}) + r_{22}^{(j)} U_m(y_j), \end{aligned} \quad (\text{A11})$$

from which we derive equations similar to Eqs. (A9). Finally, the use of Eqs. (A9)–(A11) leads to

$$\begin{aligned} U_m(0) &= R_{11}^{(j-1)} V_m(0) + R_{12}^{(j-1)} Z^{(j)} r_{12}^{(j)} U_m(y_j) \\ &\quad + R_{12}^{(j-1)} Z^{(j)} r_{11}^{(j)} R_{21}^{(j-1)} V_m(0), \\ V_m(y_j) &= r_{21}^{(j)} Y^{(j)} R_{21}^{(j-1)} V_m(0) \\ &\quad + r_{21}^{(j)} Y^{(j)} R_{22}^{(j-1)} r_{12}^{(j)} U_m(y_j) + r_{22}^{(j)} U_m(y_j), \end{aligned}$$

from which we deduce that

$$\begin{aligned} R_{11}^{(j)} &= R_{11}^{(j-1)} + R_{12}^{(j-1)} Z^{(j)} r_{11}^{(j)} R_{21}^{(j-1)}, \\ R_{12}^{(j)} &= R_{12}^{(j-1)} Z^{(j)} r_{12}^{(j)}, \\ R_{21}^{(j)} &= r_{21}^{(j)} Y^{(j)} R_{21}^{(j-1)}, \\ R_{22}^{(j)} &= r_{22}^{(j)} + r_{21}^{(j)} Y^{(j)} R_{22}^{(j-1)} r_{12}^{(j)}, \end{aligned}$$

which is the  $R'$ -matrix propagation algorithm.

## APPENDIX B: EXPRESSIONS OF THE $r$ AND THE $r'$ MATRICES FOR A DIELECTRIC SLAB

Let us consider a homogeneous slab having  $y_j - y_{j-1}$  thickness that is filled with a dielectric with refractive index  $\nu$ . Inside the slab the field is expressed by the following Rayleigh expansion:

$$U(x, y) = \sum_m [A_m \exp(-i\beta_m y) + B_m \exp(i\beta_m y)] \exp(i\alpha_m x), \quad (\text{B1})$$

where  $\beta_m^2 = [(2\pi/\lambda)\nu]^2 - \alpha_m^2$ , with use of the definitions given in Section 2. So we can write

$$\begin{aligned} U_m(y_j) &= A_m \exp(-i\beta_m y_j) + B_m \exp(i\beta_m y_j), \\ V_m(y_j) &= -i\chi\beta_m [A_m \exp(-i\beta_m y_j) - B_m \exp(i\beta_m y_j)], \end{aligned} \quad (\text{B2})$$

with  $\chi = 1$  for TE polarization and  $\chi = (\lambda/2\pi\nu)^2$  for TM polarization.

From the expressions  $V_m(y_j)$  and  $V_m(y_{j-1})$  we derive

$$\begin{aligned} A_m &= \frac{1}{i\chi\beta_m} \frac{1}{2i \sin[\beta_m(y_j - y_{j-1})]} [V_m(y_j) \exp(i\beta_m y_{j-1}) \\ &\quad - V_m(y_{j-1}) \exp(i\beta_m y_j)], \\ B_m &= \frac{1}{i\chi\beta_m} \frac{1}{2i \sin[\beta_m(y_j - y_{j-1})]} [V_m(y_j) \exp(-i\beta_m y_{j-1}) \\ &\quad - V_m(y_{j-1}) \exp(-i\beta_m y_j)]. \end{aligned} \quad (\text{B4})$$

Introducing Eqs. (B4) into Eq. (B2), we obtain

$$\begin{aligned} U_m(y_j) &= -\frac{1}{2\chi\beta_m} \left\{ \frac{1}{\sin[\beta_m(y_j - y_{j-1})]} \right\} (-2V_m(y_{j-1}) \\ &\quad + \{\exp[i\beta_m(y_{j-1} - y_j)] \\ &\quad + \exp[i\beta_m(y_j - y_{j-1})]\} V_m(y_j)), \\ U_m(y_{j-1}) &= -\frac{1}{2\chi\beta_m} \left\{ \frac{1}{\sin[\beta_m(y_j - y_{j-1})]} \right\} \\ &\quad \times \{(-\exp[i\beta_m(y_j - y_{j-1})] \\ &\quad - \exp[-i\beta_m(y_j - y_{j-1})]) V_m(y_{j-1}) \\ &\quad + 2V_m(y_j)\}. \end{aligned} \quad (\text{B5})$$

From these equations we get

$$\begin{aligned} r_{11}^{(j)} &= \left[ \frac{1}{\chi\beta_m} \frac{1}{\tan(\beta_m h_j)} \right], \\ r_{12}^{(j)} &= \left[ -\frac{1}{\chi\beta_m} \frac{1}{\sin(\beta_m h_j)} \right], \\ r_{21}^{(j)} &= \left[ \frac{1}{\chi\beta_m} \frac{1}{\sin(\beta_m h_j)} \right], \\ r_{22}^{(j)} &= \left[ -\frac{1}{\chi\beta_m} \frac{1}{\tan(\beta_m h_j)} \right], \end{aligned} \quad (\text{B6})$$

where  $h_j = y_j - y_{j-1}$ .

A similar calculation leads to

$$\begin{aligned} r_{11}^{(j')} &= \left[ \frac{i}{\chi\beta_m} \tan(\beta_m h_j) \right], \\ r_{12}^{(j')} &= \left[ \frac{1}{\cos(\beta_m h_j)} \right], \\ r_{21}^{(j')} &= \left[ \frac{1}{\cos(\beta_m h_j)} \right], \\ r_{22}^{(j')} &= [i\chi\beta_m \tan(\beta_m h_j)]. \end{aligned} \quad (\text{B7})$$

## REFERENCES

1. M. T. Gale, K. Knop, and R. Morf, "Zero order diffractive microstructure for security applications," presented at the Optical Security and Anticounterfeiting Systems Conference, Los Angeles, Calif., January 15-16, 1990.
2. E. G. Loewen, M. Nevière, and D. Maystre, "Grating efficiency theory as it applies to blazed and holographic gratings," *Appl. Opt.* **16**, 2711-2721 (1977).
3. R. Petit, ed., *Electromagnetic Theory of Gratings* (Springer-Verlag, Berlin, 1980).
4. B. Vidal, P. Vincent, P. Dhez, and M. Nevière, "Thin films and gratings theories used to optimize the high reflectivity of mirrors and gratings for x-ray optics, in *Applications of Thin Film Multilayered Structures to Figured X-Ray Optics*, G. F. Marshall, ed., *Proc. Soc. Photo-Opt. Instrum. Eng.* **563**, 142-149 (1985).
5. M. Nevière, "Bragg-Fresnel multilayer gratings: electromagnetic theory," *J. Opt. Soc. Am. A* **11**, 1835-1845 (1994).
6. F. Abelès, "Recherches sur la propagation des ondes électromagnétiques sinusoidales dans les milieux stratifiés. Application aux couches minces," *Ann. Phys. (Paris)* **5**, 596-640 and 706-782 (1950).
7. D. M. Pai and K. A. Awada, "Analysis of dielectric gratings of arbitrary profiles and thicknesses," *J. Opt. Soc. Am. A* **8**, 755-762 (1991).
8. D. J. Zvijac and J. C. Light, "R-matrix theory for collinear chemical reactions," *Chem. Phys.* **12**, 237-251 (1976).
9. J. C. Light and R. B. Walker, "An R-matrix approach to the solution of coupled equations for atom-molecule reactive scattering," *J. Chem. Phys.* **65**, 4272-4282 (1976).
10. L. F. DeSandre and J. M. Elson, "Extinction theorem analysis of diffraction anomalies in overcoated gratings," *J. Opt. Soc. Am. A* **8**, 763-777 (1991).
11. L. Li, "Multilayer modal method for diffraction gratings of arbitrary profile, depth, and permittivity," *J. Opt. Soc. Am. A* **10**, 2581-2593.
12. L. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams, and J. R. Andrewartha, "The dielectric lamellar diffraction gratings," *Opt. Acta* **28**, 413-428 (1981).
13. L. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams, and J. R. Andrewartha, "The finitely conducting lamellar diffraction grating," *Opt. Acta* **28**, 1087-1102 (1981).
14. L. Li, "A modal analysis of lamellar diffraction gratings in conical mounting," *J. Mod. Opt.* **40**, 553-573 (1993).
15. M. Nevière, P. Vincent, and R. Petit, "Sur la théorie du réseau conducteur et ses applications à l'optique," *Nou. Rev. Opt.* **5**, 65-77 (1974).
16. M. Nevière and P. Vincent, "Differential theory of gratings: answer to an objection on its validity for TM polarization," *J. Opt. Soc. Am. B* **5**, 1522-1524 (1988).
17. P. Vincent, "Differential methods," in *Electromagnetic Theory of Gratings*, R. Petit, ed. (Springer-Verlag, Berlin, 1980), pp. 101-121.
18. P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations* (Wiley, New York, 1962).
19. H. Iwaoka and K. Akiyama, "A high-resolution laser scale interferometer," in *Application, Theory, and Fabrication of Periodic Structures, Diffraction Gratings, and Moiré Phenomena II*, J. M. Lerner, ed., *Proc. Soc. Photo-Opt. Instrum. Eng.* **503**, 135-139 (1984).
20. A. Teimel, "Technology and application of grating interferometers in high-precision measurements," in *Progress in Precision Engineering*, P. Seysried, H. Kunzmann, and T. McKeown, eds. (Springer-Verlag, Berlin, 1991), pp. 131-147.

# Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings

Lifeng Li

Optical Sciences Center, University of Arizona, Tucson, Arizona 85721

Received July 20, 1995; revised manuscript received November 13, 1995; accepted December 4, 1995

Two recursive and numerically stable matrix algorithms for modeling layered diffraction gratings, the *S*-matrix algorithm and the *R*-matrix algorithm, are systematically presented in a form that is independent of the underlying grating models, geometries, and mountings. Many implementation variants of the algorithms are also presented. Their physical interpretations are given, and their numerical stabilities and efficiencies are discussed in detail. The single most important criterion for achieving unconditional numerical stability with both algorithms is to avoid the exponentially growing functions in every step of the matrix recursion. From the viewpoint of numerical efficiency, the *S*-matrix algorithm is generally preferred to the *R*-matrix algorithm, but exceptional cases are noted. © 1996 Optical Society of America

## 1. INTRODUCTION

As research in the field of diffraction gratings advances and the range of grating applications widens, the structures of gratings become more complicated than before. One of many new types of gratings that are finding more applications is layered gratings. For example, multilayer thin films were deposited onto photoresist surface-relief gratings to make high-efficiency, all-dielectric reflection gratings,<sup>1</sup> and coating polycarbonate lamellar gratings with a layer of MgF<sub>2</sub> was proposed as a means of making broadband antireflection structures.<sup>2</sup> Perhaps the most extreme cases of layered gratings are the Bragg-Fresnel gratings for use in x-ray spectroscopy<sup>3</sup> and the photonic band-gap materials.<sup>4</sup> On the other hand, in some grating models even a grating that consists of a single periodically corrugated surface is treated numerically as a layered structure. In this paper the term layered gratings will be used broadly to refer to both physically and numerically layered periodic structures.

All numerical methods for analyzing layered gratings face a common difficulty associated with the exponential functions of the spatial variable in the direction perpendicular to the grating plane. This difficulty is indicative of many problems of wave propagation and scattering in layered systems, and it is exacerbated by the fact that accurate numerical analysis of gratings usually requires a large number of eigenmodes. Recently this numerical difficulty has been overcome by many authors.<sup>4-15</sup> First, Pai and Awada<sup>5</sup> presented a Bremmer series method, based on the modal analysis with Fourier expansions, for dielectric gratings of arbitrary profile and groove depth in TE polarization. At the same time, DeSandre and Elson<sup>6</sup> presented an extinction-theorem analysis of diffraction anomalies in multilayer-coated shallow gratings by using the *R*-matrix propagation algorithm. Later, Li<sup>7</sup> applied the *R*-matrix algorithm to the classical modal method and enabled the latter to treat gratings

of arbitrary profile, depth, and permittivity. Chateau and Hugonin<sup>8</sup> proposed an algorithm, with the coupled-wave method, to model surface relief and volume gratings made of lossless and lossy dielectric materials. Montiel and Nevière<sup>9</sup> applied the *R*-matrix algorithm to the differential method and thereby eliminated "the numerical instabilities that have plagued the differential theory in TM polarization during the past 20 years (Ref. 9, p. 3241). Recently Li<sup>10</sup> applied the *R*-matrix algorithm to the differential formalism of Chandezon *et al.* (the *C* method) and thus removed a formerly existing limitation of the *C* method. The same goal was later achieved by Cotter *et al.*<sup>11</sup> using a scattering-matrix approach (*S*-matrix algorithm). The *S*-matrix algorithm was also used by Maystre<sup>4</sup> in an electromagnetic study of photonic band gaps by the integral method. Additionally, Li<sup>12</sup> showed that under certain conditions the *S*-matrix algorithm (which, unfortunately, was referred to there as the *R*-matrix algorithm) and the Bremmer series algorithm are equivalent. Very recently Moharam *et al.*<sup>13</sup> presented another stable algorithm, which they call the enhanced transmission matrix approach. For references concerning the applications of the *S*-matrix and *R*-matrix algorithms to problems of wave propagation and scattering outside the field of diffraction gratings, the reader may consult the reference list in Ref. 12.

Now there exist many stable numerical algorithms and several variants of implementation, expressed with different terminologies and applied to different grating models. There are obvious similarities and subtle differences among these algorithms and their variants. Their advantages and disadvantages, as well as interrelationships, have not been addressed in the literature. The purpose of this paper is to provide a systematic and unified presentation of the *S*-matrix and *R*-matrix algorithms, independent of the underlying grating models (integral, differential, modal, etc.) being used and the incidence conditions (TE, TM, or conical mount), and to

compare these two algorithms in terms of their physical interpretations, numerical stabilities, and numerical efficiencies. Some results presented here have already appeared in the literature, but many intricate details are new.

The algorithmic structure of the  $S$ -matrix and  $R$ -matrix algorithms is recursive, and the matrix dimension in the recursion is independent of the number of layers. Meanwhile, there exist stable and nonrecursive algorithms, for example, those in Refs. 14 and 15. In these algorithms the field amplitudes in all layers are solved together from a large linear system of equations whose matrix dimension is proportional to the number of the layers. The nonrecursive algorithms and the recursive algorithm of Moharam *et al.*,<sup>13</sup> which has a structure different from that of the  $S$ -matrix and  $R$ -matrix algorithms, are not considered in this paper.

In what follows, first the framework is laid down, in Section 2, for the development in the subsequent sections by defining the notation and the basis functions. The  $S$ -matrix algorithm and the  $R$ -matrix algorithm are presented in Sections 3 and 4, respectively. The presentations are arranged as parallel as possible for the two algorithms to bring out their similarities. Several variants of the two algorithms are then given in Section 5. In Section 6 the two algorithms are compared in terms of their numerical stabilities and efficiencies. Finally, in Section 7 some remarks are made that are specific to the applications of the algorithms to several grating models.

## 2. BACKGROUND FRAMEWORK

### A. Layer Abstraction

Figure 1 depicts a multilayer surface-relief grating. We assume that the profiles of all medium interfaces have the same periodicity in the  $x$  direction and that they are invariant in the  $z$  direction. We say that two adjacent interfaces are separable if a line  $y = \text{constant}$  can be drawn between them without crossing either interface; otherwise, we say that they are nonseparable. Thus the bottom three interfaces in Fig. 1 are separable, and the top three are not.

The  $S$ -matrix and  $R$ -matrix algorithms are applicable to all grating models, but here the discussion will be restricted to the classical modal method, the C method, the coupled-wave method, and the differential method. When it is not necessary to make the distinction, the first three methods will be referred to collectively as the modal methods, because they all rely on finding eigenmodes of Maxwell's equations. The classical modal method and the coupled-wave method approximate a continuous profile by a stack of lamellar gratings, as illustrated in the triangular grating in Fig. 1. This numerical approximation effectively introduces a number of numerical layer interfaces. The differential method does not use the multilayer lamellar grating approximation, but for numerical purposes it decomposes a grating profile into thin horizontal slices, thus also creating numerical layer interfaces. If two adjacent medium interfaces have identical functional form and amplitude, the C method does not require any numerical layer interface; otherwise, one numerical interface may be needed between the two medium interfaces.<sup>16,17</sup>

We abstract a layered grating structure by a series of parallel straight lines, each representing a real or numerical, straight or curved interface, depending on the profile of the medium interface and the grating model being used [see Fig. 2(a)]. For example, suppose that for the layered grating shown in Fig. 1 we use the classical modal method to treat the rectangular profile, the same method with a three-layer approximation to treat the triangular profile, the differential method with a three-slice decomposition to treat the asymmetrical smooth profile, and the C method to treat the top three profiles. Then, in Fig. 2(a),  $n = 15$ , 2 for the rectangular profile, 4 for the triangular profile, 4 for the asymmetrical smooth profile, and 6 for the top three profiles. The permittivities in Fig. 2(a) may be either constants or periodic functions of  $x$ , depending on the spatial region and the grating model. Media 0 and  $n + 1$  are two semi-infinite homogeneous media. The dashed line in medium 0 is a numerical interface. It can be ar-

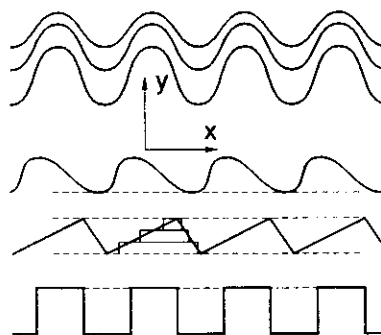


Fig. 1. General layered grating. All periodic medium interfaces share a common period, but otherwise they are arbitrary.

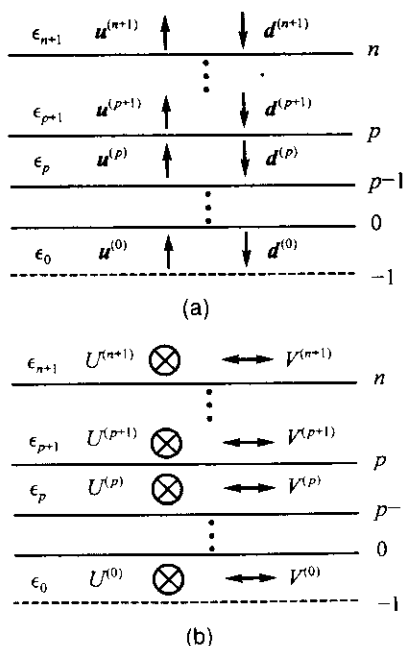


Fig. 2. Abstract layered grating structure, where the horizontal lines represent either actual material interfaces or numerical interfaces. The fields in each layer can be represented either (a) as a superposition of upward- and downward-propagating and decaying waves or (b) as a superposition of two sets of orthogonally polarized eigenmodes.

bitrarily close to interface 0. The thickness of layer  $p$  will be denoted by  $h_p$ .

### B. Basis Functions

When a modal method is used to analyze layer  $p$  in Fig. 2(a), the fields there are represented by superpositions of the eigenmodes. We assume that the eigenmodes in all layers share a common Floquet exponent, which is determined by the incident plane wave. The eigenvalue spectrum  $\sigma^{(p)}$ , having elements  $\lambda_m^{(p)}$ , can be partitioned such that  $\sigma^{(p)} = \sigma^{(p)+} \cup \sigma^{(p)-}$ , where

$$\sigma^{(p)-} = \{\lambda_m^{(p)-}; \quad \text{Re } \lambda_m^{(p)-} + \text{Im } \lambda_m^{(p)+} \geq 0, \quad \lambda_m^{(p)-} \in \sigma^{(p)+}\}. \quad (1)$$

In general, for any numerical truncation,  $\sigma^{(p)+}$  and  $\sigma^{(p)-}$  have the same number of elements. The dependence of the eigenmodes on  $y$ , i.e., the  $y$ -dependent basis functions, is given by  $\exp[i\lambda_m^{(p)\pm} y]$ , where  $\lambda_m^{(p)\pm} \in \sigma^{(p)\pm}$ . (Here  $y$  should be replaced by  $u$  if the C method is used; however, for simplicity we will ignore this minor difference.) Thus we call an eigenmode corresponding to  $\lambda_m^{(p)+}$  an up wave, and that corresponding to  $\lambda_m^{(p)-}$  a down wave. In particular, in the two semi-infinite regions and the homogeneous regions between the separable medium interfaces, the eigenmodes are simply the Rayleigh modes. In Fig. 2(a) the upward and downward arrows schematically represent the up wave and down waves, and the boldface letters  $\mathbf{u}$  and  $\mathbf{d}$  denote the column vectors whose elements are the wave amplitudes. Once the eigenmodes are determined everywhere, the grating problem reduces to a problem of determining the mode amplitudes.

