Appl. Phys. B 28, 1–14 (1982)

Applied Physics B Photophysics and Laser Chemistry

Planar Dielectric Grating Diffraction Theories

T. K. Gaylord and M. G. Moharam

School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA

Received 23 December 1981/Accepted 26 February 1982

Abstract. Various planar dielectric grating diffraction theories are reviewed for the case of a general sinusoidal permittivity planar grating with slanted fringes and plane wave incidence at an arbitrary angle. Exact formulations without approximations (rigorous coupled-wave analysis and rigorous modal analysis) are developed first. Then, using a series of fundamental assumptions, rigorous theory is shown to reduce to the various approximate theories in the appropriate limits. The implications of these fundamental assumptions are discussed.

PACS: 42.10, 42.20, 42.30

Since 1930 there have been over 400 scientific papers on the subject of grating diffraction. Many of these papers have been applicable to planar dielectric gratings. These periodic structures have been applied in numerous areas such as acousto-optics, holography, integrated optics, and spectral analysis. The diffraction of electromagnetic waves by spatially periodic media may be analyzed by numerous methods and with a wide variety of possible assumptions. The purpose of this paper is to review both rigorous and approximate planar grating diffraction theories and to show explicitly the relationships between the various theories.

The most common methods of analyzing planar dielectric grating diffraction are the coupled-wave approach [1–8] and the modal approach [9–18]. These theories have recently been treated in two extensive reviews [19, 20]. Both coupled-wave and modal approaches can produce exact formulations without approximations. In their full rigorous forms these formulations are completely equivalent [21]. They represent merely alternative methods of representing the electromagnetic fields inside the grating (Sect. 2).

Starting with the rigorous theories and using a series of fundamental assumptions, these general theories are shown to reduce to the various approximate theories [two-wave modal theory, two-wave second-order coupled-wave theory, multiwave coupled-wave theory, two-wave first-order coupled-wave theory (Kogelnik theory), Raman-Nath theory, and amplitude transmittance theory] in the appropriate limits. This is shown in Sect. 8.

1. Planar Dielectric Grating Diffraction

The general planar grating diffraction problem is depicted in Fig. 1. An electromagnetic wave is obliquely incident upon a slanted-fringe planar grating bounded by two different homogeneous media. In general, there will be simultaneously both forward-diffracted and backward-diffracted waves as shown in the figure. This geometry is applicable 1) to holographic gratings in air or other media ($\varepsilon_I = \varepsilon_{III} \neq \varepsilon_0$), 2) to acousto-optic gratings within a medium ($\varepsilon_I = \varepsilon_0 = \varepsilon_{III}$), and 3) to grating couplers such as used in integrated optics ($\varepsilon_I \neq \varepsilon_0 \neq \varepsilon_{III} \neq \varepsilon_I$). The quantities ε_I, ε_0, and ε_{III} are the average relative permittivities (dielectric constants) in regions 1, 2, and 3, respectively.

In this paper, for simplicity, the case of a lossless dielectric grating with sinusoidal permittivity is treated. The incident plane wave polarization is perpendicular to the plane of incidence (H mode). This is probably the case of widest general interest. However, these assumptions are not essential to the theories described. The relative permittivity (dielectric constant) in the

0721-7269/82/0028/0001/$02.80

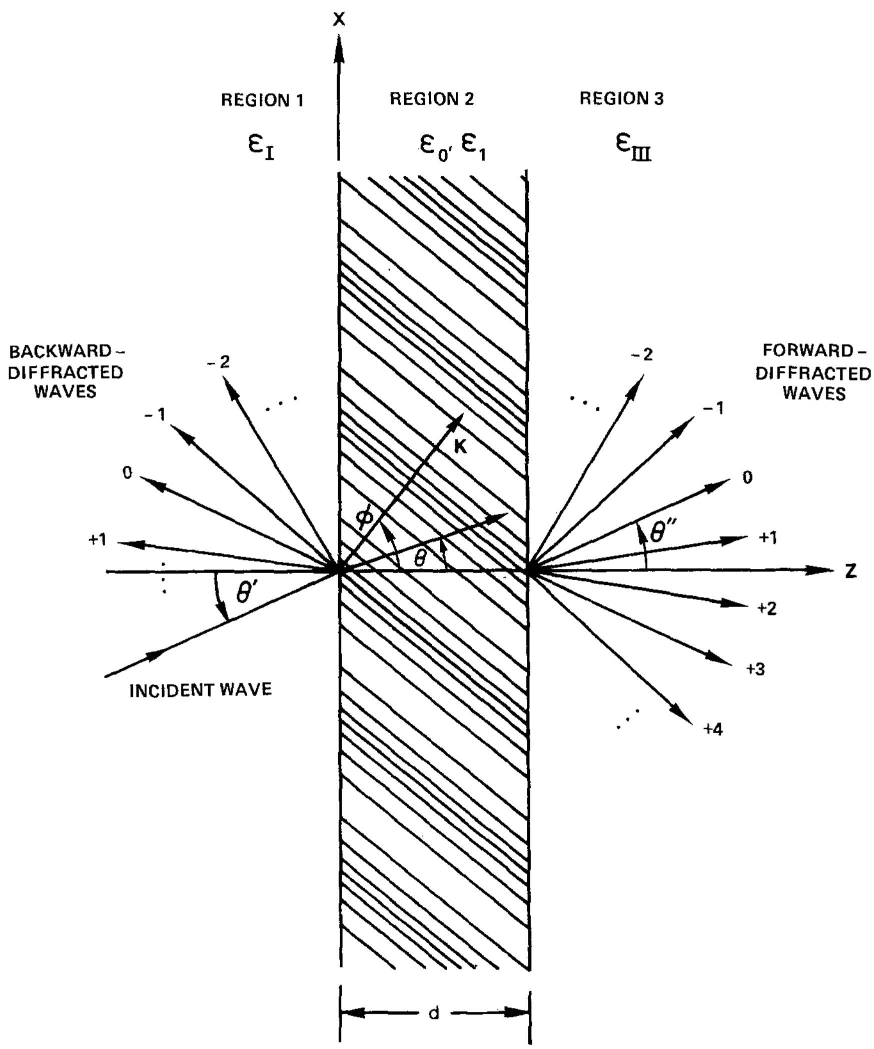

Fig. 1. Geometrical configuration of planar grating diffraction

grating region is given by

$$\begin{aligned}\varepsilon(x,z)&=\varepsilon_0+\varepsilon_1\cos(\bar{\boldsymbol{K}}\cdot\bar{\boldsymbol{r}})\\&=\varepsilon_0+\varepsilon_1\cos[K(x\sin\phi+z\cos\phi)],\end{aligned}\tag{1}$$

where ε_1 is the amplitude of the sinusoidal relative permittivity, ϕ is the grating slant angle, and $K=2\pi/\Lambda$, where Λ is the grating period. The cosinusoidal form used in (1) is common in the volume holographic grating literature. In the acousto-optics literature, a sinusoidal form for (1) is more common. Using the sinusoidal form would alter the resulting equations in the following sections as well as their amplitude solutions. However, the diffracted intensities are identical in either case.