Alternatively, we can choose  $\cos(\lambda_m^{(p)+} y)$  and  $\sin(\lambda_m^{(p)+} y)$  as the basis functions, which is always possible because  $\lambda_m^{(p)-} = -\lambda_m^{(p)+}$  with the classical modal method and the coupled-wave method and with the C method when the grating profile is symmetrical. We use  $U$  and  $V$  to denote the amplitudes of the modes that use this basis function set. The physical meaning of these amplitudes is clear: for example, if  $U$  is proportional to the  $z$  component of the electric field, then  $V$  is proportional to the  $x$  components of the magnetic field, as schematically shown in Fig. 2(b). From a mathematical point of view, the derivative of a  $U$  mode is a  $V$  mode, and vice versa. For the modal methods, both the exponential (or  $\mathbf{u}$ - $\mathbf{d}$ ) basis functions and the trigonometrical (or  $U$ - $V$ ) basis functions can be used.

In the differential method, one does not seek the eigen-solutions of Maxwell's equations. Instead, one numerically integrates  $U$  from one interface to another, where  $U$  is a column vector whose elements are the Fourier expansion coefficients of the  $z$  component of the electromagnetic fields. The numerical integration procedure gives the values of  $U$  and  $V = dU/dy$  as functions of  $y$ . Clearly, the  $U$ - $V$  basis functions described here correspond to the  $U$ - $V$  basis functions described in the preceding paragraph. Thus the schematic diagram in Fig. 2(b) applies to the differential method as well. It is possible to have a set of  $\mathbf{u}$ - $\mathbf{d}$  basis functions for the differential method if suitable linear combinations are made.<sup>9</sup> It is important to realize that, although the basis functions for the differential method do not have an explicit  $y$  dependence, their  $y$  dependence is asymptotically the same as that of the basis functions in the modal methods.

### C. Boundary Conditions

In this subsection we affix the equation numbers of all equations that apply only to the  $\mathbf{u}$ - $\mathbf{d}$  basis functions with a letter  $a$ , and those that apply only to the  $U$ - $V$  basis functions with a letter  $b$ . The same convention will be used for the formulas of the  $S$ -matrix and  $R$ -matrix algorithms in Sections 3, 4, and 5.

In the modal methods, when the boundary conditions are matched along interface  $p$ , we generally get an equation of form

$$W^{(p+1)} \begin{bmatrix} \mathbf{u}^{(p+1)}(y_p + 0) \\ \mathbf{d}^{(p+1)}(y_p + 0) \end{bmatrix} = W^{(p)} \begin{bmatrix} \mathbf{u}^{(p)}(y_p - 0) \\ \mathbf{d}^{(p)}(y_p - 0) \end{bmatrix}, \quad (2a)$$

where  $W^{(p)}$  and  $W^{(p+1)}$  are square matrices. Furthermore, by virtue of the modal fields,

$$\begin{bmatrix} \mathbf{u}^{(p)}(y_p - 0) \\ \mathbf{d}^{(p)}(y_p - 0) \end{bmatrix} = \phi^{(p)} \begin{bmatrix} \mathbf{u}^{(p)}(y_{p-1} + 0) \\ \mathbf{d}^{(p)}(y_{p-1} + 0) \end{bmatrix}, \quad (3a)$$

where

$$\phi^{(p)} = \begin{bmatrix} \exp(i\lambda_m^{(p)+} h_p) & 0 \\ 0 & \exp(i\lambda_m^{(p)-} h_p) \end{bmatrix}, \quad (4a)$$

and the exponential functions represent diagonal matrices (henceforth, all quantities with a subscript  $m$  represent diagonal matrices). Thus we obtain a recursive relation for the field amplitudes

$$\begin{bmatrix} \mathbf{u}^{(p+1)}(y_p + 0) \\ \mathbf{d}^{(p+1)}(y_p + 0) \end{bmatrix} = \tilde{t}^{(p)} \begin{bmatrix} \mathbf{u}^{(p)}(y_{p-1} + 0) \\ \mathbf{d}^{(p)}(y_{p-1} + 0) \end{bmatrix}, \quad (5a)$$

where

$$\tilde{t}^{(p)} = t^{(p)} \phi^{(p)}, \quad (6)$$

with

$$t^{(p)} = W^{(p+1)-1} W^{(p)}. \quad (7)$$

The matrices  $t^{(p)}$  and  $\tilde{t}^{(p)}$  can be fittingly called interface and layer  $t$  matrices, respectively. Note that  $t^{(p)}$  is of order  $O(1)$ .<sup>18</sup>

If the  $U$ - $V$  basis functions are used to match the boundary conditions, we have, correspondingly,

$$W^{(p+1)} \begin{bmatrix} U^{(p+1)}(y_p + 0) \\ V^{(p+1)}(y_p + 0) \end{bmatrix} = W^{(p)} \begin{bmatrix} U^{(p)}(y_p - 0) \\ V^{(p)}(y_p - 0) \end{bmatrix}, \quad (2b)$$

$$\begin{bmatrix} U^{(p)}(y_p - 0) \\ V^{(p)}(y_p - 0) \end{bmatrix} = \phi^{(p)} \begin{bmatrix} U^{(p)}(y_{p-1} + 0) \\ V^{(p)}(y_{p-1} + 0) \end{bmatrix}, \quad (3b)$$

$$\phi^{(p)} =$$

$$\begin{bmatrix} \cos(\lambda_m^{(p)} h_p) & \eta_m^{(p)} \sin(\lambda_m^{(p)} h_p) \\ -\eta_m^{(p)-1} \sin(\lambda_m^{(p)} h_p) & \cos(\lambda_m^{(p)} h_p) \end{bmatrix}, \quad (4b)$$

$$\begin{bmatrix} U^{(p+1)}(y_p + 0) \\ V^{(p+1)}(y_p + 0) \end{bmatrix} = \tilde{t}^{(p)} \begin{bmatrix} U^{(p)}(y_{p-1} + 0) \\ V^{(p)}(y_{p-1} + 0) \end{bmatrix}. \quad (5b)$$

In Eq. (4b),  $\eta_m^{(p)}$  is a constant independent of  $h_p$ , and for simplicity the plus sign has been dropped from the superscript of eigenvalue  $\lambda_m^{(p)+}$ . Also, the same notation

$W$ ,  $\phi$ , and  $t$  is used with the two different basis function sets. This, however, is not a problem because the context will tell to which basis the matrices are referring. From now on, the amplitude vectors without an explicit argument stand for the vectors that are evaluated at the lower bound of the layer. For example,  $\mathbf{u}^{(p)} \equiv \mathbf{u}^{(p)}(y_{p-1} + 0)$ .

At this point, it is most natural and mathematically simplest to proceed with solving the grating problem by the so-called  $T$ -matrix algorithm, which is obtained by repeated use of Eq. (5a) or Eq. (5b). However, it is well known that the  $T$ -matrix algorithm, with either of the basis function sets, is numerically unstable when the total layer thickness of the grating structure and the matrix dimension are large.<sup>7</sup> This numerical instability is generally attributed to the presence of the growing exponential functions in the algorithm. Fundamentally, the cause of instability is a classic one: loss of significant digits when one is computing a small number by subtracting two large numbers by a computer of finite precision. Symbolically, it is a case of  $\infty - \infty = O(1)$ . It should be emphasized that the numerical instability of the  $T$ -matrix algorithm cannot be eased or removed by simply reducing the individual layer thicknesses without lowering the total thickness, because the  $T$ -matrix algorithm accumulates the magnitudes of the exponential functions as the layer  $t$  matrices are multiplied together.

### 3. S-MATRIX ALGORITHM

The  $S$ -matrix algorithm uses the exponential basis functions. For any  $0 \leq p \leq n$ , it seeks a stack  $S$  matrix,  $S^{(p)}$ , that links the waves in layer  $p + 1$  and medium 0 in this way:

$$\begin{bmatrix} \mathbf{u}^{(p+1)} \\ \mathbf{d}^{(0)} \end{bmatrix} = S^{(p)} \begin{bmatrix} \mathbf{u}^{(0)} \\ \mathbf{d}^{(p+1)} \end{bmatrix}. \quad (8a)$$

Before moving on, it is important to describe the physical meaning of the  $S$  matrix. For this purpose we rewrite  $S^{(p)}$  in a two-by-two block form:

$$\begin{bmatrix} \mathbf{u}^{(p+1)} \\ \mathbf{d}^{(0)} \end{bmatrix} = \begin{bmatrix} T_{uu}^{(p)} & R_{ud}^{(p)} \\ R_{du}^{(p)} & T_{dd}^{(p)} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(0)} \\ \mathbf{d}^{(p+1)} \end{bmatrix}. \quad (9a)$$

The significance of the subscripts  $u$  and  $d$  becomes evident once the reader mentally carries out the matrix-vector multiplication on the right-hand side of the equation. As a rule in this paper, the use of subscripts  $u$  and  $d$  is an automatic indication that the submatrix belongs to a matrix in the  $S$ -matrix algorithm. The choice of letters  $R$  and  $T$ , instead of  $S$ , makes the physical meanings of the four submatrices of  $S^{(p)}$  self-explanatory. For example,  $T_{uu}^{(p)}$  and  $R_{ud}^{(p)}$  are the transmission matrix and reflection matrix that give the upward wave amplitudes in layer  $p + 1$  resulting from the transmission of the upward incident waves in medium 0 and from the reflection of the incident downward waves in layer  $p + 1$ , respectively, by the whole stack below layer  $p + 1$ . Alternatively, the first  $p$  layers of the layered grating can be viewed as a linear four-terminal network. Matrix  $S^{(p)}$  operates on the two sets of inputs to generate the two sets of outputs, as shown schematically in Fig. 3(a).

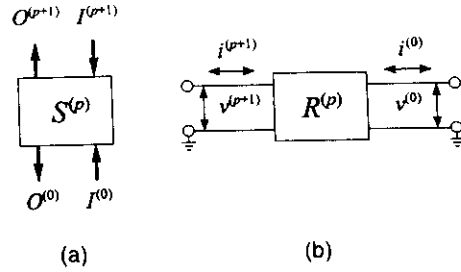


Fig. 3. Schematic diagrams for alternative interpretations of (a) the  $S$  matrix and (b) the  $R$  matrix. In Fig. 3(a),  $I$  and  $O$  stand for inputs to and outputs from the system represented by the square box. In Fig. 3(b),  $i$  and  $v$  stand for currents and voltages at the terminals of the circuit represented by the square box.

To link the waves in two adjacent layers, we can define an interface  $s$  matrix,  $s^{(p)}$ , and a layer  $s$  matrix,  $\tilde{s}^{(p)}$ , as follows:

$$\begin{bmatrix} \mathbf{u}^{(p+1)}(y_p + 0) \\ \mathbf{d}^{(p)}(y_p - 0) \end{bmatrix} = s^{(p)} \begin{bmatrix} \mathbf{u}^{(p)}(y_p - 0) \\ \mathbf{d}^{(p+1)}(y_p + 0) \end{bmatrix}, \quad (10a)$$

$$\begin{bmatrix} \mathbf{u}^{(p+1)} \\ \mathbf{d}^{(p)} \end{bmatrix} = \tilde{s}^{(p)} \begin{bmatrix} \mathbf{u}^{(p)} \\ \mathbf{d}^{(p+1)} \end{bmatrix}, \quad (11a)$$

or

$$\begin{bmatrix} \mathbf{u}^{(p+1)} \\ \mathbf{d}^{(p)} \end{bmatrix} = \begin{bmatrix} \tilde{t}_{uu}^{(p)} & \tilde{r}_{ud}^{(p)} \\ \tilde{r}_{du}^{(p)} & \tilde{t}_{dd}^{(p)} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(p)} \\ \mathbf{d}^{(p+1)} \end{bmatrix}. \quad (12a)$$

The physical interpretations of  $s^{(p)}$  and  $\tilde{s}^{(p)}$  are similar to that of  $S^{(p)}$ . Note that  $\tilde{t}_{uu}^{(p)}$  and  $\tilde{t}_{dd}^{(p)}$ , because of the notation of their subscripts, cannot be confused with the  $t$  matrix defined in Section 2. The layer  $s$  matrix is related to the interface  $s$  matrix by

$$\tilde{s}^{(p)} = \begin{bmatrix} 1 & 0 \\ 0 & \exp(-i\lambda_m^{(p)} h_p) \end{bmatrix} s^{(p)} \begin{bmatrix} \exp(i\lambda_m^{(p)} h_p) & 0 \\ 0 & 1 \end{bmatrix}, \quad (13a)$$

and the interface  $s$  matrix is in turn related to the interface  $t$  matrix by

$$s^{(p)} = \begin{bmatrix} t_{11}^{(p)} - t_{12}^{(p)} t_{22}^{(p)-1} t_{21}^{(p)} & t_{12}^{(p)} t_{22}^{(p)-1} \\ -t_{22}^{(p)-1} t_{21}^{(p)} & t_{22}^{(p)-1} \end{bmatrix}. \quad (14a)$$

Note that all four entries in Eq. (14a) contain the inverse of submatrix  $t_{22}^{(p)}$ . For this reason we call  $t_{22}^{(p)}$  the pivotal submatrix.

From Eqs. (8a) and (11a), the set of recursion formulas for the stack  $S$  matrix are

$$\begin{aligned} T_{uu}^{(p)} &= \tilde{t}_{uu}^{(p)} [1 - R_{ud}^{(p-1)} \tilde{r}_{du}^{(p-1)}]^{-1} T_{uu}^{(p-1)}, \\ R_{ud}^{(p)} &= \tilde{r}_{ud}^{(p)} + \tilde{t}_{uu}^{(p)} R_{ud}^{(p-1)} [1 - \tilde{r}_{du}^{(p-1)} R_{ud}^{(p-1)}]^{-1} \tilde{t}_{dd}^{(p)}, \\ R_{du}^{(p)} &= R_{du}^{(p-1)} + T_{dd}^{(p-1)} \tilde{r}_{du}^{(p-1)} [1 - R_{ud}^{(p-1)} \tilde{r}_{du}^{(p-1)}]^{-1} T_{uu}^{(p-1)}, \\ T_{dd}^{(p)} &= T_{dd}^{(p-1)} [1 - \tilde{r}_{du}^{(p-1)} R_{ud}^{(p-1)}]^{-1} \tilde{t}_{dd}^{(p)}. \end{aligned} \quad (15a)$$



The  $S$ -matrix recursion can be initialized by setting

$$S^{(-1)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (16a)$$

or, equivalently, by setting  $S^{(0)} = s^{(0)}$ .

The form of the factors enclosed by the square brackets in Eq. (15a) is such that the inverse matrices can be readily expanded, at least formally, into a geometrical series in terms of the product of the two reflection matrices. This fact naturally gives rise to the multiple-reflection interpretation of the  $S$ -matrix recursion formulas.<sup>12</sup> (It is quite unfortunate that in Ref. 12 the  $S$ -matrix algorithm was incorrectly called the  $R$ -matrix algorithm.) Because of the elegant form of the inverse matrices, Eqs. (15a) will be called the normalized  $S$ -matrix recursion formulas.

Equations (8a)–(16a) constitute the basic ingredients of the  $S$ -matrix algorithm. In most grating problems the quantities of interest are the field amplitudes leaving the grating structure in the two outer media, i.e.,  $\mathbf{u}^{(n-1)}$  and  $\mathbf{d}^{(0)}$ . They are simply given by

$$\begin{bmatrix} \mathbf{u}^{(n-1)} \\ \mathbf{d}^{(0)} \end{bmatrix} = \begin{bmatrix} T_{uu}^{(n)} & R_{ud}^{(n)} \\ R_{du}^{(n)} & T_{dd}^{(n)} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(0)} \\ \mathbf{d}^{(n-1)} \end{bmatrix}. \quad (17a)$$

In particular, if there are no incident waves in medium 0 ( $\mathbf{u}^{(0)} = 0$ ),

$$\begin{aligned} \mathbf{u}^{(n-1)} &= R_{ud}^{(n)} \mathbf{d}^{(n-1)}, \\ \mathbf{d}^{(0)} &= T_{dd}^{(n)} \mathbf{d}^{(n-1)}. \end{aligned} \quad (18a)$$

The numerical stability of the  $S$ -matrix algorithm is rooted in the construction of the layer  $s$  matrix. The problem-causing, growing exponential function,  $\exp(i\lambda_m^{(p)} h_p)$ , that was originally in the  $\tilde{t}$  matrix, is now inverted in Eq. (13a). Since  $s^{(p)}$  is of order  $O(1)$ , so is  $\tilde{s}^{(p)}$ . Furthermore, the submatrices of  $\tilde{s}^{(p)}$  appear in the recursion formulas only as additive or multiplicative terms. Thus the  $S$  matrices remain of order  $O(1)$ , and the numerical stability of the algorithm is ensured.

#### 4. R-MATRIX ALGORITHM

The  $R$ -matrix algorithm uses the trigonometrical basis functions. For any  $0 \leq p \leq n$ , it seeks a stack  $R$  matrix,  $R^{(p)}$ , that links the fields in layer  $p+1$  and medium 0 in this way:

$$\begin{bmatrix} U^{(p-1)} \\ U^{(0)} \end{bmatrix} = R^{(p)} \begin{bmatrix} V^{(p-1)} \\ V^{(0)} \end{bmatrix}. \quad (8b)$$

The  $R$  matrix can be physically interpreted as field impedance or field admittance (the ratio of the tangential component of the  $\mathbf{E}$  field to the tangential component of the  $\mathbf{H}$  field or its inverse). For example, in TE polarization, because  $U$  and  $V$  correspond to the  $\mathbf{E}$  and  $\mathbf{H}$  fields, respectively,  $R^{(p)}$  plays the role of field impedance. An alternative interpretation of Eq. (8b) can be made, with the aid of Fig. 3(b), in terms of currents and voltages in an electrical circuit. Here, if  $U$  is identified with the voltages and  $V$  with the currents, or vice versa, then  $R^{(p)}$  is the electrical impedance or admittance. The concept of impedance has been used previously in modeling grat-

ings that contain multiple planar interfaces but only one periodically modulated interface.<sup>19</sup>

To link the fields in two adjacent layers, we can define an interface  $r$  matrix,  $r^{(p)}$ , and a layer  $r$  matrix,  $\tilde{r}^{(p)}$ , as follows:

$$\begin{bmatrix} U^{(p-1)}(y_p + 0) \\ U^{(p)}(y_p - 0) \end{bmatrix} = r^{(p)} \begin{bmatrix} V^{(p+1)}(y_p + 0) \\ V^{(p)}(y_p - 0) \end{bmatrix}, \quad (10b)$$

$$\begin{bmatrix} U^{(p+1)} \\ U^{(p)} \end{bmatrix} = \tilde{r}^{(p)} \begin{bmatrix} V^{(p+1)} \\ V^{(p)} \end{bmatrix}. \quad (11b)$$

The layer  $r$  matrix is related to the interface  $r$  matrix by

$$\begin{aligned} \tilde{r}_{11}^{(p)} &= r_{11}^{(p)} - r_{12}^{(p)} \zeta^{(p)} r_{21}^{(p)}, \\ \tilde{r}_{12}^{(p)} &= r_{12}^{(p)} \zeta^{(p)} \eta_m^{(p)} \csc[\lambda_m^{(p)} h_p], \\ \tilde{r}_{21}^{(p)} &= \eta_m^{(p)} \csc[\lambda_m^{(p)} h_p] \zeta^{(p)} r_{21}^{(p)}, \\ \tilde{r}_{22}^{(p)} &= \eta_m^{(p)} \cot[\lambda_m^{(p)} h_p] - \eta_m^{(p)} \csc[\lambda_m^{(p)} h_p] \\ &\quad \times \zeta^{(p)} \eta_m^{(p)} \csc[\lambda_m^{(p)} h_p], \end{aligned} \quad (13b)$$

where

$$\zeta^{(p)} = [r_{22}^{(p)} + \eta_m^{(p)} \cot(\lambda_m^{(p)} h_p)]^{-1}. \quad (13b')$$

For a proof of Eq. (13b), see Appendix A. Here we have excluded the possibility that accidentally  $\lambda_m^{(p)} h_p = l\pi$ , where  $l$  is an integer. The interface  $r$  matrix is in turn related to the interface  $t$  matrix by

$$r^{(p)} = \begin{bmatrix} t_{11}^{(p)} t_{21}^{(p)-1} & t_{12}^{(p)} - t_{11}^{(p)} t_{21}^{(p)-1} t_{22}^{(p)} \\ t_{21}^{(p)-1} & -t_{21}^{(p)-1} t_{22}^{(p)} \end{bmatrix}. \quad (14b)$$

From Eqs. (13b) and (14b) it is clear that  $\tilde{r}^{(p)}$  is of order  $O(1)$ . Alternatively, we can use the layer  $t$  matrix to express the layer  $r$  matrix:

$$\tilde{r}^{(p)} = \begin{bmatrix} t_{11}^{(p)} t_{21}^{(p)-1} & t_{12}^{(p)} - t_{11}^{(p)} t_{21}^{(p)-1} t_{22}^{(p)} \\ t_{21}^{(p)-1} & -t_{21}^{(p)-1} t_{22}^{(p)} \end{bmatrix}. \quad (14b')$$

In Eqs. (14b) and (14b')  $t_{21}^{(p)}$  and  $t_{21}^{(p)}$  are the pivotal submatrices. In some cases it is possible that  $t_{21}^{(p)} \equiv 0$ , but then  $t_{21}^{(p)} \neq 0$ . In fact, such a case can be utilized beneficially (see Appendix B).

From Eqs. (8b) and (11b) the set of recursion formulas for the stack  $R$  matrix are

$$\begin{aligned} R_{11}^{(p)} &= \tilde{r}_{11}^{(p)} - \tilde{r}_{12}^{(p)} Z^{(p)} \tilde{r}_{21}^{(p)}, \\ R_{12}^{(p)} &= \tilde{r}_{12}^{(p)} Z^{(p)} R_{12}^{(p-1)}, \\ R_{21}^{(p)} &= -R_{21}^{(p-1)} Z^{(p)} \tilde{r}_{21}^{(p)}, \\ R_{22}^{(p)} &= R_{22}^{(p-1)} + R_{21}^{(p-1)} Z^{(p)} R_{12}^{(p-1)}, \end{aligned} \quad (15b)$$

where

$$Z^{(p)} = (\tilde{r}_{22}^{(p)} - R_{11}^{(p-1)})^{-1}. \quad (15b')$$

The  $R$ -matrix recursion can be initialized by setting

$$R^{(0)} = \tilde{r}^{(0)}. \quad (16b)$$

Unlike their counterparts in the  $S$ -matrix algorithm,

Eqs. (15b) do not readily subject themselves to an intuitive physical interpretation. Nonetheless, to preserve the formal symmetry between the two algorithms, we shall call Eqs. (15b) the normalized  $R$ -matrix recursion formulas.

Starting with Eq. (16b), repeated use of Eqs. (15b) until  $p = n$  leads to

$$\begin{bmatrix} U^{(n+1)} \\ U^{(0)} \end{bmatrix} = \begin{bmatrix} R_{11}^{(n)} & R_{12}^{(n)} \\ R_{21}^{(n)} & R_{22}^{(n)} \end{bmatrix} \begin{bmatrix} V^{(n+1)} \\ V^{(0)} \end{bmatrix}. \quad (17b)$$

Suppose that  $U^{(0)} = \mathbf{u}^{(0)} + \mathbf{d}^{(0)}$ ,  $V^{(0)} = \mathbf{u}^{(0)} - \mathbf{d}^{(0)}$ ,  $U^{(n+1)} = \mathbf{u}^{(n+1)} + \mathbf{d}^{(n+1)}$ , and  $V^{(n+1)} = \mathbf{u}^{(n+1)} - \mathbf{d}^{(n+1)}$ , which is always possible by definition. Then from Eq. (17b) we get the linear system that determines the out-going diffraction amplitudes in the top and bottom media:

$$\begin{bmatrix} 1 - R_{11}^{(n)} & R_{12}^{(n)} \\ -R_{21}^{(n)} & 1 + R_{22}^{(n)} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(n+1)} \\ \mathbf{d}^{(0)} \end{bmatrix} = \begin{bmatrix} -1 - R_{11}^{(n)} & R_{12}^{(n)} \\ -R_{21}^{(n)} & -1 + R_{22}^{(n)} \end{bmatrix} \begin{bmatrix} \mathbf{d}^{(n+1)} \\ \mathbf{u}^{(0)} \end{bmatrix}. \quad (18b)$$

The  $R$ -matrix algorithm is also immune to the numerical difficulties associated with growing exponential functions. This is because the submatrices of  $\tilde{r}^{(p)}$  are all of order  $O(1)$ , and they appear in the recursion formulas Eqs. (15b) and (15b') only as additive and multiplicative terms. The former fact is evident when Eq. (13b) is used to construct the layer  $r$  matrices. It is not so obvious if Eqs. (14b') are used, however; in fact, in this case the  $R$ -matrix algorithm is only conditionally stable.

It can be shown that Eq. (14b') and Eqs. (13b) are algebraically equivalent. Therefore the submatrix  $\tilde{r}_{12}^{(p)}$ , as given by Eq. (14b'), should be mathematically proportional to  $\csc(\lambda_m^{(p)} h_p)$ , which tends to 0 as  $m \rightarrow \infty$  and  $h_p \rightarrow \infty$ . On the other hand, the first term of  $\tilde{r}_{12}^{(p)}$  in Eq. (14b') is  $\tilde{t}_{12}^{(p)} = t_{11}^{(p)} \eta_m^{(p)} \sin[\lambda_m^{(p)} h_p] + t_{12}^{(p)} \cos[\lambda_m^{(p)} h_p]$ , which tends to  $\infty$ . Thus the second term of  $\tilde{r}_{12}^{(p)}$  must also tend to  $\infty$  as  $m \rightarrow \infty$  and  $h_p \rightarrow \infty$ . Clearly this mathematical arrangement presents a serious numerical problem. When the absolute values of the imaginary parts of  $\lambda_m^{(p)} h_p$  are large, the numerically calculated matrix elements of  $\tilde{r}_{12}^{(p)}$  by Eq. (14b') may not be small, as a result of round-off errors. Let  $\Lambda^{(p)}$  be the maximum of the absolute values of the imaginary parts of all eigenvalues for a given matrix truncation. As a rule of thumb, when  $\exp[\Lambda^{(p)} h_p] \sim 10^{15}$ , the numerical problem described above begins to arise (double precision is assumed here). To avoid the problem one has to choose the layer thickness so that  $\exp[\Lambda^{(p)} h_p] \ll 10^{15}$ . Therefore the  $R$ -matrix algorithm is conditionally stable when Eq. (14b') is used. Fortunately, unlike the  $T$ -matrix algorithm, here the magnitudes of the exponential functions do not accumulate. Therefore lowering the individual layer thickness is an effective remedy for the numerical instability caused by the use of Eq. (14b').

## 5. VARIANTS OF IMPLEMENTATION

### A. Variation in Matrix Manipulation

In Sections 3 and 4 the  $S$ -matrix and  $R$ -matrix algorithms were systematically presented. The presentation took

three steps: the definitions and derivations of the layer  $t$  matrices, the layer  $s$  (or  $r$ ) matrices, and the stack  $S$  (or  $R$ ) matrices. Although from a theoretical point of view the introduction of the layer  $t$  matrices and the layer  $s$  (or  $r$ ) matrices has made the presentation systematic, from a practical point of view the use of one of the two kinds of layer matrices can be eliminated, as demonstrated below.

The  $S$ -matrix recursion can be accomplished by use of the  $t$  matrices directly, without the layer  $s$  matrices. From Eqs. (5a), (6), and (9a) we can easily derive a set of nonnormalized  $S$ -matrix recursion formulas by using the interface  $t$  matrix:

$$\begin{aligned} T_{uu}^{(p)} &= [t_{11}^{(p)} - R_{ud}^{(p)} t_{21}^{(p)}] \phi_+^{(p)} T_{uu}^{(p-1)}, \\ R_{ud}^{(p)} &= [t_{12}^{(p)} + t_{11}^{(p)} \Omega^{(p)}] [t_{22}^{(p)} + t_{21}^{(p)} \Omega^{(p)}]^{-1}, \\ R_{du}^{(p)} &= R_{du}^{(p-1)} - T_{dd}^{(p)} t_{21}^{(p)} \phi_+^{(p)} T_{uu}^{(p-1)}, \\ T_{dd}^{(p)} &= T_{dd}^{(p-1)} \phi_-^{(p-1)} [t_{22}^{(p)} + t_{21}^{(p)} \Omega^{(p)}]^{-1}, \end{aligned} \quad (19a)$$

where

$$\Omega^{(p)} = \phi_+^{(p)} R_{ud}^{(p-1)} \phi_-^{(p-1)}, \quad (19a')$$

and  $\phi_{\pm}^{(p)}$  are the two diagonal submatrices in Eq. (4a). Of course, Eqs. (19a) and (15a) are algebraically equivalent. Note that the above equations have been written in terms of the interface  $t$  matrices, instead of the layer  $t$  matrices, and the appearance of  $\phi_{\pm}^{(p)}$  has been arranged properly so that there are no exponentially growing functions in the formulas. This measure avoids possible numerical overflow and ensures numerical stability of the  $S$ -matrix algorithm.

If we set  $\phi_{\pm}^{(p)} = 1$  in Eqs. (19a) and (19a') and replace all interface  $t$  submatrices by layer  $t$  submatrices, we obtain the nonnormalized  $S$ -matrix recursion formulas by using the layer  $t$  matrix:

$$\begin{aligned} T_{uu}^{(p)} &= [t_{11}^{(p)} - R_{ud}^{(p)} t_{21}^{(p)}] T_{uu}^{(p-1)}, \\ R_{ud}^{(p)} &= [t_{12}^{(p)} + t_{11}^{(p)} R_{ud}^{(p-1)}] [t_{22}^{(p)} + t_{21}^{(p)} R_{ud}^{(p-1)}]^{-1}, \\ R_{du}^{(p)} &= R_{du}^{(p-1)} - T_{dd}^{(p)} t_{21}^{(p)} T_{uu}^{(p-1)}, \\ T_{dd}^{(p)} &= T_{dd}^{(p-1)} [t_{22}^{(p)} + t_{21}^{(p)} R_{ud}^{(p-1)}]^{-1}. \end{aligned} \quad (20a)$$

The use of this set of recursion formulas should be avoided whenever possible, because the matrix to be inverted is a sum of an exponentially growing matrix and an exponentially decaying matrix.

The  $R$ -matrix recursion can also be accomplished with use of the  $t$  matrices directly, without the layer  $r$  matrices. From Eqs. (5b), (6), and (8b) we can easily derive a set of nonnormalized  $R$ -matrix recursion formulas by using the layer  $t$  matrix:

$$\begin{aligned} R_{11}^{(p)} &= [t_{12}^{(p)} + t_{11}^{(p)} R_{11}^{(p-1)}] [t_{22}^{(p)} + t_{21}^{(p)} R_{11}^{(p-1)}]^{-1}, \\ R_{12}^{(p)} &= [t_{11}^{(p)} - R_{11}^{(p)} t_{21}^{(p)}] R_{12}^{(p-1)}, \\ R_{21}^{(p)} &= R_{21}^{(p-1)} [t_{22}^{(p)} + t_{21}^{(p)} R_{11}^{(p-1)}]^{-1}, \\ R_{22}^{(p)} &= R_{22}^{(p-1)} - R_{21}^{(p)} t_{21}^{(p)} R_{12}^{(p-1)}. \end{aligned} \quad (20b)$$

Since Eqs. (20b) are algebraically equivalent to Eqs. (15b),

the  $R$  matrices obtained this way are mathematically of order  $O(1)$ . Numerically, however, devastating round-off errors could occur if the numerical layer thicknesses are set too high. The reason is the same as the one given at the end of Section 4 for the possible numerical instability resulting from the use of Eq. (14b'). Specifically, the expression of  $R_{12}^{(p)}$  in Eqs. (20b) is of type  $\infty - \infty = O(1)$ . Thus Eqs. (20b) also give a conditional stable implementation of the  $R$ -matrix algorithm. The unconditional

$$\begin{bmatrix} a_{uu} & a_{ud} \\ a_{du} & a_{dd} \end{bmatrix} * \begin{bmatrix} b_{uu} & b_{ud} \\ b_{du} & b_{dd} \end{bmatrix} = \begin{bmatrix} b_{uu}(1 - a_{ud}b_{du})^{-1}a_{uu} & b_{ud} + b_{uu}a_{ud}(1 - b_{du}a_{ud})^{-1}b_{dd} \\ a_{du} + a_{dd}b_{du}(1 - a_{ud}b_{du})^{-1}a_{uu} & a_{dd}(1 - b_{du}a_{ud})^{-1}b_{dd} \end{bmatrix}, \quad (23a)$$

stable, nonnormalized  $R$ -matrix recursion formulas obtained by using the interface  $t$  matrix are given in Appendix C.

We recall that in Subsection 2.C we formally derived the layer  $t$  matrix from the boundary equation, Eq. (2a) or Eq. (2b). In fact, in at least two important cases Eq. (5a) or Eq. (5b) is obtained without the aid of Eq. (2a) or Eq. (2b). The first case is the classical modal method in which matrix  $t^{(p)}$  is obtained directly by projecting the functional boundary equations onto a natural basis function set,<sup>7</sup> and the second case is the differential method in which matrix  $\tilde{t}^{(p)}$  is obtained from a numerical integration procedure.<sup>9</sup> In these cases the use of the non-normalized recursion formulas may be beneficial because they require fewer matrix operations.

In the C method and the coupled-wave method the boundary equations (2a) or (2b) are an integral step of the numerical treatment. In this case we can bypass the  $t$  matrix and derive the  $s$  (or  $r$ ) matrix directly from the boundary equations. Writing the two  $W$  matrices in Eq. (2a) in a two-by-two form, and rearranging the terms slightly, we have

$$\begin{bmatrix} W_{11}^{(p+1)} & -W_{12}^{(p)} \\ W_{21}^{(p+1)} & -W_{22}^{(p)} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(p+1)}(y_p + 0) \\ \mathbf{d}^{(p)}(y_p - 0) \end{bmatrix} = \begin{bmatrix} W_{11}^{(p)} & -W_{12}^{(p+1)} \\ W_{21}^{(p)} & -W_{22}^{(p+1)} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(p)}(y_p - 0) \\ \mathbf{d}^{(p+1)}(y_p + 0) \end{bmatrix}. \quad (21a)$$

Therefore from Eq. (10a),

$$s^{(p)} = \begin{bmatrix} W_{11}^{(p+1)} & -W_{12}^{(p)} \\ W_{21}^{(p+1)} & -W_{22}^{(p)} \end{bmatrix}^{-1} \begin{bmatrix} W_{11}^{(p)} & -W_{12}^{(p+1)} \\ W_{21}^{(p)} & -W_{22}^{(p+1)} \end{bmatrix}. \quad (22a)$$

The layer  $s$  matrix,  $\tilde{s}^{(p)}$ , then follows immediately from Eq. (13a). Similarly the interface  $r$  matrix,  $r^{(p)}$ , can be derived directly from boundary equation (2b); i.e.,

$$r^{(p)} = \begin{bmatrix} W_{11}^{(p+1)} & -W_{11}^{(p)} \\ W_{21}^{(p+1)} & -W_{21}^{(p)} \end{bmatrix}^{-1} \begin{bmatrix} -W_{12}^{(p+1)} & W_{12}^{(p)} \\ -W_{22}^{(p+1)} & W_{22}^{(p)} \end{bmatrix}. \quad (22b)$$

The layer  $r$  matrix,  $\tilde{r}^{(p)}$ , then follows immediately from Eq. (13b).

## B. Variation in Recursion Order

The  $S$ -matrix and  $R$ -matrix recursions do not have to be performed in the order indicated in Sections 3 and 4. In other words, one does not have to start the calculation from medium 0 and work step by step up to medium  $n + 1$ . For the normalized recursions this point can be best illustrated by the use of Redheffer's start product.<sup>20</sup> Let  $a$ ,  $b$ , and  $c$  be  $2N \times 2N$  matrices. Then the star product of  $a$  and  $b$ , in the  $S$ -matrix algorithm, is defined as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} * \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} b_{11} - b_{12}(b_{22} - a_{11})^{-1}b_{21} & b_{12}(b_{22} - a_{11})^{-1}a_{12} \\ -a_{21}(b_{22} - a_{11})^{-1}b_{21} & a_{22} + a_{21}(b_{22} - a_{11})^{-1}a_{12} \end{bmatrix}. \quad (23b)$$

It can be shown that the star multiplications are associative, i.e., that

$$a * (b * c) = (a * b) * c, \quad (24a)$$

$$a * (b * c) = (a * b) * c. \quad (24b)$$

In the remainder of this section, for simplicity, I will mention only the  $S$ -matrix recursion. The results for the  $R$ -matrix recursion can be obtained by obvious substitutions.

In terms of star products, the  $S$ -matrix algorithm can be succinctly expressed as

$$S^{(n)} = \{ \dots [(\tilde{s}^{(0)} * \tilde{s}^{(1)}) * \tilde{s}^{(2)}] * \dots \} * \tilde{s}^{(n)}. \quad (25a)$$

However, because of the associativity of the star multiplication, the product can be regrouped as follows:

$$S^{(n)} = \tilde{s}^{(0)} * \{ \dots * [\tilde{s}^{(n-2)} * (\tilde{s}^{(n-1)} * \tilde{s}^{(n)})] \dots \}. \quad (26a)$$

Equation (26a) describes the  $S$ -matrix recursion in the reverse order, starting from medium  $n + 1$  and moving downward to medium 0.

The associativity of the  $S$ -matrix recursion can be used advantageously to save computation effort or to increase computation speed. Suppose that a large number of calculations are to be carried out for a grating with a varying parameter that affects only the  $j$ th layer, for fixed  $j$ . Then the  $S$ -matrix recursion can be performed like this:

$$S^{(n)} = [\tilde{s}^{(0)} * \dots * \tilde{s}^{(j-1)}] * \tilde{s}^{(j)} * [\tilde{s}^{(j-1)} * \dots * \tilde{s}^{(n)}], \quad (27a)$$

where the two recursions inside the square brackets are performed just once and then are used repeatedly to form a star product with the changing  $\tilde{s}^{(j)}$ . If the grating calculation is done on a computer capable of parallel pro-

cessing, then the  $s$  matrices can be grouped pairwise to increase the computation speed.

## 6. COMPARISONS

Having systematically addressed the  $S$ -matrix and  $R$ -matrix algorithms and their variants, we are now ready to make some comparisons. As mentioned in Sections 3 and 4, the  $S$  matrices are related to the physical concept of reflection and transmission, and the  $R$  matrices are related to the physical concept of impedance and admittance. Furthermore, the normalized  $S$ -matrix recursion formulas can be readily interpreted in terms of multiple reflections, but the  $R$ -matrix recursion formulas and the nonnormalized  $S$ -matrix recursion formulas are not easily interpreted in physical terms. In what follows we compare the numerical stabilities and efficiencies of the two algorithms.

### A. Numerical Stabilities

Both the  $S$ -matrix and  $R$ -matrix algorithms are inherently stable because the  $S$  matrices and the  $R$  matrices are both mathematically of order  $O(1)$ . However, there are subtle numerical differences between them. In general the  $S$ -matrix algorithm is much easier to work with than the  $R$  matrix algorithm.

The implementation of the  $S$ -matrix algorithm is mostly worry free (but see Subsection 7.D), thanks to its use of the exponential basis functions. All  $s$  and  $S$  matrices are of order  $O(1)$ , not only mathematically but also numerically [we assume that Eqs. (20a) are not used]. Thus there is no limit in layer thickness. The possibility of numerical overflow associated with the exponential functions is eliminated because only the decreasing exponential functions are evaluated. Underflow can happen, but it is not a problem for most compilers. Additionally, the occurrence of  $\lambda_m^{(p)} h_p = l\pi$  is not a problem at all.

The implementation of the  $R$ -matrix algorithm requires special treatment. Although all  $r$  and  $R$  matrices are of order  $O(1)$  mathematically, they may not be so numerically. When the factorization of the layer  $t$  matrix into the product of the interface  $t$  matrix and the diagonal matrix  $\phi$  is possible, one should use Eqs. (C1) and (C2) below to perform the nonnormalized recursion or use Eqs. (13b) to compute  $\tilde{r}^{(p)}$  if the normalized recursion is to be used. When the factorization is impossible, as is the case for the differential method, the numerical layer thicknesses should be kept sufficiently low that the computation of  $\tilde{r}_{12}^{(p)}$  by Eq. (14b') or of  $R_{12}^{(p)}$  by Eq. (20b) will not suffer significant loss of accuracy. There is also a minor technical complication. Functions  $\cot$  and  $\csc$  that admit a complex argument are not intrinsic functions in most compilers. Thus the programmer has to write  $\cot$  and  $\csc$  as user-defined functions, using either the sine and cosine functions or the exponential functions. In doing so, care has to be taken to avoid overflow. In principle, the accidental occurrence of  $\lambda_m^{(p)} h_p = l\pi$  is a problem for the  $R$ -matrix algorithm. In practice, it is highly unlikely that the equality holds for  $l \neq 0$  with high precision; therefore the singularity never poses serious numerical problems. Of course, one should judiciously avoid  $h_p = 0$ , which is an uninteresting case, anyway.

### B. Numerical Efficiencies

In this subsection we consider the numerical efficiencies of different variants of the  $S$ -matrix and  $R$ -matrix algorithms. More specifically, we estimate the number of algebraic operations that each variant takes to compute the outgoing diffraction amplitudes  $\mathbf{u}^{(n+1)}$  and  $\mathbf{d}^{(0)}$ , assuming that the  $W$  matrices in the boundary equations have been obtained.