The general approach to the planar-grating diffraction problem involves finding a solution of the wave equation in each of the three regions and then matching the tangential electric and magnetic fields at the two interfaces ($z=0$ and $z=d$). In region 1, the normalized amplitude of the incident plane wave is

$$\begin{aligned}E_{\text{inc}}&=\exp(-j\bar{\boldsymbol{k}}_1\cdot\bar{\boldsymbol{r}})\\&=\exp[-jk_1(x\sin\theta'+z\cos\theta')],\end{aligned}\tag{2}$$

where θ' is the angle of incidence on region 1, $k_1=2\pi(\varepsilon_{\text{I}})^{1/2}/\lambda$, and λ is the free space wavelength. The wave equation for the H mode polarization is the scalar wave equation (Helmholtz equation)

$$\nabla^2E+k^2\varepsilon(x,z)E=0,\tag{3}$$

where $k=2\pi/\lambda$. For H mode polarization, the electric field only has a component in the y direction. The fields and the grating are unchanging in the y direction. For any arbitrary direction the grating as bounded by regions 1 and 3 is periodic only in the x component of the direction. If region 2 was infinite in all directions (not bounded) the resulting grating would be periodic in any direction that was not perpendicular to the grating vector $\bar{\boldsymbol{K}}$. The Floquet theorem [22, 23] restricts the possible fields that can exist in a periodic structure at steady state. As a result of the Floquet theorem the diffracted wavevectors inside the grating may be represented for the infinite periodic medium case by

$$\bar{\boldsymbol{\sigma}}_i=\bar{\boldsymbol{k}}_2-i\bar{\boldsymbol{K}},\tag{4}$$

where $\bar{\sigma}_i$ is the wavevector of the i-th space harmonic in the grating, i is any integer, and $\bar{k}_2$ is the wavevector of the zero-order ($i=0$) space harmonic having a magnitude of $k_2=2\pi(\varepsilon_0)^{1/2}/\lambda$. For the bounded grating that is periodic only in the x-component of direction, the Floquet theorem only requires

$$\bar{\sigma}_i \cdot \hat{x} = (\bar{k}_2 - i\bar{K}) \cdot \hat{x}\,, \tag{5}$$

where $\hat{}$ denotes a unit vector. This expression is just the x-component of (4). Only the form as represented by (5) is necessary for the present problem. However, for subtle reasons that will become clear in the next section, the Floquet requirement as expressed by (4) will be used. This is certainly acceptable since it contains the necessary (5) within it.

2. Rigorous Coupled-Wave and Modal Theories

It is possible to formulate the planar grating diffraction problem depicted in Fig. 1 in an exact manner. This may be done with the coupled-wave approach or the modal approach. The modal approach is sometimes referred to as the Floquet, Floquet-Bloch, eigenmode, characteristic-mode, or coupled-mode approach. The coupled-wave approach is confusingly also sometimes called coupled-mode approach. Both the coupled-wave and modal approaches are altenrative methods of representing the fields inside the grating medium.

In the coupled-wave representation, the fields inside the grating are expanded in terms of the space harmonics of the fields in the periodic structure. These space harmonics inside the grating correspond to diffracted orders outside of the grating. Thus, the partial fields inside the modulated medium are visualized as diffracted waves that progress through the planar slab and couple energy back and forth between each other as they progress. This picture agrees rather well with simple physical intuition about the process of diffraction by a volume grating. In the coupled-wave approach the total field is thus expressed as

$$E(x,z) = \sum_{i=-\infty}^{+\infty} S_i(z)\exp(-j\bar{\sigma}_i \cdot \bar{r})\,, \tag{6}$$

where i is the space harmonic index. Equation (6) has the general appearance of a plane wave expansion of diffracted waves with amplitudes S_i. This would be true if the S_i's were constants. However, since the S_i's are not constants but are functions of z, each i does not correspond to a single plane wave. In general there are an infinite number of plane waves associated with each i. S_i varies only in the direction perpendicular to the boundary. The sum of all of the i-th partial fields as represented by $E(x,z)$ in (6) satisfies the wave equation. However, individually the partial fields do not satisfy the wave equation.

In the modal representation, the fields inside the grating are expanded in terms of the allowable modes of the periodic medium. The fields are visualized as waveguide modes in the grating region. In the modal approach, the total electric field is expressed as a weighted summation over all possible modes,

$$E(x,z) = \sum_{\mathrm{m}=-\infty}^{+\infty} C_{\mathrm{m}} \Phi_{\mathrm{m}}(\bar{r})\exp(-j\bar{k}_{2\mathrm{m}} \cdot \bar{r})\,, \tag{7}$$

where m is the mode index. The function $\Phi_{\mathrm{m}}(\bar{r})$ is periodic with a period equal to the grating period. That is $\Phi_{\mathrm{m}}(\bar{r}) = \Phi_{\mathrm{m}}(\bar{r}+\bar{\Lambda})$. The summation includes both forward and backward propagating modes. The backward propagating modes are due to diffraction in the grating volume (when the grating fringes are slanted) and due to reflections at the $z=d$ boundary. Each individual m-th mode satisfies the wave equation and may be either evanescent or propagating. These modes in the grating are precisely analogous to modes in a waveguide. Each waveguide mode satisfies the wave equation by itself and it may be either cutoff or propagating. Each mode (m) consists of an infinite number of space harmonics (i) and each mode propagates through the medium without change. The space harmonics may be viewed as arising from the Fourier expansion of the periodic function $\Phi_{\mathrm{m}}(\bar{r})$.