As is evident from the presentations in Sections 3 and 4, after the  $S$ -matrix recursion is completed,  $\mathbf{u}^{(n+1)}$  and  $\mathbf{d}^{(0)}$  are readily given in a solved form. However, with the  $R$ -matrix algorithm the completion of the  $R$ -matrix recursion only gives a system of linear equations that has yet to be solved to yield  $\mathbf{u}^{(n+1)}$  and  $\mathbf{d}^{(0)}$ . This initial comparison is already in favor of the  $S$ -matrix algorithm.

We now take a closer look at the structure of the matrix recursion formulas. We say that a subset of the four submatrices of  $S^{(p)}$  is a closed set with respect to the  $S$ -matrix recursion if every element of the set is determined by the elements in the same set. Thus  $S^{(p)}$  has four closed proper subsets:

$$\begin{aligned} \{R_{ud}^{(p)}\}, \quad \{R_{ud}^{(p)}, T_{dd}^{(p)}\}, \quad \{R_{ud}^{(p)}, T_{uu}^{(p)}\}, \\ \{R_{ud}^{(p)}, T_{uu}^{(p)}, T_{dd}^{(p)}\}. \end{aligned} \quad (28a)$$

If there are incident plane waves in both media 0 and  $n+1$ , then from Eq. (17a) the  $S$ -matrix recursion of all four submatrices has to be performed. Suppose that  $\mathbf{u}^{(0)} = 0$ ; then only the recursion of  $\{R_{ud}^{(p)}, T_{dd}^{(p)}\}$  is necessary if both  $\mathbf{u}^{(n+1)}$  and  $\mathbf{d}^{(0)}$  are needed, and only the recursion of  $R_{ud}^{(p)}$  is necessary if only  $\mathbf{u}^{(n+1)}$  is needed. We call the recursions above the full, half, and quarter  $S$ -matrix recursions, respectively.

Similarly,  $R^{(p)}$  also has four closed proper subsets:

$$\begin{aligned} \{R_{11}^{(p)}\}, \quad \{R_{11}^{(p)}, R_{12}^{(p)}\}, \quad \{R_{11}^{(p)}, R_{21}^{(p)}\}, \\ \{R_{11}^{(p)}, R_{12}^{(p)}, R_{21}^{(p)}\}. \end{aligned} \quad (28b)$$

With the  $R$ -matrix algorithm, if both  $\mathbf{u}^{(n+1)}$  and  $\mathbf{d}^{(0)}$  are needed, the full matrix recursion has to be performed even when  $\mathbf{u}^{(0)} = 0$ . If  $\mathbf{d}^{(0)}$  is not needed and  $\mathbf{u}^{(0)} = 0$ , then quarter  $R$ -matrix recursion with  $R_{11}^{(p)}$  is possible, but the  $R$  matrix has to be initialized by

$$R^{(-1)} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \quad (29b)$$

With this initialization,  $R_{12}^{(p)} = R_{21}^{(p)} = 0$  and  $R_{22}^{(p)} = -1$  for all  $p$ .

Finally, we shall provide operation counts per grating layer for the variants of the two recursive matrix algorithms that have been presented in this paper. The operation counts will be given in units of flops.<sup>21</sup> The counts do not include the effort in assembling the  $W$  matrices and in solving the final linear system to yield  $\mathbf{u}^{(n+1)}$  and  $\mathbf{d}^{(0)}$ . For convenience, we shall consider only the operations that are proportional to  $N^3$ , where  $N$  is the truncation order, the dimension of the submatrices. The method of counting is based on well-established rules<sup>21</sup>: suppose that  $A$ ,  $B$ , and  $C$  are  $N \times N$  nonsparse matrices.

**Table 1. Operation Counts (in  $N^3$  Flops) per Grating Layer for Different Variants of the S-Matrix and R-Matrix Algorithms**

Algorithm Number	Algorithm	Stability	Operation Counts		
			Full	Half	Quarter
1a	$W \rightarrow t \rightarrow s \rightarrow \bar{s} \rightarrow S$	Unconditional	76/3	20	19
2a	$W \rightarrow t \rightarrow S$	Unconditional	19	15	14
3a	$W \rightarrow s \rightarrow \bar{s} \rightarrow S$	Unconditional	64/3	16	15
4a	$W \rightarrow \bar{i} \rightarrow S$	Conditional	19	15	14
1b	$W \rightarrow t \rightarrow r \rightarrow \bar{r} \rightarrow R$	Unconditional	25	—	21
2b	$W \rightarrow t \rightarrow R$	Unconditional	23	—	15
3b	$W \rightarrow r \rightarrow \bar{r} \rightarrow R$	Unconditional	21	—	17
4b	$W \rightarrow \bar{i} \rightarrow \bar{r} \rightarrow R$	Conditional	21	—	17
5b	$W \rightarrow \bar{i} \rightarrow R$	Conditional	19	—	14

ces; then  $AB + C$ ,  $A^{-1}$ , and  $A^{-1}B + C$  take  $N^3$ ,  $N^3$ , and  $(4/3)N^3$  flops, respectively.

The results are summarized in Table 1 where implementation variants of the S-matrix and R-matrix algorithms that have been described in Sections 3, 4, and 5 are represented symbolically. For example,  $W \rightarrow t \rightarrow S$  represents the variant of the S-matrix algorithm that uses Eq. (7) to compute the interface  $t$  matrix and then uses Eqs. (19a) to perform the S-matrix recursion. The broken arrows indicate that in some grating models the  $t$  matrices are obtained without using the  $W$  matrices. In this case,  $(32/3)N^3$  flops should be subtracted from the operation counts in Table 1. The subheadings of the last three columns stand for full-, half-, and quarter-matrix recursions, respectively. Since the half-matrix recursion of the  $R$  matrices serves no useful purpose, the corresponding operation counts are not given. Clearly, algorithms 2a and 3a are the most efficient, assuming that we start with the  $W$  matrices.

## 7. REMARKS

### A. Algorithm of Chateau and Hugonin

It is easy to see that the algorithm proposed by Chateau and Hugonin,<sup>8</sup> except for the notational differences, is algorithm 2a in Table 1 for the special case in which  $\mathbf{u}^{(0)} = 0$ . It is one of the most efficient variants of the general S-matrix algorithm, but it can be slightly improved. In Ref. 8 each of the three factors of the layer  $t$  matrix [see Eqs. (6) and (7)] is passed through the recursion formula separately. So the operation count, including the inversion of  $W^{(p-1)}$ , is  $(50/3)N^3$  for the half-recursion. In comparison, the use of product  $t = W^{(p-1)^{-1}}W^{(p)}$  in Eq. (19a) costs  $15N^3$  flops.

### B. R-Matrix Algorithm and Differential Method

For the differential method it is natural to use the  $R$ -matrix algorithm because here the  $U$ - $V$  basis is the natural basis. The factorization of the layer  $t$  matrices is unavailable in the differential method, so the application of the  $R$ -matrix starts with the layer  $t$  matrices. As explained in Sections 4 and 5, use of the  $\bar{i}$  matrix in the  $R$ -matrix algorithm makes the stability of the algorithm conditional. Although the modal methods were assumed when we analyzed the cause of numerical instability of Eq. (14b'), the conclusion applies to the differential method as well, because the basis functions are asymptotically the same in the two cases.

The first few rows of Tables 1, 2, and 3 in Ref. 9 clearly indicate that if the numerical layer thicknesses are not kept low, the  $R$ -matrix algorithm fails when applied to the differential method.

The  $R$ -matrix algorithm that is used in Ref. 9 is algorithm 4b in Table 1 of this paper, which takes  $(31/3)N^3$  flops per layer for the full-matrix recursion. It can be improved slightly by using algorithm 5b, which takes  $(25/3)N^3$  flops per layer. Instead, if the  $\mathbf{u}$ - $\mathbf{d}$  basis functions and algorithm 4a are used, significant improvement can be achieved. The operation count for the S-matrix algorithm is only  $(13/3)N^3$  flops per layer for the half-matrix recursion. Furthermore, the extra work of solving the final system of linear equations, Eq. (17b), is avoided.

### C. R-Matrix Algorithm and Classical Modal Method

In the classical modal method,<sup>7</sup> thanks to the orthogonality of the modal functions and the fact that the pivotal submatrix  $t_{21}^{(p)} = 0$ , not only are the  $t$  matrices obtained analytically without the  $W$  matrices, but the  $r$  matrices can also be determined analytically from the  $t$  matrices without numerical inversion of the pivotal submatrix. The result is the most efficient variant of the  $R$ -matrix algorithm, which can be symbolized simply as  $\bar{r} \rightarrow R$ . Only one of the four submatrices of  $\bar{r}^{(p)}$  takes  $N^3$  flops to construct; the rest involve only  $N^2$  processes. Thus the overall operation counts are only  $(22/3)N^3$  and  $(10/3)N^3$  per layer for the full- and quarter-matrix recursions, respectively.

### D. S-Matrix Algorithm and Classical Modal Method

The S-matrix algorithm can be applied to the classical modal method, the most efficient variant being algorithm 2a of Table 1. The combination of the S-matrix algorithm and the classical modal method has a peculiar problem, which I shall describe below.

In the classical modal method, the  $t$  matrices that use the exponential basis functions can also be obtained analytically without the  $W$  matrices, but the  $s$  matrices in general cannot. Without going into any detail, suffice it to say that the elements of  $t^{(p)}$  at a numerical interface are all of the form

$$\int \psi_i^{(p+1)}(x) f_p(x) \psi_n^{(p)}(x) dx, \quad (30)$$

where the integration is over one grating period,  $\psi_i^{(p+1)}$

and  $\psi_n^{(p)}$  are modal functions in layers  $p + 1$  and  $p$ , respectively, and  $f_p$  is a function that depends only on the permittivity distribution of the two layers. In what follows, we consider the evaluation of Eq. (30) under a specific set of conditions: (1) the permittivity distributions in two adjacent layers are symmetrical with respect to the origin of the  $x$  axis, (2) the grating is in the first-order Littrow mount, and (3)  $N = 2M + 1$ , where  $M$  is a natural number. Under condition (1),  $f_p$  is a symmetrical function. Under condition (2),  $\psi_l^{(p+1)}$  and  $\psi_n^{(p)}$  are either symmetrical or antisymmetrical functions. Thus if integers  $l$  and  $n$  correspond to modal functions of different parities, the corresponding  $t$  matrix element is identically zero. In the classical modal method, one normally indexes the eigenvalues in the order of increasing absolute values. Let  $N_e^{(p)}$  and  $N_o^{(p)}$  be the numbers of even and odd eigenvalues, respectively. Numerical experiments show that, under condition (2) and for a given truncation order  $N = N_e^{(p)} + N_o^{(p)}$ ,  $N_e^{(p)}$  and  $N_o^{(p)}$  never differ by more than 1. Thus under condition (3), either  $N_e^{(p)} = M$  and  $N_o^{(p)} = M + 1$ , or  $N_e^{(p)} = M + 1$  and  $N_o^{(p)} = M$ . It can be easily shown that if  $N_e^{(p)} \neq N_e^{(p+1)}$  then all submatrices of  $t^{(p)}$ , in particular the pivotal submatrix  $t_{22}^{(p)}$ , are mathematically singular. Numerically, the condition  $N_e^{(p)} \neq N_e^{(p+1)}$  does often occur; therefore algorithm 1a of Table 1 cannot be applied to the classical modal method when conditions (1), (2), and (3) are met simultaneously. It can be verified numerically that the matrix sum that is to be inverted in Eqs. (19a) sometimes becomes numerically ill-conditioned under the above three conditions; therefore algorithm 2a fails too, even though it does not involve the inversion of the pivotal submatrix  $t_{22}^{(p)}$ .

Fortunately, the singular matrix problem described above can be easily avoided by the use of an even truncation order. If  $N = 2M$ , then the characteristics of the eigenvalue distribution automatically guarantee that  $N_e^{(p)} = N_o^{(p)} = M$ .

Another interesting difference between the  $R$ -matrix algorithm and the  $S$ -matrix algorithm, as they are applied to the classical modal method, is that the law of energy conservation (in the case of dielectric gratings) is satisfied automatically by the former, but it is satisfied only with increasing truncation orders by the latter.

### E. Other Possibilities

The essence of the  $R$ -matrix and  $S$ -matrix algorithms is to avoid the presence of the growing exponential functions in the matrix manipulations. In this spirit, several other stable algorithms have recently been presented in the literature. For example, Montiel and Nevère<sup>9</sup> presented an algorithm that they called the  $R'$ -matrix algorithm. In view of the current paper, it can be considered an  $S$ -matrix algorithm that uses the  $U-V$  basis functions. In Ref. 10 the  $R$ -matrix algorithm was used with the exponential basis functions (maybe it can be called the  $S'$ -matrix algorithm?). The scattering-matrix approach of Cotter *et al.*<sup>11</sup> is essentially algorithm 2a of Table 1, except that their  $t$  matrix is the inverse of the  $t$  matrix in this paper. Clearly there are many other possibilities, but it is pointless to enumerate all of them.

## 8. SUMMARY

The mathematical formulations of the  $S$ -matrix and  $R$ -matrix algorithms have been systematically presented. The presentation is given in a unified fashion, independent of underlying grating models, grating geometries, and grating mountings. The physical interpretations of the algorithms are illustrated. In addition, many variants of the algorithms are presented and their numerical stabilities and efficiencies analyzed.

The  $S$ -matrix and  $R$ -matrix algorithms are inherently stable because they avoid the appearance of the exponentially growing submatrices in the recursion formulas. However, to further ensure that the algorithms be unconditionally stable, effort should be made to avoid the exponentially growing submatrices in the intermediate steps, i.e., in the calculation of the layer  $s$  or  $r$  submatrices. Whenever the factorization, as given in Eq. (6), of the layer  $t$  matrix is possible, the interface  $t$  matrix should be used directly in the constructions of layer  $s$  (or  $r$ ) matrix or in the nonnormalized  $S$ -matrix (or  $R$ -matrix) recursion. When factorization is impossible, the  $S$ -matrix and  $R$ -matrix algorithms are stable under the condition that the layer thicknesses and the truncation order be kept low, as quantified at the end of Section 4.

The comparative study of the two matrix algorithms presented here seems to favor the  $S$ -matrix algorithm. The physical interpretation of the  $S$  matrix in terms of reflections and transmissions is more intuitive than that of the  $R$  matrix in terms of the impedance and admittance. The exponential basis functions adopted by the  $S$ -matrix algorithm are numerically much easier to handle than the trigonometrical basis functions adopted by the  $R$  matrix algorithm. Based on the operation counts, the  $S$ -matrix algorithm is more efficient than the  $R$ -matrix algorithm.

Many implementation variants of the algorithms are presented in this paper. The variants that use all intermediate matrices, algorithms 1a and 1b in Table 1, are the least efficient ones. They have only pedagogical value. The variants that bypass some of the intermediate matrices, for example, algorithms 2a, 3a, and 2b, are the most efficient ones. However, as exemplified in Section 7, which algorithm and variant are more efficient often depends on the grating model being used. It is the hope of the author that the information provided here will enable the reader to apply the most efficient algorithm to the grating model at his or her disposal.

## APPENDIX A

To derive Eq. (13b), let us imagine that layer  $p$  is a sum of two layers. The first layer has zero thickness, with the layer  $t$  and  $r$  matrices given by Eqs. (7) and (14b), respectively. The second layer has thickness  $h_p$ , but it does not cross a material boundary. Its equivalent layer  $t$  matrix is just  $\phi^{(p)}$ , given by Eq. (4b). Denoting the equivalent layer  $r$  matrix corresponding to  $\phi^{(p)}$  by  $\hat{r}^{(p)}$ , from Eq. (14b) we have

$$\hat{r}^{(p)} = \begin{bmatrix} -\eta_m^{(p)} \cot(\lambda_m^{(p)} h_p) & \eta_m^{(p)} \csc(\lambda_m^{(p)} h_p) \\ -\eta_m^{(p)} \csc(\lambda_m^{(p)} h_p) & \eta_m^{(p)} \cot(\lambda_m^{(p)} h_p) \end{bmatrix}. \quad (\text{A1})$$

Equations (15b) can be viewed as a set of rules that

combine matrix  $\tilde{r}^{(p)}$  in relation

$$\begin{bmatrix} U^{(p+1)} \\ U^{(p)} \end{bmatrix} = \tilde{r}^{(p)} \begin{bmatrix} V^{(p+1)} \\ V^{(p)} \end{bmatrix} \quad (\text{A2})$$

and matrix  $R^{(p-1)}$  in relation

$$\begin{bmatrix} U^{(p)} \\ U^{(0)} \end{bmatrix} = R^{(p-1)} \begin{bmatrix} V^{(p)} \\ V^{(0)} \end{bmatrix} \quad (\text{A3})$$

to obtain matrix  $R^{(p)}$  in relation

$$\begin{bmatrix} U^{(p+1)} \\ U^{(0)} \end{bmatrix} = R^{(p)} \begin{bmatrix} V^{(p+1)} \\ V^{(0)} \end{bmatrix}. \quad (\text{A4})$$

Here we have

$$\begin{bmatrix} U^{(p+1)}(y_p + 0) \\ U^{(p)}(y_p - 0) \end{bmatrix} = r^{(p)} \begin{bmatrix} V^{(p+1)}(y_p + 0) \\ V^{(p)}(y_p - 0) \end{bmatrix}, \quad (\text{A5})$$

$$\begin{bmatrix} U^{(p)}(y_p - 0) \\ U^{(p)}(y_{p-1} + 0) \end{bmatrix} = \hat{r}^{(p)} \begin{bmatrix} V^{(p)}(y_p - 0) \\ V^{(p)}(y_{p-1} + 0) \end{bmatrix}, \quad (\text{A6})$$

and what we want is the matrix in

$$\begin{bmatrix} U^{(p+1)}(y_p + 0) \\ U^{(p)}(y_{p-1} + 0) \end{bmatrix} = \tilde{r}^{(p)} \begin{bmatrix} V^{(p+1)}(y_p + 0) \\ V^{(p)}(y_{p-1} + 0) \end{bmatrix}. \quad (\text{A7})$$

Through comparison of Eqs. (A2)–(A4) with Eqs. (A5)–(A7), it is evident that the two sets of equations have identical algebraic structures. Therefore Eqs. (13b) can be obtained from Eqs. (15b) provided that  $r^{(p)}$ ,  $\hat{r}^{(p)}$ , and  $\tilde{r}^{(p)}$  are identified with  $\tilde{r}^{(p)}$ ,  $R^{(p-1)}$ , and  $R^{(p)}$ , respectively.

## APPENDIX B

When  $t_{21}^{(p)} = 0$ , which happens in the classical modal method, Eq. (14b) cannot be used to compute  $r^{(p)}$ , but in this case for sure  $t_{22}^{(p)} \neq 0$ . Suppose that  $t_{22}^{(p)}$  is nonsingular and  $\lambda_m^{(p)} h_p \neq l\pi$ , then Eq. (14b') can be used to derive the layer  $r$  matrix. After some simple algebra, we have

$$\tilde{r}^{(p)} = \begin{bmatrix} [t_{12}^{(p)} - t_{11}^{(p)} \eta_m^{(p)} \cot(\lambda_m^{(p)} h_p)] t_{22}^{(p)-1} & t_{11}^{(p)} \eta_m^{(p)} \csc(\lambda_m^{(p)} h_p) \\ -\eta_m^{(p)} \csc(\lambda_m^{(p)} h_p) t_{22}^{(p)-1} & \eta_m^{(p)} \cot(\lambda_m^{(p)} h_p) \end{bmatrix}. \quad (\text{B1})$$

This expression of  $\tilde{r}^{(p)}$ , like Eqs. (13b), contains no exponentially growing functions, and therefore is suitable for the unconditionally stable  $R$ -matrix recursion.

## APPENDIX C

Similarly to the treatment in Appendix A, we consider that each of the two factors in Eq. (6) corresponds to a layer, one with zero thickness and the other with the full thickness  $h_p$ . Substituting  $\phi^{(p)}$  for  $\tilde{r}^{(p)}$  in Eqs. (20b), and making some simple algebraic rearrangement, we obtain

an intermediate  $R$  matrix,  $\hat{R}^{(p)}$ , which is given by

$$\begin{aligned} \hat{R}_{11}^{(p)} &= -\eta_m^{(p)} \cot(\lambda_m^{(p)} h_p) + \eta_m^{(p)} \csc(\lambda_m^{(p)} h_p) \\ &\quad \times \omega^{(p)} \eta_m^{(p)} \csc(\lambda_m^{(p)} h_p), \\ \hat{R}_{12}^{(p)} &= \eta_m^{(p)} \csc(\lambda_m^{(p)} h_p) \omega^{(p)} R_{12}^{(p-1)}, \\ \hat{R}_{21}^{(p)} &= R_{21}^{(p-1)} \omega^{(p)} \eta_m^{(p)} \csc(\lambda_m^{(p)} h_p), \\ \hat{R}_{22}^{(p)} &= R_{22}^{(p-1)} + R_{21}^{(p-1)} \omega^{(p)} R_{12}^{(p-1)}, \end{aligned} \quad (\text{C1})$$

where

$$\omega^{(p)} = [\eta_m^{(p)} \cot(\lambda_m^{(p)} h_p) - R_{11}^{(p-1)}]^{-1}. \quad (\text{C1}')$$

Clearly, all submatrices of  $\hat{R}^{(p)}$  are of order  $O(1)$  and numerically stable. To complete the nonnormalized  $R$ -matrix recursion using the interface  $t$  matrix, we need only to use Eqs. (20b) again, this time replacing  $R^{(p-1)}$  by  $\hat{R}^{(p)}$  and  $\tilde{r}^{(p)}$  by  $t^{(p)}$ . The result is

$$\begin{aligned} R_{11}^{(p)} &= [t_{12}^{(p)} + t_{11}^{(p)} \hat{R}_{11}^{(p)}] [t_{22}^{(p)} + t_{21}^{(p)} \hat{R}_{11}^{(p)}]^{-1}, \\ R_{12}^{(p)} &= [t_{11}^{(p)} - R_{11}^{(p)} t_{21}^{(p)}] \hat{R}_{12}^{(p)}, \\ R_{21}^{(p)} &= \hat{R}_{21}^{(p)} [t_{22}^{(p)} + t_{21}^{(p)} \hat{R}_{11}^{(p)}]^{-1}, \\ R_{22}^{(p)} &= \hat{R}_{22}^{(p)} - R_{21}^{(p)} t_{21}^{(p)} \hat{R}_{12}^{(p)}. \end{aligned} \quad (\text{C2})$$

## ACKNOWLEDGMENTS

This research was supported by the Optical Data Storage Center, University of Arizona, and by the advanced Technology Program of the U.S. Department of Commerce through a grant to the National Storage Industry Consortium.

## REFERENCES AND NOTES

1. L. Li and J. Hirsh, "All-dielectric high-efficiency reflection gratings made with multilayer thin film coatings," *Opt. Lett.* **20**, 1349–1351 (1995).
2. C. Heine and R. H. Morf, "Submicrometer gratings for solar energy applications," *Appl. Opt.* **34**, 2476–2482 (1995).
3. M. Nevière, "Bragg-Fresnel multilayer gratings: electromagnetic theory," *J. Opt. Soc. Am. A* **11**, 1835–1845 (1994).
4. D. Maystre, "Electromagnetic study of photonic band gaps," *Pure Appl. Opt.* **3**, 975–993 (1994).
5. D. M. Pai and K. A. Awada, "Analysis of dielectric gratings of arbitrary profiles and thicknesses," *J. Opt. Soc. Am. A* **8**, 755–762 (1991).
6. L. F. DeSandre and J. M. Elson, "Extinction-theorem analysis of diffraction anomalies in overcoated gratings," *J. Opt. Soc. Am. A* **8**, 763–777 (1991).
7. L. Li, "Multilayer modal method for diffraction gratings of arbitrary profile, depth, and permittivity," *J. Opt. Soc. Am. A* **10**, 2581–2591 (1993).
8. N. Chateau and J. P. Hugonin, "Algorithm for the rigorous coupled-wave analysis of grating diffraction," *J. Opt. Soc. Am. A* **11**, 1321–1331 (1994).
9. 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**, 3241–3250 (1994).
10. L. Li, "Multilayer-coated diffraction gratings: differential method of Chandezon *et al.* revisited," *J. Opt. Soc. Am. A* **11**, 2816–2828 (1994).
11. N. P. K. Cotter, T. W. Preist, and J. R. Sambles, "Scattering-matrix approach to multilayer diffraction," *J. Opt. Soc. Am. A* **12**, 1097–1103 (1995).

12. L. Li, "Bremmer series,  $R$ -matrix propagation algorithm, and numerical modeling of diffraction gratings," *J. Opt. Soc. Am. A* **11**, 2829–2836 (1994).
13. M. G. Moharam, D. A. Pommet, E. B. Grann, and T. K. Gaylord, "Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmission matrix approach," *J. Opt. Soc. Am. A* **12**, 1077–1086 (1995).
14. M. G. Moharam and T. K. Gaylord, "Diffraction analysis of dielectric surface-relief gratings," *J. Opt. Soc. Am.* **72**, 1385–1392 (1982).
15. R. H. Morf, "Exponentially convergent and numerically efficient solution of Maxwell's equations for lamellar gratings," *J. Opt. Soc. Am. A* **12**, 1043–1056 (1995).
16. G. Granet, J. P. Plumey, and J. Chandezon, "Scattering by a periodically corrugated dielectric layer with non-identical faces," *Pure Appl. Opt.* **4**, 1–5 (1995).
17. L. Li, G. Granet, J. P. Plumey, and J. Chandezon, "Some topics in extending the C method to coated gratings with different profiles," *Pure Appl. Opt.* **5**, 141–156 (1996).
18. The statement "matrix  $A$  is of order  $O(1)$ " means that every element of  $A$  is of order  $O(1)$ . In the context of this paper saying that a matrix is of order  $O(1)$  means that it contains no exponentially growing functions with respect to layer thickness and matrix truncation order.
19. A. K. Cousins and S. C. Gottschalk, "Application of the impedance formalism to diffraction gratings with multiple coating layers," *Appl. Opt.* **29**, 4268–4271 (1990).
20. R. Redheffer, "Difference equations and functional equations in transmission-line theory," in *Modern Mathematics for the Engineer*, E. F. Beckenbach, ed. (McGraw-Hill, New York, 1961), Chap. 12, pp. 282–337.
21. G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore, 1983), Chap. 3, p. 30, and Chap. 4, p. 52.



# Highly improved convergence of the coupled-wave method for TM polarization

Philippe Lalanne

*Institut d'Optique Théorique et Appliquée, Centre National de la Recherche Scientifique,  
BP 147, 91403 Orsay Cedex, France*

G. Michael Morris

*The Institute of Optics, University of Rochester, Rochester, New York 14627*

Received June 12, 1995; revised manuscript received August 28, 1995; accepted August 30, 1995

The coupled-wave method formulated by Moharam and Gaylord [J. Opt. Soc. Am. **73**, 451 (1983)] is known to be slowly converging, especially for TM polarization of metallic lamellar gratings. The slow convergence rate has been analyzed in detail by Li and Haggans [J. Opt. Soc. Am. A **10**, 1184 (1993)], who made clear that special care must be taken when coupled-wave methods are used for TM polarization. By reformulating the eigenproblem of the coupled-wave method, we provide numerical evidence and argue that highly improved convergence rates similar to the TE polarization case can be obtained. The discussion includes both nonconical and conical mountings.

*Key words:* coupled-wave methods, rigorous grating analysis, diffractive optics. © 1996 Optical Society of America

## 1. INTRODUCTION

In a recent publication Li and Haggans<sup>1</sup> provided strong numerical evidence that the rigorous coupled-wave analysis (RCWA) formulated by Moharam and Gaylord<sup>2</sup> converges slowly for one-dimensional (1-D) metallic gratings and TM polarization (magnetic-field vector parallel to the grating vector). They argued that the slow convergence is caused by the slowly convergent Fourier expansions for the permittivity and the electromagnetic field inside the grating. The RCWA computation is twofold. First, the Fourier expansion of the field inside the grating provides a system of differential equations. Then once the eigenvalues and the eigenvectors of this system are found, the boundary conditions at the grating interfaces are matched to compute the diffraction efficiencies. In this paper we focus on the eigenproblem of 1-D gratings for TM polarization. By reformulating the eigenproblem, we report on highly improved convergence rates even for highly conductive gratings. We also reveal that the slow convergence is due not to the use of Fourier expansions but to an inadequate formulation of the conventional eigenproblem.

In Section 2 we review briefly the previous eigenproblem formulations used in coupled-wave analysis for nonconical mountings; these include the original formulation provided in Ref. 2 and an updated formulation by the same authors.<sup>3</sup> In Section 3 we propose a new formulation for the eigenproblem. This new formulation can be straightforwardly extended to any modified method<sup>4,5</sup> that is based on a Fourier expansion of the field in the grating. Section 4 provides numerical evidence that the new formulation significantly improves the convergence rate. Two examples showing the improved convergence rate are provided. The first one is taken from Ref. 6 in which Peng and Morris showed

that a very large number of orders must be retained to analyze accurately a wire-grid-polarizer problem. The second example is taken from Ref. 1, in which poor and oscillating convergence rates were observed with a highly conductive grating. In Section 5 a simple intuitive argument is used to explain the observed improvement, and in Section 6 the generalization to conical mounts is briefly derived. Concluding remarks are given in Section 7.

## 2. CONVENTIONAL EIGENPROBLEM

Let us consider a 1-D periodic structure along the  $x$  axis with an arbitrary permittivity profile  $\epsilon(x)$  (see Fig. 1). The  $z$  axis is perpendicular to the grating boundaries. The diffraction problem is invariant in the  $y$  direction. Magnetic effects are not considered in this paper, and the constant  $\mu_0$  denotes the permeability of the periodic structure.  $\epsilon_0$  is the permittivity of the vacuum. The period of the structure is denoted by  $\Lambda$ , and the length of the grating vector  $K$  is equal to  $2\pi/\Lambda$ . An incident plane wave with wavelength  $\lambda$  in the incident medium makes an angle  $\theta$  with the  $z$  direction in a nonconical mounting. We denote the magnitude of the wave vector of the incident wave by  $k$  ( $k = 2\pi/\lambda$ ),  $\beta$  ( $\beta = k \sin \theta$ ) is its  $x$  component, and  $k_0$  represents the magnitude of the incident plane-wave vector in a vacuum. A temporal dependence of  $\exp(i\omega t)$  of the wave is assumed ( $j^2 = -1$ ).  $\epsilon_m$  denotes the  $m$ th Fourier coefficient of  $\epsilon(x)/\epsilon_0$ , and  $a_m$  is used to denote the  $m$ th Fourier coefficients of  $\epsilon_0/\epsilon(x)$ .

Using the Floquet theorem, the  $x$  component  $E_x$  and the  $z$  component  $E_z$  of the electric field and the  $y$  component  $H_y$  of the magnetic field inside the grating can be expressed as<sup>2</sup>

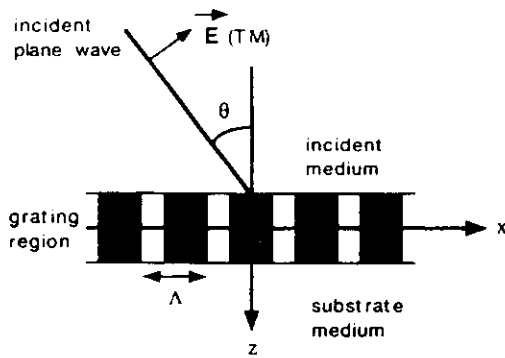


Fig. 1. Geometry for the nonconical grating diffraction problem analyzed in Sections 2 and 3 for TM polarization.

$$\begin{aligned} E_x &= \sum_m S_m(z) \exp j(Km + \beta)x, \\ E_z &= \sum_m f_m(z) \exp j(Km + \beta)x, \\ H_y &= \sqrt{\frac{\epsilon_0}{\mu_0}} \sum_m U_m(z) \exp j(Km + \beta)x. \end{aligned} \quad (1)$$

Maxwell's curl equations are

$$-\frac{\partial E_z}{\partial x} + \frac{\partial E_x}{\partial z} = -j\omega\mu_0 H_y, \quad (2a)$$

$$\frac{\partial H_y}{\partial z} = -j\omega\epsilon E_x, \quad (2b)$$

$$\frac{1}{\epsilon} \frac{\partial H_y}{\partial x} = j\omega E_z. \quad (2c)$$

In the following equations we denote the first derivative in the  $z$  variable by a prime. Consistently a double prime denotes the second derivative. Identified in the quasi-plane-wave basis, Eqs. (1) and (2) are used to obtain

$$-j(mK + \beta)f_m + S_m' = -jk_0 U_m, \quad (3a)$$

$$U_m' = -jk_0 \sum_p \epsilon_{m-p} S_p, \quad (3b)$$

$$\sum_p (pK + \beta) a_{m-p} U_p = k_0 f_m. \quad (3c)$$

By substituting  $f_m$  from Eq. (3c) into Eq. (3a), we obtain<sup>2</sup>

$$S_m' = -jk_0 U_m + j(mK/k_0 + \beta/k_0) \sum_p (pK + \beta) a_{m-p} U_p. \quad (4)$$

Equations (3b) and (4) provide a complete set of first-order differential equations and constitute an eigenproblem of size  $2(2M + 1)$  when  $\pm M$  orders are retained in the computation. As was noted by Li and Haggans<sup>1</sup> and was systematically exploited by Peng and Morris<sup>6</sup> and Moharam *et al.*,<sup>3</sup> it can be an advantage to solve the set of second-order differential equations. This solution easily takes into account the double degeneracy of the eigenproblem and decreases the computational effort. Using Eqs. (3b) and (4) we obtain the infinite set of second-order differential equations for the magnetic field:

$$\begin{aligned} U_m'' &= -k_0^2 \sum_p \epsilon_{m-p} \left[ U_p - (pK/k_0 + \beta/k_0) \right. \\ &\quad \left. \times \sum_r (rK/k_0 + \beta/k_0) a_{p-r} U_r \right]. \end{aligned} \quad (5)$$

Except for minor notation disparities, Eqs. (3b) and (4) were originally introduced by Moharam and Gaylord.<sup>2</sup> Equation (5) can be found in Refs. 3 and 6. Equation (5) can be written in the compact form

$$k_0^{-2}[U''] = [\mathbf{E}(\mathbf{K}_x \mathbf{A} \mathbf{K}_x - \mathbf{I})][\mathbf{U}], \quad (6a)$$

where  $\mathbf{I}$  is the identity matrix,  $\mathbf{E}$  is the matrix formed by the permittivity harmonic coefficients,  $\mathbf{K}_x$  is a diagonal matrix with the  $i, i$  element being  $(iK + \beta)/k_0$ , and  $\mathbf{A}$  is the matrix formed by the inverse-permittivity harmonic coefficients.  $\mathbf{K}_x$ ,  $\mathbf{E}$ , and  $\mathbf{I}$  are notations of Ref. 3. When a finite number of orders are retained in the numerical computation, the authors of Ref. 3 prefer to implement the eigenproblem by numerically inverting matrix  $\mathbf{E}$  instead of directly taking the inverse-permittivity coefficients  $a_m$ . Replacing  $\mathbf{A}$  by  $\mathbf{E}^{-1}$  in Eq. (6a), we obtain

$$k_0^{-2}[U''] = [\mathbf{E}(\mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_x - \mathbf{I})][\mathbf{U}]. \quad (6b)$$

Equation (6b) is the same as Eqs. (35) and (36) of Ref. 3. For the following comparison, the eigenproblem of Eq. (6b) is used in the RCWA implementation.

### 3. REFORMULATION OF THE EIGENPROBLEM

In this section we derive a new set of differential equations and reformulate the eigenproblem. Equations (3b) and (3c) can be written as

$$-\sum_p a_{m-p} U_p' = jk_0 S_m, \quad (7a)$$

$$(mK + \beta)U_m = k_0 \sum_p \epsilon_{m-p} f_p. \quad (7b)$$

By substituting  $f_m$  from Eq. (3a) into Eq. (7b) and then eliminating  $S_m$  with Eq. (7a), we obtain another infinite set of second-order differential equations:

$$\begin{aligned} (mK/k_0 + \beta/k_0)U_m - \sum_p \frac{\epsilon_{m-p}}{pK/k_0 + \beta/k_0} U_p \\ = \frac{1}{k_0^2} \sum_{l,p} \frac{\epsilon_{m-p} a_{p-l}}{pK/k_0 + \beta/k_0} U_l''. \end{aligned} \quad (8)$$

Note that Eq. (8) is not valid for normal incidence ( $\beta = 0$ ) and must be replaced by Eqs. (15) as discussed in Section 5. In a compact form, Eq. (8) becomes  $k_0^{-2}[\mathbf{E} \mathbf{K}_x^{-1} \mathbf{A}][\mathbf{U}'] = [\mathbf{K}_x - \mathbf{E} \mathbf{K}_x^{-1}][\mathbf{U}]$ , which is written by multiplying both sides by  $(\mathbf{E} \mathbf{K}_x^{-1} \mathbf{A})^{-1}$ :

$$k_0^{-2}[U''] = [\mathbf{A}^{-1}(\mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_x - \mathbf{I})][\mathbf{U}], \quad (9)$$

with  $\mathbf{E}$ ,  $\mathbf{K}_x$ ,  $\mathbf{A}$ , and  $\mathbf{I}$  being defined as in Eqs. (6a) and (6b). Since  $\mathbf{A}^{-1}$  is identical to  $\mathbf{E}$  when an infinite number of orders are retained, Eqs. (6b) and (9) are fully equivalent. As will be shown with numerical examples in the next Section, and as will be argued in Section 5, this equivalence is true only when an infinite number of orders is retained. When truncating the matrices for simulation purposes, we can see that the two eigenproblem formulations provide highly different convergence-rates.

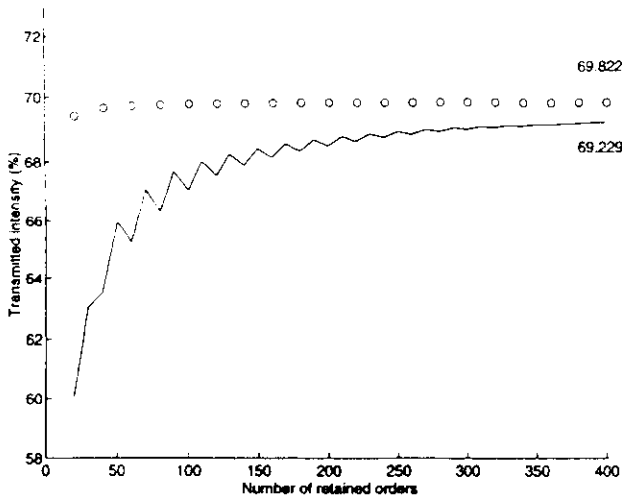


Fig. 2. Diffraction efficiency of the transmitted zeroth order of a metallic grating with TM polarized light. The solid curve is obtained by using the conventional eigenproblem formulation of Eq. (6b). The circles are provided by the new eigenproblem of Eq. (9). The grating parameters and the geometry problem are defined in Fig. 2 of Ref. 6.

#### 4. NUMERICAL EXAMPLES

In our implementation of the new eigenproblem formulation, we form matrices  $\mathbf{A}$  and  $\mathbf{E}$  by directly using the analytical values of the harmonic coefficients  $\epsilon_m$  and  $a_m$ . Matrices  $\mathbf{A}$  and  $\mathbf{E}$  are then inverted, and the eigenproblem of Eq. (9) is solved with standard library programs. When  $(2M + 1)$  orders are retained in the computation, we obtain  $(2M + 1)$  eigenvectors  $\mathbf{u}_i$  and  $(2M + 1)$  eigenvalues  $\lambda_i^2$ . Using Eq. (7a), we derive the  $2(2M + 1)$  eigenvectors  $[\mathbf{u}_i^{\lambda_i, \mathbf{A} \mathbf{u}_i}]$  and  $[\mathbf{u}_i^{-\lambda_i, \mathbf{A} \mathbf{u}_i}]$  with eigenvalues  $\lambda_i$  and  $-\lambda_i$ , respectively. The first numerical example is related to a metallic lamellar grating deposited on a glass substrate, which acts as a polarizer in the visible region. It was provided by Peng and Morris in Ref. 6. The lamellar grating is composed of chrome (index of refraction equals  $3.18 - j4.41$ ) and air and is acting as a zeroth-order filter for normal incidence (see the caption of Fig. 2 in Ref. 6 for more details). Figure 2 shows the transmitted intensity of the zeroth order as a function of the number of retained orders. The solid curve is obtained by solving the eigenproblem of Eq. (6b). A detailed explanation of the algorithm implementation can be found in Ref. 6. Note that a slow and oscillating convergence is obtained. The amplitude of the oscillations decreases as the number of retained orders increases. The circles are obtained by solving the eigenproblem of Eq. (9). No oscillation is observed. We are grateful to Mike Miller at the Institut d'Optique Théorique et Appliquée in Orsay, who computed for us the zeroth-order transmitted diffraction efficiency using his modal method.<sup>7</sup> He found a transmitted intensity of 70.28% when retaining 90 modes in his numerical computation. If we consider that 70.28% is the exact diffraction efficiency, it is clear from Fig. 2 that the new eigenproblem formulation with as few as 20 retained orders provides a more accurate result than the conventional formulation with 400 retained orders.

The second numerical example is taken from Ref. 1,

where the convergence rate of a highly conductive grating on gold substrate was investigated (see Fig. 1 in Ref. 1 for additional details on the grating geometry). The diffraction configuration is a  $30^\circ$  incident angle, which corresponds to the first-order Bragg condition. Only the negative first and zeroth reflected orders are propagating. Figure 3 shows the diffraction efficiencies of the negative first and zeroth orders when the new eigenproblem of Eq. (9) is used for the numerical computation. As the same scale is used in Fig. 3 of this publication and in Figs. 3(a) and 3(b) of Ref. 1, a visual comparison of the convergence rates can be made. It is obvious that the convergence rate is drastically improved in that particularly stringent example. For example, when 51 orders are retained for the computation, the conventional eigenproblem provides diffraction efficiencies of 25% and 55% for the negative first and zeroth orders, respectively. With the new formulation, the diffraction efficiencies are 10% and 84%. In Fig. 3 the numerical value of the diffraction efficiencies obtained with 25, 51, 75, and 125 retained orders are given. They can be compared with the exact values 84.843% and 10.162%, obtained by Li and Haggans,<sup>1</sup> when 125 modes are retained in the modal decomposition of the field. When only 25 orders are retained with the new eigenproblem, the diffraction-efficiency differences between the modal method and the new eigenproblem formulation are less than 0.009 for the reflected zeroth order and 0.002 for the negative first order (relative errors less than 1% and 2%, respectively). We conclude that the new eigenproblem formulation provides highly improved convergence rates.