The coupled-wave representation or expansion in terms of space harmonics, (6), and the modal representation or expansion in terms of modes, (7), are merely alternative representations of the same physical problem. Both approaches are complete and both are rigorous formulations (without approximations). These two approaches are thus completely equivalent and this will be demonstrated in Sect. 5.

3. Rigorous Coupled-Wave Equations

The rigorous coupled-wave equations are developed by starting with the field expansion in terms of space harmonics, (6). There are mutliple versions of the rigorous coupled-wave equations for the physical situation depicted in Fig. 1 depending on the form of the modulation chosen (sinusoidal or cosinusoidal permittivity), the form of the Floquet condition chosen, (4) or (5), and the polarization chosen (H mode or E mode). The basic case treated in this paper is cosinusoidal permittivity and H mode polarization. Both forms of the Floquet condition are analyzed in this section.

The Floquet condition for an infinite periodic medium, (4), will be treated first since this gives rise to the very useful, common form of the rigorous coupled-wave equations. In fact, only the component of the wavevectors along the boundary in the direction of periodicity

(x) needs to satisfy the Floquet condition. This is given by (5) and will be treated below. The Floquet condition for an infinite grating, (4), contains the required x dependence. The z dependence in (4) will be shown to be very useful in obtaining a directly solvable form of the coupled-wave equations. Substituting (4) into (6) yields

$$E(x,z)=\sum_{i=-\infty}^{+\infty} S_i(z)\exp[-j(\bar{k}_2-i\bar{K})\cdot\bar{r}]$$
$$=\sum_{i=-\infty}^{+\infty} S_i(z)\exp\{-j[(k_2\sin\theta-iK\sin\phi)x$$
$$+(k_2\cos\theta-iK\cos\phi)z]\}\,, \qquad (8)$$

where θ is the angle of refraction of the incident beam from region 1. Thus θ is related to θ' through

$$k_1\sin\theta'=k_2\sin\theta\,. \qquad (9)$$

Substituting (1) and (8) into (3) and performing the indicated differentiations gives

$$\sum_{i=-\infty}^{+\infty}\left\{\frac{\partial^2 S_i(z)}{\partial z^2}-j2(k_2\cos\theta-iK\cos\phi)\frac{\partial S_i(z)}{\partial z}\right.$$
$$-[(k_2\sin\theta-iK\sin\phi)^2+(k_2\cos\theta-iK\cos\phi)^2]S_i(z)$$
$$\left.+k^2\varepsilon_0 S_i(z)+\frac{k^2\varepsilon_1}{2}S_{i-1}(z)+\frac{k^2\varepsilon_1}{2}S_{i+1}(z)\right\}$$
$$\cdot\exp\{-j[(k_2\sin\theta-iK\sin\phi)x+(k_2\cos\theta$$
$$-iK\cos\phi)z]\}=0\,. \qquad (10)$$

This equation must be satisfied for all values of the variables. Thus the coefficient of each exponential must individually be zero for nontrivial solutions. Using this and the definitions of k, k_2, and K, (10) reduces to the rigorous coupled-wave equations:

$$\frac{1}{2\pi^2}\frac{d^2S_i(z)}{dz^2}-j\frac{2}{\pi}\left[\frac{(\varepsilon_0)^{1/2}\cos\theta}{\lambda}-\frac{i\cos\phi}{\Lambda}\right]\frac{dS_i(z)}{dz}$$
$$+\frac{2i(m-i)}{\Lambda^2}S_i(z)+\frac{\varepsilon_1}{\lambda^2}[S_{i+1}(z)+S_{i-1}(z)]=0\,. \qquad (11)$$

This is an infinite set of second-order coupled difference-differential equations. By inspection, it is seen that the wave corresponding to each value of i (space harmonic inside the grating or diffracted order outside of the grating) is coupled to its adjacent ($i+1$ and $i-1$) space harmonics. There is no direct coupling between nonadjacent orders.

In the rigorous coupled-wave equations the quantity m has been defined as

$$m=\frac{2\Lambda(\varepsilon_0)^{1/2}}{\lambda}\cos(\theta-\phi)\,. \qquad (12)$$

The quantity m may have any value in general. For the case when m is an integer, (12) becomes the Bragg condition. However, it is important to realize that the Bragg condition is not specifically an input into this theory. The approach applies to an arbitrary angle of incidence and wavelength. Only if the angle of incidence and wavelength are such that m is an integer does Bragg incidence occur. The rigorous coupled-wave equations given by (11) are in a from that is directly solvable using a state variables approach from linear systems theory. This method of solution will be used in the next section.

The Floquet condition as given by (5) may also be used in space harmonic field expansion, (6), to obtain an alternative set of rigorous coupled-wave equations. Substituting (5) into (6) yields

$$E(x,z)=\sum_{i=-\infty}^{+\infty} S_i(z)\exp[-j(k_{2x}-iK_x)x]$$
$$=\sum_{i=-\infty}^{+\infty} S_i(z)\exp[-j(k_2\sin\theta-iK\sin\phi)x]\,. \qquad (13)$$

The field expansion given by (13) is the same as that in (8) except that the z dependent part of the exponential has been included in the $S_i(z)$ functions in (13). Substituting (1) and (13) into (3), performing the indicated differentiations, and setting the coefficients of each exponential equal to zero as before, produces

$$\frac{1}{2\pi^2}\frac{d^2S_i(z)}{dz^2}-2\left\{\left[\frac{(\varepsilon_0)^{1/2}\sin\theta}{\lambda}-\frac{i\sin\phi}{\Lambda}\right]^2-\frac{\varepsilon_0}{\lambda^2}\right\}S_i(z)$$
$$+\frac{\varepsilon_1}{\lambda^2}\exp\left(+j\frac{2\pi z\cos\phi}{\Lambda}\right)S_{i-1}(z)$$
$$+\frac{\varepsilon_1}{\lambda^2}\exp\left(-j\frac{2\pi z\cos\phi}{\Lambda}\right)S_{i+1}(z)=0\,. \qquad (14)$$

This set of coupled-wave equations contains no first derivative terms in contrast to the coupled-wave equations given in (11). In addition, (14) is a nonconstant coefficient differential equation due to the presence of z in the coefficients of the $S_{i-1}(z)$ and $S_{i+1}(z)$ terms. The equations in the form of (14) represent a linear shift-variant system and direct solution would be difficult. For the case of an unslanted grating ($\phi=\pi/2$, fringes perpendicular to the surface), the equations become constant coefficient differential equations. For this limiting case, the equations become identical to the coupled-wave equations of Kong [Ref. 6, Eqs. (6a) and (6b)] if only two waves are retained ($i=0$, 1).