The improved convergence rates illustrated in Figs. 2 and 3 are not isolated cases. All our simulation results show an improvement even for dielectric and nonlamellar gratings and for small or large period-to-wavelength ratios.

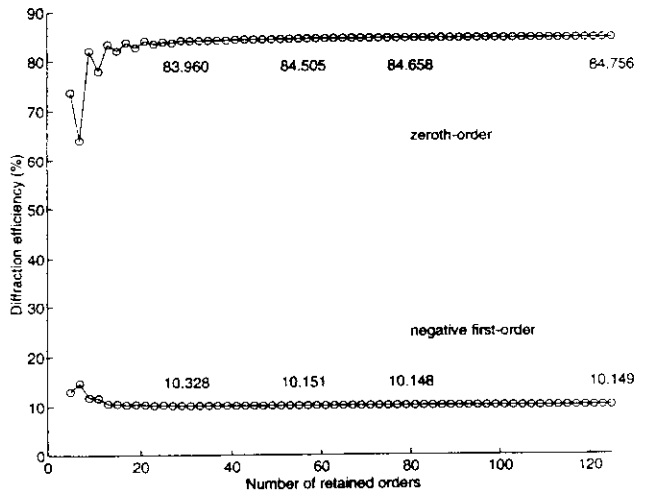


Fig. 3. Diffraction efficiencies of the reflected negative first and zeroth orders of a metallic grating with TM polarization. The circles are provided by the new eigenproblem method of Eq. (9). The grating parameters and the geometry problem are defined in Fig. 1 of Ref. 1. A direct comparison can be applied with Figs. 3(a) and 3(b) of Ref. 1, where simulation results obtained with the conventional eigenproblem and modal methods are presented.

## 5. INTERPRETATION

In this section we give a simple interpretation of the convergence-rate differences between the conventional and the new eigenproblem formulations. The interpretation is given in the quasi-static limit, i.e., when the period-to-wavelength ratio tends to zero. We show that, with the conventional eigenproblem formulation, an adequate description of the quasi-static limit requires that an infinite number of orders be retained in the computation. We also show that, with the new eigenproblem formulation, the quasi-static limit is accurately described with a finite number of retained orders.

As was shown by Li and Haggans,<sup>1</sup> the convergence of RCWA is directly related to the convergence of the eigensolution. Therefore any method to improve the convergence of the eigenproblem should improve the convergence of the diffraction efficiencies. Among the eigenvalues there is at least one that can be interpreted physically. It takes advantage of the equivalence between gratings and homogeneous media in the quasi-static limit. By quasi-static limit we mean situations for which the grating period is infinitely small compared with the wavelength. The equivalence was rigorously derived by Bouchitte and Petit.<sup>8</sup>

For the sake of simplicity we restrict the discussion to normal incidence. In the quasi-static limit and for TM polarization, the grating is equivalent to a thin layer with an effective relative permittivity equal to  $1/a_0$ , where  $a_0$  is the zeroth Fourier coefficient of  $\epsilon_0/\epsilon(x)$ . The field in the grating can be written as a linear combination of two counterpropagating plane waves, namely,  $\exp jk_0\sqrt{1/a_0}z$  and  $\exp -jk_0\sqrt{1/a_0}z$ . These two plane waves must be solutions of Eqs. (5) and (8). So in the quasi-static limit,  $-k_0^2/a_0$  must be an eigenvalue of Eqs. (5) and (8). Let us note  $U_m' = -jk_0nU_m$  and  $S_m' = -jk_0nS_m$ , where  $-k_0^2n^2$  is the degenerated eigenvalue expected to be equal to  $-k_0^2/a_0$ .

Let us first start with the conventional eigenproblem formulation. In the quasi-static limit, i.e., when  $K/k_0$  tends to infinity, Eq. (4) reduces to

$$nS_0^{(0)} = U_0^{(0)}, \quad (10a)$$

$$\forall m \neq 0, \quad \sum_p p a_{m-p} U_p^{(0)} = 0, \quad (10b)$$

where superscript (0) holds for the quasi-static notation of the fields and  $\beta$  was taken equal to zero in Eq. (4). If  $\pm M$  orders are retained in the computation, Eq. (10b) provides a homogeneous system of  $2M$  linear equations with  $2M$  unknowns,  $U_p^{(0)}$  with  $p \neq 0$ . Except for a possible unexpected degeneracy, the solutions are zeros. So in the quasistatic limit Eq. (3b) becomes

$$\forall m \neq 0, \quad \sum_{p \neq 0} \epsilon_{m-p} S_p^{(0)} = -\epsilon_m S_0^{(0)}, \quad (11a)$$

$$nU_0^{(0)} = \sum_{p \neq 0} \epsilon_{-p} S_p^{(0)} + \epsilon_0 S_0^{(0)}. \quad (11b)$$

Multiplying both sides of Eq. (11a) by  $a_{-m}$  and then summing over all  $m$ , we obtain

$$\sum_{p \neq 0} \left( \sum_{m \neq 0} a_{-m} \epsilon_{m-p} \right) S_p^{(0)} = - \sum_{m \neq 0} a_{-m} \epsilon_m S_0^{(0)}. \quad (12a)$$

When an infinite number of orders is retained, because  $\sum_m a_{-m} \epsilon_{m-p} = 0$  as  $\mathbf{E}$  and  $\mathbf{A}$  are inverse matrices, the left-hand side of Eq. (12a) reduces to  $-a_0 \sum_{p \neq 0} \epsilon_{-p} S_p^{(0)}$ . When we truncate the number of orders and retain  $\pm M$  orders in the computation, this is no longer true, and we note the left side of Eq. (12a)  $-a_0^* \sum_{p \neq 0} \epsilon_{-p} S_p^{(0)}$ . Similarly, the right-hand side of Eq. (12a) can be written as  $-(1 - a_0 \epsilon_0) S_0^{(0)}$  when an infinite number of orders are retained and is noted as  $-(1 - a_0^* \epsilon_0) S_0^{(0)}$  during truncating. So Eq. (12a) can be written as

$$-a_0^* \sum_{p \neq 0} \epsilon_{-p} S_p^{(0)} = -(1 - a_0^* \epsilon_0) S_0^{(0)}. \quad (12b)$$

In Eq. (12b),  $a_0^*$  and  $a_0^*$  denote two slightly different values of  $a_0$ , which depend on the truncation rank  $M$ .  $a_0^*$  and  $a_0^*$  tend to  $a_0$  when the number of retained orders tends to infinity. By eliminating  $\sum_{p \neq 0} \epsilon_{-p} S_p^{(0)}$  between Eqs. (11b) and (12b), we obtain

$$a_0^* (nU_0^{(0)} - \epsilon_0 S_0^{(0)}) = (1 - a_0^* \epsilon_0) S_0^{(0)}. \quad (13)$$

Using Eq. (10a) to substitute  $S_0^{(0)}$  for  $U_0^{(0)}$  in Eq. (13), and looking for a nonzero solution in  $S_0^{(0)}$ , we obtain

$$n^2 = \frac{1}{a_0^*} + \epsilon_0 \left( 1 - \frac{a_0^*}{a_0^*} \right). \quad (14)$$

Equation (14) shows that an infinite number of orders must be retained for the numerical computation of the exact eigenvalue  $-k_0^2 n^2 = -k_0^2/a_0$ . The effect of the truncation is not negligible. Figure 4 shows the real and the imaginary parts of the absolute error  $e = n - \sqrt{1/a_0}$  as a function of the number of retained orders. It was obtained by solving the system of Eqs. (10) and (11) for the problem of Fig. 2. The error  $e$  is quite large even when 200 orders are retained, especially for thick gratings for which a small error on  $n$  is responsible for a large error on

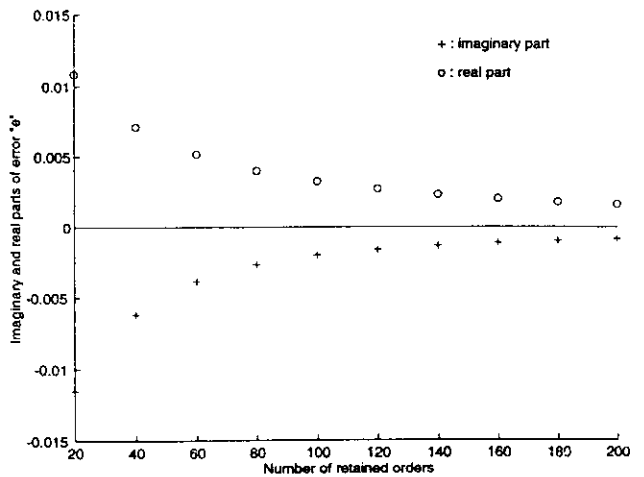


Fig. 4. Effect of the truncation on the accuracy of the conventional eigenproblem. The pluses and circles correspond to the imaginary and the real parts, respectively, of the error  $e = n - \sqrt{1/a_0}$ . The results are obtained by solving the system of linear equations defined by Eqs. (10a) and (11).

$\exp(-jknz)$  when the boundary conditions at the grating interface are being matched.

Let us now consider the quasi-static-limit situation with the new eigenproblem formulation. For normal incidence ( $\beta = 0$ ) the system of second-order differential equations given by Eq. (8) is not valid. This is because  $f_0 = 0$  for normal incidence. It is easily shown that Eq. (8) must be replaced by

$$\forall m \neq 0, \quad m \frac{K^2}{k_0^2} U_m - \sum_{p \neq 0} \frac{\epsilon_{m-p}}{p} U_p = \sum_{p \neq 0, l} \frac{\epsilon_{m-p} a_{p-l}}{p k_0^2} U_l'', \quad (15a)$$

$$k_0^2 U_0 + \sum_p a_{-p} U_p'' = 0. \quad (15b)$$

Equations (15a) and (15b) constitute the set of second-order differential equations for normal incidence. Proceeding to the quasi-static limit in Eq. (15a) results in  $U_m^{(0)} = 0$  for any nonzero  $m$ .  $n^2$  then becomes  $1/a_0$  in Eq. (15b); this result holds for any number of retained orders.

For nonnormal incidence, a similar argument can be provided. The eigenvalue of the quasi-static limit must be equal to  $-k_0^2(\sin^2 \theta/\epsilon_0 + a_0 \cos^2 \theta)^{-1}$  instead of  $-k_0^2/a_0$ ; this is because, in the quasi-static limit, the equivalent homogeneous medium is uniaxial, with the optic axis parallel to the  $x$  axis (see Ref. 8). The faster convergence rate of the eigenvalue problem defined by Eq. (8) was justified only in the quasi-static limit. For nonzero period-to-wavelength ratios and for TM polarization, although the eigenvalues are more difficult to interpret, it is possible to derive an eigenvalue that approximately satisfies the eigenproblem.<sup>9</sup> This approximate solution is expressed as a power series of  $\Lambda/\lambda$ . It is clear that the power series' zeroth order, which corresponds to the quasi-static limit, is given by  $-k_0^2/a_0$ . The result is that the conventional eigenproblem formulation, which is able to provide the zeroth-order term only when an infinite number of orders are retained in the computation, is also inadequate for accurately describing the eigenproblem of gratings with nonzero period-to-wavelength ratios. Although the derivation given in this section is restricted to the quasi-static limit, we believe that it provides good insight for understanding the improved convergence rates for nonzero period-to-wavelength ratios.

## 6. GENERALIZATION TO CONICAL MOUNTINGS

The new eigenproblem formulation can be generalized in a straightforward way to the case of conical mountings. We have to interpret the conical diffraction eigenproblem as a combination of TE and TM polarization eigenproblems, and we note that the conventional TE eigenproblem formulation<sup>3</sup> must not be changed since it provides good convergence rates. Also note that the conventional formulation for TE polarization, like the new formulation for TM polarization, provides the adequate eigenvalue in the quasi-static limit for any number of retained orders. Using strictly the notation of Ref. 3, it is then easily shown that a useful eigenproblem formulation is

$$k_0^{-1} \begin{bmatrix} S_y' \\ S_x' \\ U_y' \\ U_x' \end{bmatrix} = \begin{bmatrix} 0 & 0 & K_y E^{-1} K_x & I - K_y E^{-1} K_y \\ 0 & 0 & K_x E^{-1} K_x & -K_x E^{-1} K_y \\ K_x K_y & A^{-1} - K_y^2 & 0 & 0 \\ K_x^2 - E & -K_x K_y & 0 & 0 \end{bmatrix} \times \begin{bmatrix} S_y \\ S_x \\ U_y \\ U_x \end{bmatrix}. \quad (16)$$

In Eq. (16)  $S_x$ ,  $S_y$ ,  $U_x$ ,  $U_y$ ,  $K_x$ ,  $K_y$ ,  $E$ , and  $I$  are defined as in Ref. 3.  $A$  denotes again the matrix formed by the inverse-permittivity harmonic coefficients. The only difference between the conventional formulation [see Eq. (57) of Ref. 3] and the new formulation of Eq. (16) is in the third row of the second column, where matrix  $E - K_y^2$  has been replaced by  $A^{-1} - K_y^2$ . In Fig. 5 the diffraction efficiencies of the negative first and zeroth orders of a conical mounting are shown as functions of the number of retained orders. The grating used to obtain the result in Fig. 5 is the same as that discussed in the second example of Section 4 (see Fig. 1 of Ref. 1). The diffraction configuration is a  $30^\circ$  angle of incidence, a  $30^\circ$  azimuthal angle, and a  $45^\circ$  angle between the electric-field vector and the plane of incidence. Using the notation of Ref. 3,  $\theta = 30^\circ$ ,  $\phi = 30^\circ$ , and  $\psi = 45^\circ$ . The solid curves are obtained with the conventional formulation of Eq. (57) in Ref. 3, and the dotted curves are obtained with the new formulation of Eq. (16). As in the two examples above we note that the new formulation provides faster and smoother convergence rates. For example, for the zeroth-order diffracted plane wave the numerical values of the diffraction efficiencies are 10.58%, 10.11%, 10.08%,

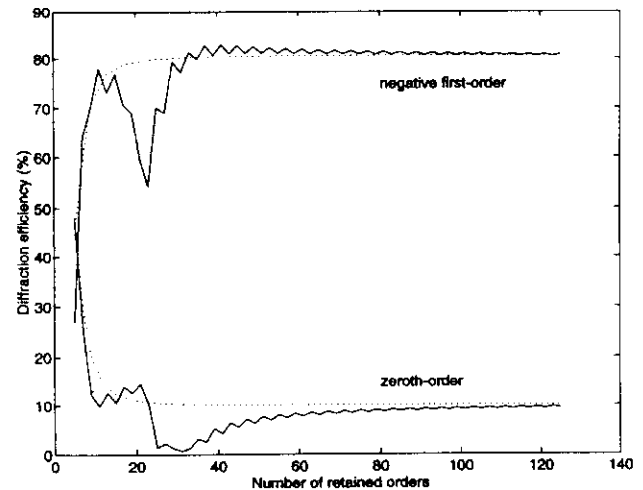


Fig. 5. Diffraction efficiencies of the reflected negative first and zeroth orders of a metallic grating for conical mount ( $\theta = 30^\circ$ ,  $\phi = 30^\circ$ , and  $\psi = 45^\circ$ ). The grating parameters are defined in Fig. 1 of Ref. 1. The solid curves are obtained with the conventional eigenproblem formulation of Ref. 3. The dotted curves are obtained with the new formulation of Eq. (16).

and 10.07% when 25, 51, 75, and 125 orders, respectively, are retained with the new formulation. With the conventional formulation the corresponding diffraction efficiencies are 1.38%, 7.70%, 8.99%, and 9.42%. We conclude that the new formulation with 25 retained orders provides more accurate results than the conventional formulation with 125 retained orders. By use of the second derivative of the field vector, the eigenproblem of Eq. (16) reduces to

$$\begin{aligned} k_0^{-2}[\mathbf{U}_x''] &= [\mathbf{K}_y^2 + \mathbf{K}_z^2 - \mathbf{E}][\mathbf{U}_x], \\ k_0^{-2}[\mathbf{S}_x''] &= [\mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_x \mathbf{A}^{-1} + \mathbf{K}_y^2 - \mathbf{A}^{-1}][\mathbf{S}_x]. \end{aligned} \quad (17)$$

Equations (17) are new formulations of Eq. (60) in Ref. 3 and can be used to save computational time.

## 7. CONCLUSION AND DISCUSSION

By reformulating the eigenproblem of RCWA, we show that good convergence rates can be achieved for TM polarization of 1-D metallic gratings. In Ref. 1, Li and Haggans interpreted the oscillating and poor convergence rates of conventional RCWA by invoking truncation effects that are due to the slowly convergent Fourier expansions of the permittivity and the field inside the grating, but they noted that their interpretation poses a difficulty in understanding why convergence rates are much slower with TM than with TE polarization. Because the new eigenproblem is also based on a truncated Fourier expansion of the permittivity, the poor convergence rates observed for TM polarization must not be attributed to truncation effects. In Section 5, by examining the eigenproblem in the quasi-static limit, we show that the conventional eigenproblem requires an infinite Fourier expansion to provide an accurate description of the quasi-static limit diffraction problem; this can be considered to be a kind of bad conditioning of the conventional eigenproblem. However, as shown in Fig. 3, the effect of the truncation remains slightly visible with the new eigenproblem formulation. When the number of retained orders increases from 25 to 125, the zeroth-order diffraction efficiency keeps increasing from 83.96% to 84.76% and is expected ultimately to reach the approximate value of 84.84%. This convergence rate is similar to that observed for TE polarization of the same grating problem. The approach developed in this paper can be applied to any numerical techniques using a Fourier expansion and is not restricted to the implementation of RCWA.

With respect to computational effort, the new eigenproblem formulations of Eqs. (9) and (17) are more demanding than their corresponding conventional formulations [Eqs. (6b) and (60) of Ref. 3]. They additionally require the numerical computation of matrices  $\mathbf{A}$  and  $\mathbf{A}^{-1}$ . However, for a given reasonable accuracy, the new eigenproblem formulation saves considerable time and computer memory because fewer orders have to be retained. This is especially true when continuous profile gratings or stacks of lamellar gratings are considered or when several grating depths are studied for a given diffraction problem.

## ACKNOWLEDGMENTS

When this work was completed, Philippe Lalanne was a visiting scientist at The Institute of Optics of the University of Rochester. He is pleased to acknowledge the Direction Générale de l'Armement for financial support under contract DRET-DGA 94-1123. This work was also supported in part by the U.S. Army Research Office.

## REFERENCES

1. L. Li and C. W. Haggans, "Convergence of the coupled-wave method for metallic lamellar diffraction gratings," *J. Opt. Soc. Am. A* **10**, 1184-1189 (1993).
2. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of grating diffraction—E-mode polarization and losses," *J. Opt. Soc. Am.* **73**, 451-455 (1983).
3. M. G. Moharam, E. B. Grann, D. A. Pommet, and T. K. Gaylord, "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," *J. Opt. Soc. Am. A* **12**, 1068-1086 (1995).
4. C. B. Burckhardt, "Diffraction of a plane wave at a sinusoidally stratified dielectric grating," *J. Opt. Soc. Am.* **56**, 1502-1509 (1966).
5. K. Knop, "Rigorous diffraction theory for transmission phase gratings with deep rectangular grooves," *J. Opt. Soc. Am.* **68**, 1206-1210 (1978).
6. S. Peng and G. M. Morris, "Efficient implementation of rigorous coupled-wave analysis for surface relief gratings," *J. Opt. Soc. Am. A* **12**, 1087-1096 (1995).
7. J. M. Miller, J. Turunen, E. Noponen, A. Vasara, and M. R. Taghizadeh, "Rigorous modal theory for multiply grooved lamellar gratings," *Opt. Commun.* **111**, 526-535 (1994).
8. G. Bouchitte and R. Petit, "Homogenization techniques as applied in the electromagnetic theory of gratings," *Electromagnetics* **5**, 17-36 (1985).
9. Ph. Lalanne and D. Lemerrier-Lalanne, "On the effective medium theory of subwavelength periodic structures," submitted to *J. Mod. Opt.*

# Use of Fourier series in the analysis of discontinuous periodic structures

Lifeng Li

Optical Sciences Center, University of Arizona, Tucson, Arizona 85721

Received November 20, 1995; accepted March 5, 1996; revised manuscript received April 4, 1996

The recent reformulation of the coupled-wave method by Lalanne and Morris [J. Opt. Soc. Am. A 13, 779 (1996)] and by Granet and Guizal [J. Opt. Soc. Am. A 13, 1019 (1996)], which dramatically improves the convergence of the method for metallic gratings in TM polarization, is given a firm mathematical foundation in this paper. The new formulation converges faster because it uniformly satisfies the boundary conditions in the grating region, whereas the old formulations do so only nonuniformly. Mathematical theorems that govern the factorization of the Fourier coefficients of products of functions having jump discontinuities are given. The results of this paper are applicable to any numerical work that requires the Fourier analysis of products of discontinuous periodic functions. © 1996 Optical Society of America.

## 1. INTRODUCTION

The determination of the eigensolutions of Maxwell's equations in a periodic, piecewise-constant medium, as shown in Fig. 1, is the most crucial step in the analysis of surface-relief gratings by modal methods. Among the existing modal methods, the most popular one is the modal method by Fourier expansion,<sup>1,2</sup> commonly referred to as the coupled-wave method (CWM). In the CWM, both the electromagnetic fields and the permittivity function are expanded into Fourier series, and thereby the boundary-value problem is reduced to an algebraic eigenvalue problem. In an earlier paper<sup>3</sup> Li and Haggans provided strong numerical evidence to show that the CWM converged slowly for metallic gratings in TM polarization. The authors attributed the slow convergence of the CWM to the slow convergence of the Fourier expansions. However, they also admitted that "the convergence-rate difference [between TE and TM] cannot be completely explained by such a simplistic convergence analysis of the Fourier expansions" (p. 1188). Recently Lalanne and Morris<sup>4</sup> and Granet and Guizal<sup>5</sup> numerically achieved truly dramatic improvement in the convergence rate for TM polarization by reformulating the algebraic eigenvalue problem of the CWM. Their work convincingly proved that the cause of the slow convergence of the CWM for TM polarization is not the use of the Fourier series but the way in which the Fourier series of the permittivity and the reciprocal permittivity functions are used.

Whenever a  $\Sigma$  sign is used in this paper without the summation range explicitly given, a sum from  $-M$  to  $M$  is understood. Similarly, a matrix without an indication of its dimension is understood to be a  $(2M + 1) \times (2M + 1)$  square matrix. The Gaussian system of units, the coordinate system of Fig. 1, and the time dependence  $\exp(-i\omega t)$  are used.

In the old formulation,<sup>1,2</sup> one solves the coupled first-order differential system,

$$\frac{1}{i} \frac{dH_{zn}}{dy} = -k_0 \sum_m \epsilon_{n-m} E_{xm}, \quad (1a)$$

$$\frac{1}{i} \frac{dE_{zn}}{dy} = -k_0 \mu_0 H_{zn} + \frac{\alpha_n}{k_0} \sum_m \left( \frac{1}{\epsilon} \right)_{n-m} \alpha_m H_{zm}, \quad (1b)$$

or better yet, the equivalent second-order system

$$\frac{d^2 H_{zn}}{dy^2} = \sum_m \epsilon_{n-m} \sum_p \left[ \alpha_m \left( \frac{1}{\epsilon} \right)_{m-p} \alpha_p - \mu_0 k_0^2 \delta_{mp} \right] H_{zp}. \quad (2)$$

Here,  $k_0$  is the vacuum wave number;  $\mu_0 = 1$ ;  $\delta_{mp}$  is the Kronecker symbol;  $\epsilon_n$  and  $(1/\epsilon)_n$  are the Fourier coefficients of the permittivity and the reciprocal permittivity functions, respectively;  $E_{zn}$  and  $H_{zn}$  are the  $y$ -dependent Fourier coefficients of the fields; and  $\alpha_n = \alpha_0 + nK$ , with  $K = 2\pi/d$  and  $\alpha_0$  being the Floquet exponent. In the new formulation,<sup>4,5</sup> one solves the coupled first-order system,

$$\frac{1}{i} \frac{dH_{zn}}{dy} = -k_0 \sum_m \left[ \frac{1}{\epsilon} \right]_{nm}^{-1} E_{xm}, \quad (3a)$$

$$\frac{1}{i} \frac{dE_{zn}}{dy} = -k_0 \mu_0 H_{zn} + \frac{\alpha_n}{k_0} \sum_m [\epsilon]_{nm}^{-1} \alpha_m H_{zm}, \quad (3b)$$

or the second-order system,

$$\frac{d^2 H_{zn}}{dy^2} = \sum_m \left[ \frac{1}{\epsilon} \right]_{nm}^{-1} \sum_p (\alpha_m [\epsilon]_{mp}^{-1} \alpha_p - \mu_0 k_0^2 \delta_{mp}) H_{zp}, \quad (4)$$

where  $[f]$  denotes the Toeplitz matrix generated by the Fourier coefficients of  $f$  such that its  $(n, m)$  entry is  $f_{n-m}$ , and  $-1$  denotes the matrix inverse. Thus the only difference between the new and the old formulations is the manner in which the permittivity function appears in the equations: The new formulation uses  $[1/\epsilon]^{-1}$  and  $[\epsilon]^{-1}$  instead of  $[\epsilon]$  and  $[1/\epsilon]$ , respectively. It should be mentioned that there is another version of the old formulation, recently presented by Moharam *et al.*,<sup>6</sup> in which the matrix  $[1/\epsilon]$  in Eq. (2) is replaced by  $[\epsilon]^{-1}$ .

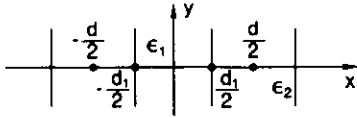


Fig. 1. Periodic, piecewise-constant medium. The periodicity of the permittivity is  $d$ , and its discontinuities are located at  $x = \pm d/2$ .

$$\frac{d^2 H_{zn}}{dy^2} = \sum_m \epsilon_{n-m} \sum_p (\alpha_m [\epsilon]_{mp}^{-1} \alpha_p - \mu_0 k_0^2 \delta_{mp}) H_{2p}. \quad (5)$$

The close similarity in equation structure and the striking difference in performance between the old and the new formulations poses an intriguing question: What is the fundamental difference between the two formulations? The authors of Refs. 4 and 5 did not provide any answer, although the former offered an ingenious demonstration of the plausibility of the new formulation in the quasi-static limit. They also did not say how they discovered the new equations. Indeed, their discovery appears empirical.

In this paper I show that the reason for the success of the new formulation is that it uniformly preserves the continuity of the appropriate field components across the discontinuities of the permittivity function; by inference, the old formulations do so only nonuniformly. I will provide the mathematical basis for the new formulation. Furthermore, I will describe the correct procedures for Fourier analyzing the electromagnetic-field components in Maxwell's equations such that the required field continuity is preserved across the discontinuities of the permittivity function.

In Section 2 I give three mathematical theorems concerning the Fourier factorization of a product of two periodic functions. The contents of these theorems are rather subtle, but they have extremely important implications to the theory of gratings. The proofs of the theorems will not be given here because they are lengthy. The reader who is interested in the proofs may refer to Ref. 7. To help the reader better understand the abstract mathematical results, some discussions and several graphical illustrations are given in the latter part of Section 2. The mathematical results of Section 2 are applied to our grating problem in Section 3, where Eqs. (3) and (4) are derived and Eqs. (1), (2), and (5) are proven to be incorrect. The correct procedures for Fourier analyzing Maxwell's equations such that the field continuity is preserved are also established in Section 3. In Section 4 I make some remarks on the results obtained from this research.

## 2. STATEMENT AND ILLUSTRATION OF THE MATHEMATICAL RESULTS

### A. Notation and Statement of the Problem

Let  $\mathbf{P}$  be the set of piecewise-continuous, piecewise-smooth, bounded, periodic functions of  $x$  with period  $2\pi$ . For every  $f(x) \in \mathbf{P}$  and  $g(x) \in \mathbf{P}$ ,

$$h(x) = f(x)g(x) \quad (6)$$

is obviously also in  $\mathbf{P}$ . Let

$$U_f = \{x_j | f(x_j + 0) \neq f(x_j - 0), \quad j = 1, 2, \dots\} \quad (7)$$

be the set of the abscissas of the discontinuities of  $f(x)$ , and let  $U_g$  be similarly defined for  $g(x)$ . Then,

$$U_{fg} = U_f \cap U_g \quad (8)$$

is the set of the abscissas of the concurrent discontinuities of  $f(x)$  and  $g(x)$ . If  $h(x)$  is such that

$$h(x_p - 0) = h(x_p + 0) \quad (x_p \in U_{fg}), \quad (9)$$

$f(x)$  and  $g(x)$  are said to have a pair of complementary jumps at  $x_p$ . In this case the discontinuity of  $h(x)$  at  $x_p$  is removable. The amount of discontinuity of  $f$  at  $x_j$  will be denoted by  $\hat{f}_j$ ,

$$\hat{f}_j = f(x_j + 0) - f(x_j - 0), \quad (10)$$

and similarly the jump of  $g$  at  $x_j$  by  $\hat{g}_j$ . If we assign the functional values of  $f(x)$ ,  $g(x)$ , and  $h(x)$  at their respective discontinuities to be the arithmetic means of their limiting values from the two sides of the discontinuities, then these functions are represented everywhere by their Fourier series. As in Section 1, a function name with a subscript in lowercase letter is used to denote the complex Fourier coefficients of the function. The term Fourier factorization means the expression of  $h(x)$  or its Fourier coefficients in terms of the Fourier coefficients of  $f(x)$  and  $g(x)$ .

For a large class of functions, including those in  $\mathbf{P}$ , the Fourier coefficients of  $h(x)$  can be obtained from the Fourier coefficients of  $f(x)$  and  $g(x)$  by Laurent's rule:<sup>8</sup>

$$h_n = \sum_{m=-\infty}^{+\infty} f_{n-m} g_m. \quad (11)$$

The Fourier factorization of  $h(x)$  is then given by

$$\begin{aligned} h(x) &= \sum_{n=-\infty}^{+\infty} h_n \exp(inx) \\ &= \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} f_{n-m} g_m \exp(inx). \end{aligned} \quad (12)$$

To be more precise, Eq. (12) should be understood in the following sense:

$$h(x) = \lim_{N \rightarrow \infty} \sum_{n=-N}^N \left( \lim_{M \rightarrow \infty} \sum_{m=-M}^M f_{n-m} g_m \right) \exp(inx). \quad (13)$$

The above equation, in the way it is written, emphasizes two important points. First, the two limits are independent of each other and the inner limit is to be taken first. Second, the upper and lower bounds in each sum should tend to infinity simultaneously; in other words, the sums converge in general only restrictedly.<sup>9</sup>

In solving a practical problem on a computer, the truncation of the infinite series is inevitable. In this section subscript  $M$  or superscript  $M$  enclosed in parentheses will be used to denote the symmetrically truncated partial sums. Then, corresponding to Eqs. (11) and (12), we have



Laurent's rule: 
$$h_n^{(M)} = \sum_{m=-M}^M f_{n-m} g_m, \quad (14)$$

$$h^{(M)}(x) = \sum_{n=-M}^M h_n^{(M)} \exp(inx), \quad (15)$$

$$h_M(x) = \sum_{n=-M}^M h_n \exp(inx). \quad (16)$$

Note that in Eq. (15) the same positive integer  $M$  is used both for the summation bounds and for the superscript of the coefficients, which is the most commonly adopted truncation convention in numerical analysis. This condition is of fundamental importance to the validity of the theorems to be given below. What a practitioner hopes is that  $h^{(M)}(x)$  converges as  $M \rightarrow \infty$  and that

$$h^{(\infty)}(x) = h(x). \quad (17)$$

Although the mathematical theory on the multiplication of Fourier series is well developed,<sup>9</sup> to the best of my knowledge the special and practically important problem that is posed by letting  $N$  and  $M$  in Eq. (13) tend to infinity simultaneously has not been addressed in the literature.

**B. Theorems of Fourier Factorization**

*Theorem 1.* If  $f(x) \in \mathbf{P}$  and  $g(x) \in \mathbf{P}$  have no concurrent jump discontinuities and  $h_n^{(M)}$  is given by Eq. (14), then Eq. (17) is valid.

*Theorem 2.* If  $f(x) \in \mathbf{P}$  and  $g(x) \in \mathbf{P}$  have concurrent jump discontinuities and  $h_n^{(M)}$  is given by Eq. (14), then

$$h^{(M)}(x) = h_M(x) - \sum_{x_p \in U_{fg}} \frac{\hat{f}_p \hat{g}_p}{2\pi^2} \Phi_M(x - x_p) + o(1), \quad (18)$$

where the term  $o(1)$  uniformly tends to zero, and

$$\Phi_M(x) = \sum_{n=1}^M \frac{\cos nx}{n} \sum_{m>M} \frac{1}{m-n}. \quad (19)$$

Furthermore,

$$\lim_{M \rightarrow \infty} \Phi_M(x) = 0 \quad (x \neq 0), \quad (20)$$

but

$$\lim_{M \rightarrow \infty} \Phi_M(0) = \frac{\pi^2}{4}. \quad (21)$$

*Theorem 3.* Let  $S$  be a subinterval or a collection of subintervals of  $[0, 2\pi)$ , and  $\bar{S}$  be its complement ( $S$  or  $\bar{S}$  may be empty). We assume that  $f(x) \neq 0$  and denote by  $\llbracket 1/f \rrbracket^{(M)}$  the symmetrically truncated Toeplitz matrix generated by the Fourier coefficients of  $1/f$ . If all the discontinuities of  $h(x)$  are removable and if  $f(x)$  satisfies either one of the two following conditions: (a)  $\text{Re} \{1/f\}$  does not change sign in  $[0, 2\pi)$ ,  $\text{Re} \{1/f\} \neq 0$  in  $S$ , and  $\text{Im} \{1/f\}$  does not change sign in  $\bar{S}$ ; (b)  $\text{Im} \{1/f\}$  does not change sign in

$[0, 2\pi)$ ,  $\text{Im} \{1/f\} \neq 0$  in  $S$ , and  $\text{Re} \{1/f\}$  does not change sign in  $\bar{S}$ —then Eq. (17) is valid provided that, instead of Eq. (14), the inverse rule

$$\text{Inverse Rule: } h_n^{(M)} = \sum_{m=-M}^M \left[ \frac{1}{f} \right]_{nm}^{(M)-1} g_m \quad (22)$$

is used in Eq. (15).

**C. Discussion**

In less formal language, theorem 1 says that if  $f$  and  $g$  have no concurrent jumps, then the difference between  $h_M(x)$ , the partial sum of the Fourier series that uses the exact Fourier coefficients, and  $h^{(M)}(x)$ , the partial sum that uses the approximate Fourier coefficients obtained by the finite Laurent rule, vanishes everywhere as the orders of the partial sums increase. Theorem 3 says that the same is true if all the jumps of  $f$  and  $g$  are pairwise complementary provided that, instead of Laurent's rule, the inverse multiplication rule is used. However, theorem 2 says that if  $f$  and  $g$  have concurrent jumps and Laurent's rule is used, then the difference between the two partial sums does not vanish everywhere; at the locations of the concurrent jumps,  $h^{(M)}(x)$  refuses to converge to  $h_M(x)$ .

As a manifestation of the nonconvergence of  $h^{(M)}(x)$  to  $h_M(x)$  at  $x_p \in U_{fg}$ , the convergence of  $h^{(M)}(x)$  to  $h_M(x)$  in the neighborhood of  $x_p$  is nonuniform. In other words, for any  $\epsilon > 0$ , one cannot find an  $M^*$  such that  $|h^{(M)}(x) - h_M(x)| < \epsilon$  not only for all  $M > M^*$  but also for all  $x \in (x_p - \delta, x_p) \cup (x_p, x_p + \delta)$ , where  $\delta > 0$  is a constant. From Eq. (18) the convergence of  $h^{(M)}(x) - h_M(x)$  is equivalent to the convergence of  $\Phi_M(x)$ . The nonuniform convergence of  $\Phi_M(x)$  can be easily seen because the sum of a uniformly convergent infinite series of continuous terms should be a continuous function. Since  $\Phi_x(x)$  is discontinuous at  $x = 0$ , the convergence of  $\Phi_M(x)$  cannot be uniform in the neighborhood of  $x = 0$ .

The function  $\Phi_M(x)$  has many interesting properties. Its limit as  $M \rightarrow \infty$  is  $\pi^2/4$  at  $x = 0$  and zero everywhere else in  $[0, 2\pi)$ .  $\Phi_M(x)$  is unique in the sense that if there is another function,  $\Phi'_M(x)$  that satisfies Eq. (18), then the difference between  $\Phi_M(x)$  and  $\Phi'_M(x)$  must converge uniformly to zero everywhere. A few graphs of  $\Phi_M(x)$  will help the reader to see its general behavior. Figures 2(a), 2(b), and 2(c) are graphs of  $\Phi_M(x)$  in the neighborhood of  $x = 0$  for  $M = 10, 100,$  and  $1000$ , respectively. Note that although the same vertical scale is used in all three graphs, the horizontal scales are different from one another by a factor of 10. Although there are visible minor differences between the two curves in Figs. 2(a) and 2(b), no differences between Figs. 2(b) and 2(c) can be easily detected. In other words, in the neighborhood of  $x = 0$ , the graph of  $\Phi_{nM}(x)$  is approximately the same as the graph of  $\Phi_M(x)$  for sufficiently large  $M$ , if the scale of the horizontal axis of the former is  $n$  times as large as that of the latter. If we index the extrema of  $\Phi_M(x)$  from the origin outward, not counting the central maximum, by  $\pm 1, \pm 2, \dots$ , with positive and negative signs for  $x > 0$  and  $x < 0$ , respectively, then these figures suggest that for an extremum of fixed index, its function value tends to a con-

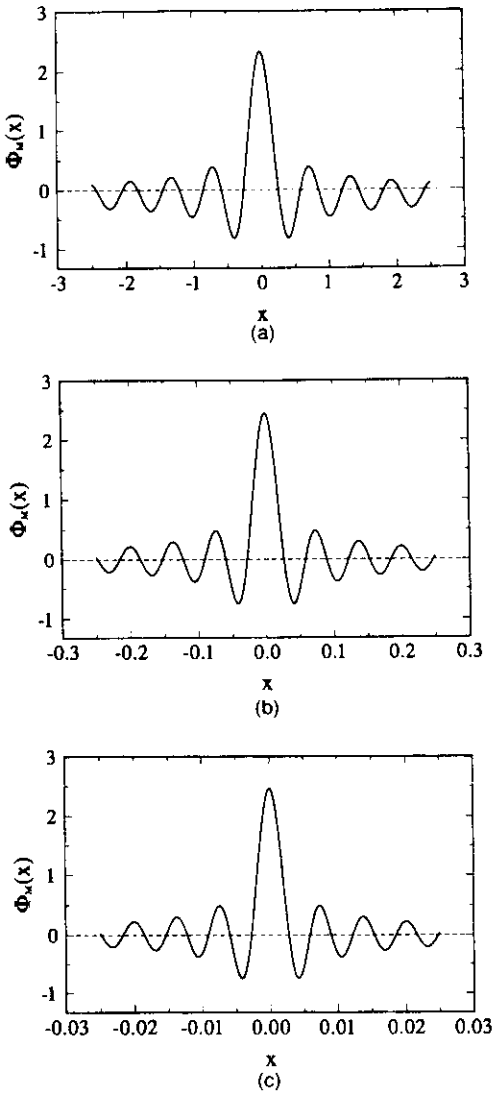


Fig. 2. Graphs of  $\Phi_M(x)$  in the neighborhood of  $x = 0$  for (a)  $M = 10$ , (b)  $M = 100$ , and (c)  $M = 1000$ . Note the change of scale for the horizontal axes.

stant but its position tends to  $x = 0$  as  $M \rightarrow \infty$ . This observation is of course consistent with our earlier conclusion that the convergence of  $\Phi_M(x)$  is nonuniform near  $x = 0$ .

From a graphical point of view, Eq. (18) of theorem 2 says that the graph of  $h^{(M)}(x)$  can be obtained by superimposing a series of properly scaled graphs of  $\Phi_M(x)$  centered at  $x_p \in U_{fg}$  on top of the graph of  $h_M(x)$ . Here for ease of visualization we may assume that both  $f(x)$  and  $g(x)$  are real-valued functions. The effect of such a superposition is most prominent when  $h(x)$  is continuous. In that case,  $h^{(M)}(x)$  will have an overshoot (if  $\hat{f}_p \hat{g}_p < 0$ ) or an undershoot (if  $\hat{f}_p \hat{g}_p > 0$ ) from the graph of  $h_M(x)$  at  $x_p \in U_{fg}$ , whose magnitude tends to  $1/8$  of  $|\hat{f}_p \hat{g}_p|$  as  $M \rightarrow \infty$ . On the other hand, theorem 3 says that when  $h(x)$  is continuous,  $h^{(M)}(x)$  calculated by the inverse rule preserves well the characteristics of  $h(x)$ , including its continuity at  $x_p \in U_{fg}$ . If we set

$f(x_p + 0)/f(x_p - 0) = \alpha$  and again assume that  $h(x)$  is continuous at  $x_p \in U_{fg}$ , then

$$\hat{f}_p \hat{g}_p = -h(x_p) \frac{(1 - \alpha)^2}{\alpha}. \quad (23)$$

Thus the magnitude of the overshoot can be arbitrarily large as  $\alpha \rightarrow 0$  or  $\alpha \rightarrow \pm\infty$ . As illustrations of what has just been said, let us consider two graphical examples.