4. Solution of the Rigorous Coupled-Wave Equations

The rigorous coupled-wave equations as given by (11) represent a set of second-order linear differential equations with constant coefficients. Using the methods of linear systems analysis [24] this differential equation description of this continuous system may be transformed into a state space description and a solution obtained directly. By defining the state variables as

$$S_{1,i}(z) = S_i(z), \tag{15}$$

$$S_{2,i}(z) = \frac{dS_i(z)}{dz} \tag{16}$$

the infinite set of second-order differential equations (11) are transformed into two infinite sets of first-order differential equations:

$$\frac{dS_{1,i}(z)}{dz} = S_{2,i}(z), \tag{17}$$

$$\frac{dS_{2,i}(z)}{dz} = -\frac{2\pi^2\varepsilon_1}{\lambda^2} S_{1,i-1}(z) + \frac{4\pi^2 i(i-m)}{\Lambda^2} S_{1,i}(z) - \frac{2\pi^2\varepsilon_1}{\lambda^2} S_{1,i+1}(z) + j4\pi\left(\frac{(\varepsilon_0)^{1/2}\cos\theta}{\lambda} - \frac{i\cos\phi}{\Lambda}\right) S_{2,i}(z). \tag{18}$$

Equations (17) and (18) are the state equations corresponding to the rigorous coupled-wave equations (11). Since these are homogeneous equations, they correspond to unforced state equations. State equations that are linear differential equations with constant coefficients such as these, may be solved for closed-form expressions for the state variables. In this case, only the homogeneous solution is necessary as there are no driving terms in these equations. The homogeneous solutions are

$$S_{l,i}(z) = \sum_{\mathrm{m}=-\infty}^{+\infty} C_{\mathrm{m}} w_{l,i\mathrm{m}} \exp(\lambda_{\mathrm{m}} z) \tag{19}$$

for $l = 1, 2$. The coefficients C_{m} are unknown constants to be determined from the boundary conditions. The quantity $w_{l,i\mathrm{m}}$ is an element of an eigenvector and λ_{m} is an eigenvalue. These quantities are determined as described below. The solution for the wave amplitudes (the "output equation" in linear systems terminology) is $S_i(z) = S_{1,i}(z)$.

The constituent state equations (17) and (18) may be written in matrix form as

$$\begin{bmatrix} \vdots \\ \dot{S}_{1,-2} \\ \dot{S}_{1,-1} \\ \dot{S}_{1,0} \\ \dot{S}_{1,1} \\ \dot{S}_{1,2} \\ \vdots \\ \dot{S}_{2,-2} \\ \dot{S}_{2,-1} \\ \dot{S}_{2,0} \\ \dot{S}_{2,1} \\ \dot{S}_{2,2} \\ \vdots \end{bmatrix} = \begin{bmatrix} & & & \vdots & & & & & & \vdots & & & \\ & 0 & 0 & 0 & 0 & 0 & & 1 & 0 & 0 & 0 & 0 & \\ & 0 & 0 & 0 & 0 & 0 & & 0 & 1 & 0 & 0 & 0 & \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 & 0 & \cdots \\ & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 1 & 0 & \\ & 0 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0 & 1 & \\ & & & \vdots & & & & & & \vdots & & & \\ & b_{-2} & a & 0 & 0 & 0 & & c_{-2} & 0 & 0 & 0 & 0 & \\ & a & b_{-1} & a & 0 & 0 & & 0 & c_{-1} & 0 & 0 & 0 & \\ \cdots & 0 & a & b_0 & a & 0 & \cdots & 0 & 0 & c_0 & 0 & 0 & \cdots \\ & 0 & 0 & a & b_1 & a & & 0 & 0 & 0 & c_1 & 0 & \\ & 0 & 0 & 0 & a & b_2 & & 0 & 0 & 0 & 0 & c_2 & \\ & & & \vdots & & & & & & \vdots & & & \end{bmatrix} \begin{bmatrix} \vdots \\ S_{1,-2} \\ S_{1,-1} \\ S_{1,0} \\ S_{1,1} \\ S_{1,2} \\ \vdots \\ S_{2,-2} \\ S_{2,-1} \\ S_{2,0} \\ S_{2,1} \\ S_{2,2} \\ \vdots \end{bmatrix}, \tag{20}$$

where $a = -2\pi^2\varepsilon_1/\lambda^2$, $b_i = 4\pi^2 i(i-m)/\Lambda^2$, and $c_i = j4\pi[(\varepsilon_0)^{1/2}\cos\theta/\lambda - i\cos\phi/\Lambda]$. This equation may be represented concisely as $\dot{\mathbf{S}} = \mathbf{AS}$ where $\dot{\mathbf{S}}$ and $\mathbf{S}$ are the column vectors in (20) and $\mathbf{A}$ is the coefficient matrix. The needed eigenvalues and eigenvectors are determined from this coefficient matrix. Although $\mathbf{A}$ is an infinite matrix, results may be obtained in practice to an arbitrary level of accuracy with a truncated matrix. Each of the four submatrices is truncated to $n \times n$. As the integer n increases, the calculated results rapidly converge to the exact results. The quantity n corresponds to the total number of space harmonics retained in the analysis. This in turn means that the analysis includes n diffracted waves in region 1 and n diffracted waves in region 3. To put the four submatrices into standard form, the integers i and m are replaced with the new integers p and q that run from 1 to n. For example, if an odd number of waves are retained symmetrically about $i = 0$ (the undiffracted transmitted wave) in the analysis, then $p = i + (n+1)/2$ and $q = \mathrm{m} + (n+1)/2$. The $2n$ solutions may then be expressed

$$S_{l,p}(z) = \sum_{r=1}^{2} \sum_{q=1}^{n} C_{r,q} w_{l,p;r,q} \exp(\lambda_{r,q} z) \tag{21}$$

for $l = 1, 2$ and $p = 1$ to n. The eigenvalues $\lambda_{r,q}$ are determined by solving the determinantal equation

$$|\mathbf{A} - \lambda_{r,q}\mathbf{I}| = 0, \tag{22}$$

where $\mathbf{I}$ is the unit matrix. The eigenvector corresponding to a particular eigenvalue $\lambda_{r,q}$ is determined by substituting $2n$ expressions ($l = 1, 2$ and $p = 1$ to n) for $S_{l,p}$ of the form $S_{l,p} = B_{l,p;r,q}\exp(\lambda_{r,q})$ into the state equation (20), performing the indicated differentiations, and then solving for each element of the eigenvector as $w_{l,p;r,q} = B_{l,p;r,q}/B_{1,1\pi;r,q}$ using Cramer's rule and thus expressing each element as a ratio of determinants. The eigenvalues and eigenvectors for a matrix are typically calculated numerically using a computer library program [25].

5. Equivalence of Coupled-Wave and Modal Representations

The total field inside the grating may be expressed in coupled-wave form, (6), or in modal form, (7), depending whether the field is expanded in terms of space harmonics or in terms of modes, respectively. These two forms are alternative representations that are completely equivalent. This equivalence has been discussed previously [14, 19, 21]. It can be shown mathematically in a simple manner as follows.

Substituting $S_i(z)=\sum_{m=-\infty}^{+\infty} C_m w_{1,im}\exp(\lambda_m z)$ into the coupled-wave expansion (8) gives

$$E(x,z)=\sum_{i=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty} C_m w_{1,im} \cdot\exp\{-j[(k^2\sin\theta-iK\sin\phi)x+(k_2\cos\theta-iK\cos\phi+j\lambda_m)z]\}. \quad (23)$$

Changing the order of the summation this may be rewritten as

$$E(x,z)=\sum_{m=-\infty}^{+\infty} C_m \sum_{i=-\infty}^{+\infty} w_{1,im}\exp[-j(\bar{k}_{2m}-i\bar{K})\cdot\bar{r}], \quad (24)$$

where $\bar{k}_{2m}=k_2\sin\theta\hat{x}+(k_2\cos\theta+j\lambda_m)\hat{z}$. Identifying the complex Fourier series and its representation of a periodic function

$$\sum_{i=-\infty}^{+\infty} w_{1,im}\exp(ji\bar{K}\cdot\bar{r})=\Phi_m(\bar{r})=\Phi_m(\bar{r}+\bar{\Lambda}) \quad (25)$$

gives the modal expansion

$$E(x,z)=\sum_{m=-\infty}^{+\infty} C_m\Phi_m(\bar{r})\exp(-j\bar{k}_{2m}\cdot\bar{r}) \quad (26)$$

which is identical to (7). Thus the coupled-wave and modal representations are seen to be equivalent.

6. Phase Matching and Boundary Conditions

Each i-th field in region 1 and 3 must be phase matched to the i-th space harmonic field inside the grating. In addition, the magnitude of the fields in regions 1 and 3 must be such that the electromagnetic boundary conditions are satisfied at the two grating boundaries ($z=0$ and $z=d$).

The total electric field in region 1 is the sum of the incident and the backward-traveling waves. The normalized total electric field in region 1 may be expressed as

$$E_1=\exp(-j\bar{k}_1\cdot\bar{r})+\sum_{i=-\infty}^{\infty} R_i\exp(-j\bar{k}_{1i}\cdot\bar{r}), \quad (27)$$

where R_i is the normalized amplitude of the i-th reflected wave in region 1 with wavevector $\bar{k}_{1i}$. Likewise the normalized total electric field in region 3 is

$$E_3=\sum_{i=-\infty}^{\infty} T_i\exp[-j\bar{k}_{3i}\cdot(\bar{r}-d\hat{z})], \quad (28)$$

where T_i is the normalized amplitude of the i-th transmitted wave in region 3 with wavevector $\bar{k}_{3i}$. These fields in regions 1 and 3 are phased matched to the field in the grating, (8). Thus the x components of the wavevectors of the i-th wave (regions 1 and 3) and the x component of the wavevector of the i-th space harmonic field (region 2) must be the same. That is

$$\bar{k}_{1i}\cdot\hat{x}=(\bar{k}_2-i\bar{K})\cdot\hat{x}=\bar{k}_{3i}\cdot\hat{x}$$
$$k_1\sin\theta'_i=k_2\sin\theta-iK\sin\phi=k_3\sin\theta''_i. \quad (29)$$

In the homogeneous regions (1 and 3) the backward- and forward-diffracted waves have wavevectors with magnitudes

$$|\bar{k}_{1i}|=|\bar{k}_1| \quad\text{and}\quad |\bar{k}_{3i}|=|\bar{k}_3|, \quad (30)$$

where $k_3=2\pi(\varepsilon_{\text{III}})^{1/2}/\lambda$. Knowing the total amplitudes and the x components of the diffracted wavevectors, the z components are then determined to be

$$\bar{k}_{1i}\cdot\hat{z}=[|\bar{k}_1|^2-(\bar{k}_{1i}\cdot\hat{x})^2]^{1/2}=[k_1^2-(k_2\sin\theta-iK\sin\phi)^2]^{1/2} \quad (31)$$

and

$$\bar{k}_{3i}\cdot\hat{z}=[|\bar{k}_3|^2-(\bar{k}_{3i}\cdot\hat{x})^2]^{1/2}=[k_3^2-(k_2\sin\theta-iK\sin\phi)^2]^{1/2}. \quad (32)$$

These quantities are either positive real (propagating wave) or negative imaginary (evanescent wave).