In the first example, we choose

$$f(x) = \begin{cases} a & |x| < \frac{\pi}{2} \\ \frac{a}{2} & \frac{\pi}{2} < |x| \leq \pi \end{cases}, \quad (a \neq 0), \quad (24)$$

and  $g(x) = 1/f$ . Then it is obvious that the discontinuities of  $f$  and  $g$  are pairwise complementary and  $h(x) = 1$ . Figure 3(a) shows what happens when the partial sum  $h^{(M)}(x)$  is computed with the coefficients  $h_n^{(M)}$  given by the finite Laurent rule. In this and the next example,  $M = 200$ . Figure 3(b) shows an enlarged view of the same partial sum in the neighborhood of  $x = \pi/2$ . As the theory predicted, it is just a graph of  $\Phi_M(x - \pi/2)$  superimposed on  $h_M(x) = 1$ . The peak value of the overshoot is also as predicted because in this case  $(-1/8)\hat{f}_p \hat{g}_p = 1/16 = 0.0625$ . The straight horizontal

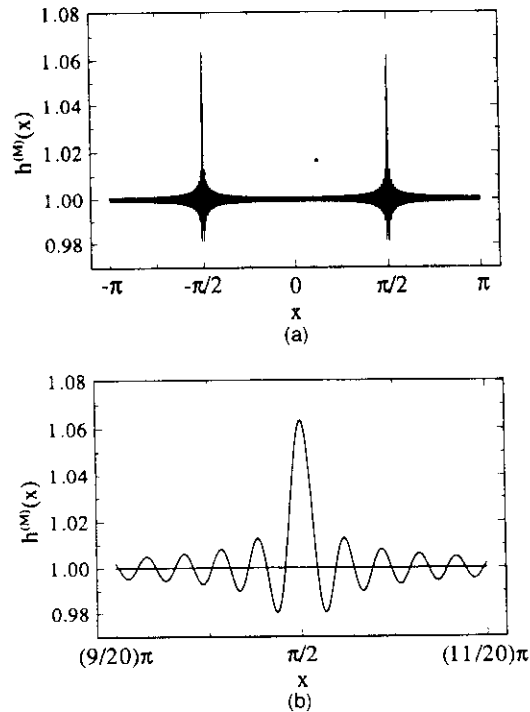


Fig. 3. (a) Graph of  $h^{(M)}(x)$  that is Fourier factorized by the finite Laurent rule, with  $f(x)$  given by Eq. (24),  $g(x) = 1/f(x)$ , and  $M = 200$ . (b) Enlarged view of Fig. 3(a) in the neighborhood of  $x = \pi/2$ . The straight horizontal line in Fig. 3(b) is obtained by the inverse rule.

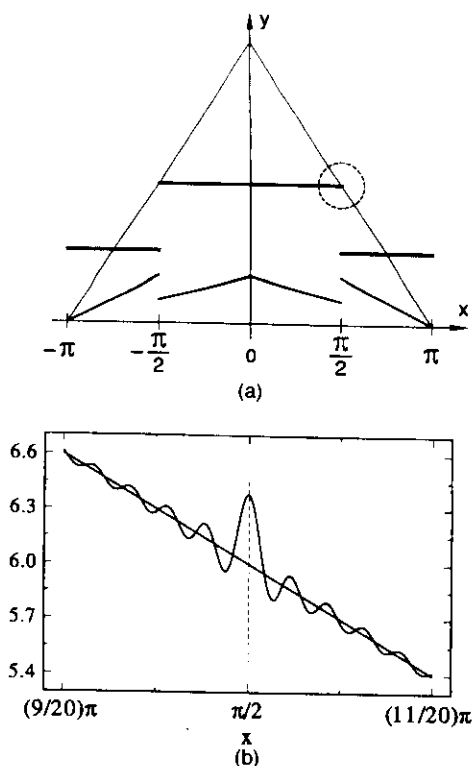


Fig. 4. (a) Schematic representations of functions  $f(x)$  and  $g(x)$  in Eqs. (24) and (25) and their product  $h(x)$  in order of decreasing line thickness. Here  $a = 6$  and  $b = 2$ . (b) Function  $h^{(M)}(x)$ , with  $M = 200$ , in the neighborhood of  $x = \pi/2$ . The oscillatory curve is obtained by Laurent's rule, and the nonoscillatory line is obtained by the inverse rule.

line in Fig. 3(b) is  $h^{(M)}(x)$  computed with  $h_n^{(M)}$  given by the inverse rule. The perfect preservation of the continuity of  $h(x)$  at  $x = \pi/2$  is evident.

Perhaps the above example, in which Eq. (22) gives the exact Fourier coefficient of  $h(x)$ ,  $h_n = \delta_{n0}$ , is too special. In the second example, we keep  $f(x)$  as given by Eq. (24) but choose

$$g(x) = \begin{cases} b \left( 1 - \frac{|x|}{\pi} \right) & |x| < \frac{\pi}{2} \\ 2b \left( 1 - \frac{|x|}{\pi} \right) & \frac{\pi}{2} < |x| \leq \pi \end{cases}, \quad (b \neq 0). \tag{25}$$

Thus the function  $h(x)$  is again continuous. In Fig. 4(a),  $f(x)$ ,  $g(x)$ , and  $h(x)$  are shown schematically in order of decreasing line thickness. Here,  $a = 6$  and  $b = 2$ . Figure 4(b) shows  $h^{(M)}(x)$  in the region enclosed by the dashed circle in Fig. 4(a). The oscillatory curve is obtained by using Laurent's rule, and the straight line is obtained by using the inverse rule. Once again, the inverse rule gives a perfect reconstruction of  $h(x)$ , but Laurent's rule gives a reconstruction that suffers from overshoot and ringing in the neighborhood of the complementary discontinuity.

We say that a product  $f(x)g(x)$  can be Fourier factorized only when Eq. (17) is valid everywhere. If the three

types of product that theorems 1, 3, and 2 are concerned with are referred to as products of type 1, 2, and 3, respectively, then from an operational point of view the three theorems can be summarized as follows:

1. A product of type 1 (two piecewise-smooth, bounded, periodic functions that have no concurrent jump discontinuities) can be Fourier factorized by Laurent's rule.
2. A product of type 2 (two piecewise-smooth, bounded, periodic functions that have only pairwise-complementary jump discontinuities) cannot be Fourier factorized by Laurent's rule, but in most cases it can be Fourier factorized by the inverse rule.
3. A product of type 3 (two piecewise-smooth, bounded, periodic functions that have concurrent but not complementary jump discontinuities) can be Fourier factorized by neither Laurent's rule nor the inverse rule.

### 3. APPLICATION TO THE GRATING PROBLEM

Strictly speaking, a modal field in a periodic medium is representable only by a pseudo-Fourier series, which differs from a Fourier series by the Floquet factor  $\exp(i\alpha_0 x)$ . It is easy to verify that the mathematical results of Section 2 apply to pseudoperiodic functions as well, except for a few changes in the terminology. Therefore for simplicity I will use the term Fourier series in this section to refer broadly to the pseudo-Fourier series of the fields and the Fourier series of the permittivity. The piecewise smoothness and boundedness of the functions required by the theorems in Section 2 are guaranteed here by the physics of the grating problem.

The  $x$ -dependent equations corresponding to Eqs. (1)–(5) are

$$\frac{1}{i} \frac{\partial H_z}{\partial y} = -k_0 \epsilon E_x, \tag{26a}$$

$$\frac{1}{i} \frac{\partial E_x}{\partial y} = -k_0 \mu_0 H_z - \frac{1}{k_0} \frac{\partial}{\partial x} \left( \frac{1}{\epsilon} \frac{\partial H_z}{\partial x} \right), \tag{26b}$$

$$-\frac{\partial^2 H_z}{\partial y^2} = \epsilon \left[ \frac{\partial}{\partial x} \left( \frac{1}{\epsilon} \frac{\partial H_z}{\partial x} \right) + \mu_0 k_0^2 H_z \right]. \tag{27}$$

Now a reader, well equipped with the mathematical theory of Section 2, can immediately see why Eqs. (1), (2), and (5) are incorrect and why Eqs. (3) and (4) are correct. Let us look at the above three equations one by one.

On the basis of the physics, we know that the product  $\epsilon E_x$  in Eq. (26a) should be continuous in  $x$ . Since  $\epsilon$  is discontinuous at  $x = \pm d/2$ ,  $\epsilon$  and  $E_x$  must together have two pairs of complementary jumps there. Equation (1a) is incorrect because it derives from the use of Laurent's rule, which does not apply to a product of type 2. As a result, the left-hand side of Eq. (1a) is the coefficient of a uniformly convergent Fourier series, but the right-hand side is the coefficient of a nonuniformly convergent trigonometric series. The two series converge at different rates to functions that are not equal everywhere. Hence the required continuity of  $\epsilon E_x$  is not uniformly preserved. In contrast, Eq. (3a) can be derived by applying the in-

verse rule to Eq. (26a). Both sides of Eq. (3a) tend to the same mathematical quantity, and the continuity of  $\epsilon E_x$  is uniformly preserved.

The Fourier analysis of Eq. (26b) can be done similarly. Here  $(1/\epsilon)(\partial H_z/\partial x)$  is a product of type 2. Equation (3b) handles this product correctly, but Eq. (1b) does not. For Eq. (27), the term involving  $(1/\epsilon)(\partial H_z/\partial x)$  should be handled just as in Eq. (3b), of course. The entire right-hand side of Eq. (27) should be viewed as the product of  $\epsilon$  and the term in the square brackets. This product is once again of type 2, because the left-hand side of Eq. (27) is continuous with respect to  $x$ . It is incorrectly handled by Eqs. (2) and (5) and correctly handled by Eq. (4). Note that there is no ambiguity in the way that Eqs. (26) and (27) can be Fourier analyzed. For example, if the right-hand side of Eq. (27) is multiplied out to yield two or more terms, then there will be terms that are products of type 3, which cannot be Fourier factored.

For the sake of completeness, I provide two more examples. For TE polarization, the  $z$  component of the electric field obeys the Helmholtz equation:

$$-\frac{\partial^2 E_z}{\partial y^2} = \frac{\partial^2 E_z}{\partial x^2} + \mu_0 k_0^2 \epsilon E_z. \quad (28)$$

Here the product  $\epsilon E_z$  is type 1, so Laurent's rule can be applied, just as every author on this subject has done. In the conical mount the  $x$  component of the electric field of an  $H_{\perp}$  mode (meaning the mode for which  $H_x = 0$ ) obeys the equation

$$k_z^2 E_x - \frac{\partial^2 E_x}{\partial y^2} = \frac{\partial}{\partial x} \left[ \frac{1}{\epsilon} \frac{\partial}{\partial x} (\epsilon E_x) \right] + \mu_0 k_0^2 \epsilon E_x, \quad (29)$$

where  $k_z$  is the  $z$  component of the incident wave vector. Based on either the physics or a mathematical analysis, the products  $\epsilon E_x$  and  $(1/\epsilon)(\partial \epsilon E_x/\partial x)$  must be continuous. Therefore by the inverse rule, Eq. (29) becomes

$$\frac{\partial^2 E_{xn}}{\partial y^2} = k_z^2 E_{xn} + \sum_m (\alpha_n [\epsilon]_{nm}^{-1} \alpha_m - \mu_0 k_0^2 \delta_{nm}) \sum_p \left[ \frac{1}{\epsilon} \right]_{mp}^{-1} E_{xp}. \quad (30)$$

Equation (30) corresponds to Eq. (60) of Ref. 6, but here the field continuities are well preserved.

On the basis of the above examples, the procedure for Fourier analyzing Maxwell's equations that contain a discontinuous permittivity function can be summarized as follows:

1. From the basic Maxwell equations, derive the coupled first-order equations or the second-order equation in terms of the vector field component(s) of interest.
2. Arrange the resulting equation(s) in such a way that the combinations of the permittivity function and the field components form products of type 1 and type 2 only; avoid type 3 products.
3. Substitute the Fourier coefficients for the field components that are not multiplied or divided by the permittivity function, and apply Laurent's rule and the inverse rule to the products of type 1 and 2, respectively.

#### 4. DISCUSSION

My research into the fundamental reason for the success of the new formulation discovered by the authors of Refs. 4 and 5 initially led me onto a path different from the one that has been presented here. Since the convergence of the CWM depends on the convergence of the solutions of the algebraic eigenvalue problem, it is natural for someone to focus attention first on the coefficient matrices on the right-hand side of Eqs. (1)–(5). After all, it is the structure and composition of these matrices that determine the convergence rates. However, such an effort seemed to be difficult and turned out to be unsuccessful for me.

Looking at the problem from a different perspective led to brighter prospects. On the basis of physical understanding and experience, we know that the difficulty of the problem lies at the permittivity discontinuities. If the solutions of the eigenvalue problem converge, they must converge to the modal fields that, by definition, satisfy the boundary conditions. If, in the construction of the eigenvalue problem, no assurance of fast convergence with satisfaction of the boundary conditions is provided, then it would be hopeless to expect the solutions of the eigenvalue problem to converge rapidly. In this sense, the new formulation provides a much better condition for the convergence of the solutions than does the old formulation.

The significance of this paper is by no means limited to the CWM. In a broad sense, any numerical work that requires the Fourier analysis of a product of discontinuous periodic functions could benefit. In particular, this research may have important implications for the classical differential method for gratings.<sup>10</sup> At first glance, it may appear that the results here do not apply to the differential method when the grating profiles are not rectangular. Indeed, as the differential method does not use the so-called multilayer approximation,  $\epsilon E_x$  and  $(1/\epsilon)(\partial H_z/\partial x)$  are not continuous across the grating profile where the surface normal is not in the  $x$  direction. However, since the method relies on numerical integration, in the  $y$  direction, of the unknown field amplitudes, the permittivity is assumed to be independent of  $y$  within each integration step. Thus the multilayer approximation is implicitly used. Therefore I expect that if Eqs. (4.30) and (4.31) of Ref. 10 are replaced by Eqs. (3a) and (3b), respectively, of this paper, the convergence of the differential method will be improved.

I have successfully applied the theorems and procedures developed in this paper to improve the convergence of the coordinate transformation method of Chandezon *et al.*<sup>11</sup> in the case in which the grating profiles have sharp edges. This result will be presented in a separate publication.<sup>12</sup>

From Eq. (11), it follows that if  $\epsilon(x) \neq 0$ , then

$$\sum_{l=-M}^M \epsilon_{m-l} \left( \frac{1}{\epsilon} \right)_{l-n} = \delta_{mn} + \Delta_{mn}, \quad (31)$$

where

$$\Delta_{mn} = \sum_{|l|>M} \epsilon_{m-l} \left( \frac{1}{\epsilon} \right)_{l-n}. \quad (32)$$

Thus, for discontinuous  $\epsilon(x)$ ,  $\Delta_{mn} = 0$  only if  $m$  and  $n$  are such that  $(M \pm n) = x$  and  $(M \pm m) = x$  as  $M = x$ . In other words, the matrix elements of  $\Delta_{mn}$  in the vicinity of the two ends of the main diagonal remain finite as  $M \rightarrow x$ . Therefore

$$\|\epsilon\|^{M+1} \neq \left\| \frac{1}{\epsilon} \right\|^{M+1}, \quad M \rightarrow x. \quad (33)$$

The incorrect assumption of equality between matrices  $\|\epsilon\|^{-1}$  and  $\|1/\epsilon\|$  might have inadvertently played a positive role in the discovery made by the authors of Refs. 4 and 5. It might also be the reason that the authors of Ref. 6 derived Eq. (5).

The work of Refs. 4 and 5 has clearly shown that the improved convergence rate more than offsets the additional computational effort needed to invert the matrices  $\|\epsilon\|$  and  $\|1/\epsilon\|$ . Actually, because these matrices are of the Toeplitz type, the extra work is minimal. There are efficient numerical algorithms<sup>13</sup> that can invert Toeplitz matrices in  $O(M^2)$  instead of  $O(M^3)$  operations. Incidentally, the inverse of a Toeplitz matrix is not necessarily a Toeplitz matrix. This is why double indices  $nm$ , instead of a single index  $n = m$ , have been used to denote the elements of the inverse matrices in this paper.

The subject of this paper serves well to illustrate certain aspects of the relationship among physics, mathematics, and numerics. The physical laws certainly do not insist that their mathematical expressions be held everywhere in the mathematical sense, nor do they require uniform convergence, if infinite series are used in the expressions. From a mathematical point of view, both the old and the new formulations of the CWM are rigorous because they are equal almost everywhere. However, the mathematical difference between everywhere convergence and almost-everywhere convergence and between uniform convergence and nonuniform convergence makes a world of difference in the numerical implementations, as demonstrated by the numerical examples in Refs. 4 and 5.

## 5. CONCLUSION

The success of the new formulation of the coupled-wave method (CWM) recently presented by Lalanne and Morris<sup>4</sup> and by Granet and Guizal<sup>5</sup> is due to the fact that it uniformly preserves the continuity of the electromagnetic-field quantities that should be continuous across permittivity discontinuities. I have given two different rules for Fourier factorizing two different types of products. Furthermore, I have described the procedures for correctly converting Maxwell's equations into linear algebraic systems in discrete Fourier space. As a result, the new formulation of the CWM is placed on a solid mathematical foundation.

Fourier series have been used for a long time to represent the periodic, piecewise-constant permittivity function and its reciprocal in grating analysis. Ironically, the mistake of using Laurent's rule to factor the Fourier coefficient of a product of functions with complementary

jumps has been made by every researcher who has used these series expansions. The lesson learned from this research is that, in converting Maxwell's equations in spatial variables to equations in the discrete Fourier space, one cannot blindly substitute the Fourier series of every term and every factor into the spatial equations; appropriate factorization rules must be applied when discontinuities are present in the factors of the products.

## ACKNOWLEDGMENTS

I am indebted to the authors of Ref. 4, P. Lalanne and G. M. Morris, and the authors of Ref. 5, G. Granet and B. Guizal, for making the preprints of their papers available to me. I am grateful to my colleagues at the Optical Sciences Center, J. J. Burke, N. Ramanujam, and M. Rivera, for their careful proofreading of the manuscript. This research was supported by the Optical Data Storage Center, University of Arizona, and by the Advanced Technology Program of the U.S. Department of Commerce through a grant to the National Storage Industry Consortium.

## REFERENCES

1. K. Knop, "Rigorous diffraction theory for transmission phase gratings with deep rectangular grooves," *J. Opt. Soc. Am.* **68**, 1206-1210 (1978).
2. M. G. Moharam and T. K. Gaylord, "Diffraction analysis of dielectric surface-relief gratings," *J. Opt. Soc. Am.* **72**, 1385-1392 (1982).
3. L. Li and C. W. Haggans, "Convergence of the coupled-wave method for metallic lamellar diffraction gratings," *J. Opt. Soc. Am. A* **10**, 1184-1189 (1993).
4. P. Lalanne and G. M. Morris, "Highly improved convergence of the coupled-wave method for TM polarization," *J. Opt. Soc. Am. A* **13**, 779-784 (1996).
5. G. Granet and B. Guizal, "Efficient implementation of the coupled-wave method for metallic lamellar gratings in TM polarization," *J. Opt. Soc. Am. A* **13**, 1019-1023 (1996).
6. M. G. Moharam, E. B. Grann, D. A. Pommet, and T. K. Gaylord, "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," *J. Opt. Soc. Am. A* **12**, 1068-1076 (1995).
7. L. Li, "Fourier factorization of a product of discontinuous periodic functions," submitted to *SIAM J. Anal. Math.*
8. A. Zygmund, *Trigonometric Series* (Cambridge U. Press, Cambridge, 1977), Vol. 1, Chap. 4, Sec. 8, p. 159.
9. G. H. Hardy, *Divergent Series* (Oxford U. Press, London, 1949), Chap. 10, Secs. 12-15, pp. 239-246.
10. P. Vincent, "Differential methods," in *Electromagnetic Theory of Gratings*, Vol. 22 of Topics in Current Physics, R. Petit, ed. (Springer-Verlag, Berlin, 1980), pp. 101-121.
11. J. Chandezon, M. T. Dupuis, G. Cornet, and D. Maystre, "Multicoated gratings: a differential formalism applicable in the entire optical region," *J. Opt. Soc. Am.* **72**, 839-846 (1982).
12. L. Li and J. Chandezon, "Improvement of the coordinate transformation method for surface-relief gratings with sharp edges," *J. Opt. Soc. Am. A* (to be published).
13. See, for example, G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins U. Press, Baltimore, Md., 1983), Chap. 5, Sec. 7, pp. 125-135, or G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators* (Birkhauser Verlag, Basel, Switzerland, 1984), Chap. 1, pp. 14-33, and the references therein.