Using the phase matching conditions (29), (31), and (32), the total fields in regions 1 and 3, (27) and (28), may be rewritten

$$E_1=\exp\{-j[k_1(\sin\theta' x+\cos\theta' z)]\}+\sum_{i=-\infty}^{\infty} R_i\exp[-j\{(k_2\sin\theta-iK\sin\phi)x-[k_1^2-(k_2\sin\theta-iK\sin\phi)^2]^{1/2}z\}] \quad (33)$$

and

$$E_3=\sum_{i=-\infty}^{\infty} T_i\exp\{-j\{(k_2\sin\theta-iK\sin\phi)x+[k_3^2-(k_2\sin\theta-iK\sin\phi)^2]^{1/2}(z-d)\}\}. \quad (34)$$

Electromagnetic boundary conditions require that the tangential electric and tangential magnetic fields be continuous across the two boundaries ($z=0$ and $z=d$). For the H mode polarization described in this paper, the electric field only has a component in the y

direction and so it is the tangential electric field directly. The magnetic field intensity, however, must be obtained through the Maxwell equation $\nabla \times \bar{E} = -\partial \bar{B}/\partial t$. The tangential component of H is in the x direction and is thus given by $H_x = (-j/\omega\mu)\partial E_y/\partial z$. For each value of i, the four quantities to be matched and the resulting boundary condition are:

1) tangential E at $z=0$:

$$\delta_{i0} + R_i = S_i(0), \tag{35}$$

2) tangential H at $z=0$:

$$j[k_1^2 - (k_2 \sin\theta - iK \sin\phi)^2]^{1/2}(R_i - \delta_{i0}) = \frac{dS_i(0)}{dz}, \tag{36}$$

3) tangential E at $z=d$:

$$T_i = S_i(d) \exp[-j(k_2 \cos\theta - iK \cos\phi)d], \tag{37}$$

4) tangential H at $z=d$:

$$-j[k_3^2 - (k_2 \sin\theta - iK \sin\phi)^2]^{1/2} T_i = \left[\frac{dS_i(d)}{dz} - j(k_2 \cos\theta - iK \cos\phi) S_i(d)\right] \cdot \exp[-j(k_2 \cos\theta - iK \cos\phi)d], \tag{38}$$

where δ_{i0} is the Kronecker delta function.

7. Diffraction Efficiency

The quantity commonly measured in grating diffraction is the diffraction efficiency. It is defined as the diffracted intensity of a particular order divided by the input intensity. In the above formulation, the incident plane wave amplitude was normalized to unity. Thus the diffraction efficiencies in regions 1 and 3 are

$$\begin{aligned} DE_{1i} &= \mathrm{Re}\{(\bar{k}_{1i} \cdot \hat{z})/(\bar{k}_{10} \cdot \hat{z})\} R_i R_i^* \\ &= \mathrm{Re}\{\{1 - [\sin\theta' - i\lambda \sin\phi/(\varepsilon_I)^{1/2}\Lambda]^2\}^{1/2}/\cos\theta'\} R_i R_i^* \end{aligned} \tag{39}$$

and

$$\begin{aligned} DE_{3i} &= \mathrm{Re}\{(\bar{k}_{3i} \cdot \hat{z})/(\bar{k}_{10} \cdot \hat{z})\} T_i T_i^* \\ &= \mathrm{Re}\{\{(\varepsilon_{III}/\varepsilon_I) - [\sin\theta' - i\lambda \sin\phi/(\varepsilon_I)^{1/2}\Lambda]^2\}^{1/2}/\cos\theta'\} T_i T_i^* . \end{aligned} \tag{40}$$

The real part of the ratio of the propagation constants occurs when the time-average power-flow density is obtained by taking the real part of the complex Poynting vector. For an unslanted grating ($\phi = \pi/2$) with the same medium on both sides ($\varepsilon_I = \varepsilon_{III}$), the real part of the ratio of the propagation constants is just the usual ratio of the cosine of the diffraction angle for the i-th diffracted wave to the cosine of the incidence angle.

If n values of i are retained in the analysis, then there will be n forward-diffracted waves (n values of T_i) and n backward-diffracted waves (n values of R_i). Correspondingly, there will be $2n$ unknown values of C_m. This is because the coefficient matrix in (20) is a $2n \times 2n$ matrix and therefore has $2n$ eigenvalues and thus there are $2n$ unknown values of C_m. Also this may be viewed as being due to the n coupled-wave equations, each being a second-order differential equation, and thus there are $2n$ roots or eigenvalues and $2n$ unknown constants C_m to be determined from the boundary conditions. Therefore, the total number of unknowns is $4n$. Substituting $S_i(z)$, as given by (15) and (19), into the equations for the boundary conditions (35)–(38) produces n linear equations containing the $4n$ unknowns. An efficient procedure to solve these equations is to eliminate R_i and T_i from these equations and to solve the resulting $2n$ equations for the $2n$ values of C_m using a technique such as Gauss elimination. Then the n values of R_i and n values of T_i may be determined from (35) and (37) respectively. Finally, the diffraction efficiencies DE_{1i} and DE_{3i} are calculated using (39) and (40). For phase gratings the input power is conserved and thus the sum of all of the efficiencies for the propagating waves is unity. That is,

$$\sum_i (DE_{1i} + DE_{3i}) = 1 . \tag{41}$$

Equation (41) may be used to verify the convergence of the numerical calculations.

8. Approximate Theories

The vast majority of the papers on grating diffraction theory have dealt with approximate theories. There are a large number of possible approximations and assumptions that can be made. These generally lead to enormous simplifications in the analyses. In some cases, these simplifications allow analytic solutions to be obtained. A number of famous analytic expressions occur for special limiting cases.

In this section, a large number of planar grating diffraction theories are classified in terms of the fundamental assumptions: 1) neglect of higher-order waves, 2) neglect of second derivatives of the field amplitudes, 3) neglect of boundary effects, 4) neglect of dephasing from the Bragg condition, and 5) the small grating modulation approximation. In addition to these assumptions, a number of other approximations such as normal incidence, unslanted gratings, and large grating period compared to a wavelength, may also be made. However, in this section, only the fundamental assumptions enumerated above are treated. Thus all of the approximate theories are presented in their general form allowing for arbitrary angle of incidence (θ'), arbitrary grating slant angle (ϕ), and arbitrary grating period (Λ). The various further reductions can then be easily formulated, if desired, from these general forms of the approximate theories.

In region 1 of Fig. 1, backward-traveling waves exist. In general, these waves are produced both by diffraction from within the grating volume and by boundary effects (diffraction and reflection from the periodic boundaries at $z=0$ and $z=d$). These physical processes produce a spectrum of plane waves traveling back into region 1 ($z<0$). For the general planar grating of Fig. 1, neglecting the second derivatives of the field amplitudes in the wave equation reduces the number of waves in the analysis from $2n$ to n. The bulk diffracted orders are retained and the boundary-produced waves are eliminated. Thus for a planar grating, the neglect of second derivatives and the neglect of boundary effects are absolutely linked together. When these assumptions are made, the resulting first-order coupled-wave analyses have the amplitudes of the diffracted waves calculated *inside* the modulated region. Then the amplitudes T_i of the forward-diffracted output waves are obtained (approximately) by arguing that they are equal to $S_i(d)$, the space harmonic field amplitude at a distance d from the input surface $z=0$. Likewise for those values of i that represent backward-diffracted waves, the amplitudes R_i are estimated to be $S_i(0)$. However, in the physical problem being analyzed, there are no boundaries at $z=0$ and $z=d$. These planes just represent reference locations. There are no reflected or diffracted waves resulting from these planes and thus there are no physical boundaries at these locations! Thus, the assumptions of neglecting the second derivatives of field amplitudes and neglecting boundary effects have transformed the problem into a filled-space problem (a grating filling all space) with imaginary boundaries at $z=0$ and $z=d$ that are used only to obtain an approximate mathematical formulation of the problem. The first-order theory approaches are not capable of solving the problem of the general planar slab grating bounded by two media different from the grating medium. These two linked assumptions therefore, unmistakably imply the filled-space problem. *After* the filled-space problem is solved, *then* it is assumed that the grating terminates at $z=0$ and $z=d$ and, as a result, that $T_i \cong S_i(d)$ for the forward-diffracted waves and $R_i \cong S_i(0)$ for the backward-diffracted waves. This is obviously only an approximation to the actual situation.

Another consequence of neglecting second derivatives is the exclusion of some propagating waves. In first-order theory, only half of the waves can be retained in the analysis. That is, only one set of i values (as opposed to two sets) is included. For a genral slanted grating, some of these waves may be forward-diffracted and some of them may be backward-diffracted. From (8), if $k_2 \cos\theta - iK\cos\phi$ is positive, the wave is forward-diffracted and if negative, it is backward-diffracted. For forward-diffracted waves, the boundary condition used must be $S_i(0)=0$. For backward-diffracted waves, the appropriate boundary condition is $S_i(d)=0$. The second set of waves (set of i values) are phase matched to these waves. This second set of waves is, of course, neglected in any first order analysis. For the example depicted in Fig. 2, the backward-diffracted waves for $-1 \leqq i \leqq +4$ would all be neglected in first-order theory. The diffraction efficiencies of these backward-diffracted waves are arbitrarily set equal to zero. For the case of a slanted-fringe grating, the power in the neglected phase-matched waves has been shown to be very significant in some cases [8]. Thus the errors introduced by using first order theory can be particularly significant for slant angles away from $\phi=0$ and $\phi=\pi/2$.

Still another consequence of neglecting second derivatives is the exclusion of evanescent waves from the analysis. In first order theory, the filled-space nature of the grating being analyzed, causes one complete set of diffracted orders (i) to exist inside the grating, since all of the $S_i(z)$'s exist there. These calculated values of $S_i(z)$ may have wavevectors with components either in the $+$ or $-$ z directions. However, many of the wavevectors of the $S_i(z)$'s cannot be phased matched to plane waves outside of the grating (regions 1 and 3). This may be seen from Fig. 2. For this example, the values $-1 \leqq i \leqq +4$ correspond to propagating plane waves in regions 1 and 3. The values $i \leqq -2$ and $i \geqq +5$ correspond to evanescent waves in region 1 and 3. However, in first order analysis (without second derivatives) all values of i are treated as representing propagating waves. This is obviously not true. Nevertheless, diffraction efficiencies can be calculated for these evanescent waves as though they were propagating. These predicted efficiencies are clearly incorrect since they should be zero. If the grating period is much larger than a wavelength ($\Lambda \gg \lambda$), then there will be a large number of propagating waves and the effect of excluding evanescent waves would be reduced.

Therefore, it is concluded that all first-order theories inherently contain: 1) the approximate method for calculating diffracted amplitudes described above, 2) neglect of phase-matched waves, and 3) neglect of evanescent waves.

A depiction of various planar grating diffraction theories and their interrelationships in terms of fundamental assumptions is shown in Fig. 3. Most of the literature on planar grating diffraction theory can be connected with a particular block in this diagram. The importance of the various assumptions cannot always

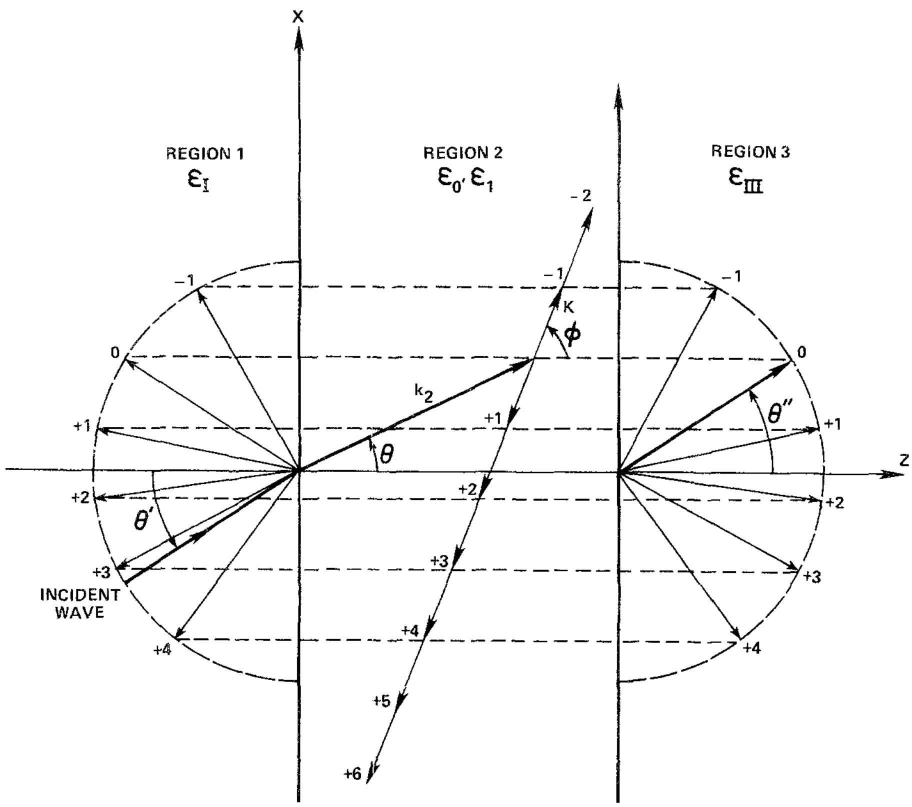

Fig. 2. Allowed wavevectors in regions 1 and 3 due to the presence of a slanted grating with wavevector $\bar{K}$. Phase matching of the diffracted waves outside of the grating with the boundary components of the wavevectors inside the grating (via Floquet construction) is shown. For $-1 \leqq i \leqq +4$, propagating diffracted orders exist in regions 1 and 3 whereas for $i \leqq -2$ and $i \geqq +5$, the waves are evanescent

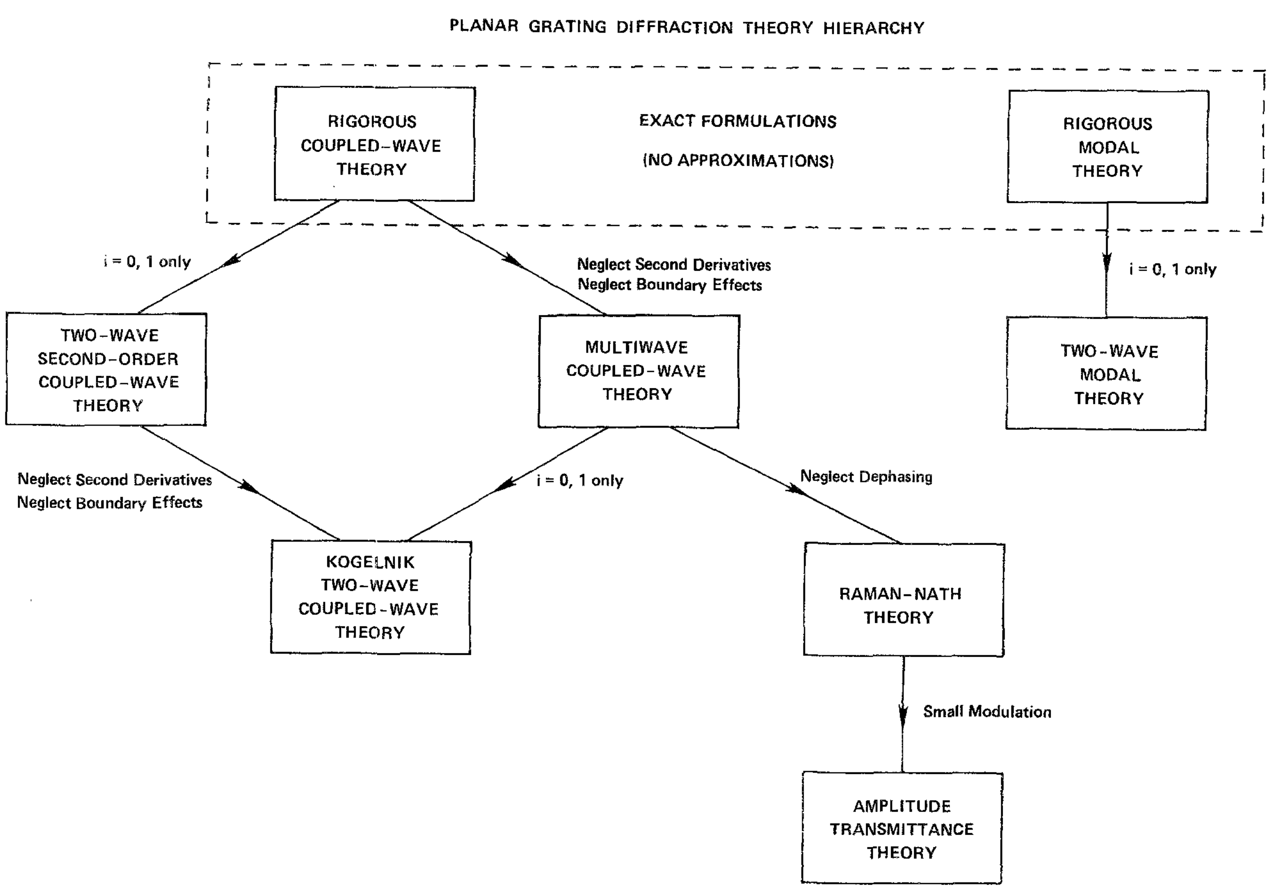

Fig. 3. Interrelationships between various planar grating diffraction theories in terms of fundamental approximations

be isolated. For example, retaining only two waves ($i=0, 1$) and the small modulation approximation can be linked for the case of a sufficiently "thick" grating.

8.1. Two-Wave Modal Theory

If only the zero and first order waves ($i=0, 1$) are retained and all higher-order waves are neglected, a two-wave regime is being assumed. There are actually a total of four waves in this analysis since there are two more waves phased matched to these. Modal theory solutions in the two-wave regime were first obtained by Bergstein and Kermisch [13] with more recent results being contributed by Lederer and Langbein [26], and Russell [19]. In this approch, the standard modal expansion (7) is used to represent the fields in the grating. However, in the two-wave case only the first two Fourier components ($i=0, 1$) of the periodic function $\Phi_m(\bar{r})$ are retained in the analysis, (25). Comparison of two-wave modal theory with exact rigorous theory [8] has shown that this can be valid near Bragg incidence in reflection gratings (backward-diffracted waves dominate). Comparison data are shown in [Ref. 8, Fig. 9].

8.2. Two-Wave Second-Order Coupled-Wave Theory

Two-wave second-order coupled-wave theory and two-wave modal theory represent exactly the same approximation. Both representations include second derivatives of field amplitudes and boundary effects. Both theories retain only the transmitted wave ($i=0$) and the fundamental diffracted wave ($i=1$) and their phased matched waves and neglect higher-order waves. This approximate theory has been used by Kong [6]. Additional approximations in this theory have been made by Kessler and Kowarschik [27–29], and by Jaaskelainen et al. [30]. The two governing equations may be obtained directly from the rigorous coupled-wave equations (11) by keeping only terms in S_0 and S_1 and neglecting all other field amplitudes. The resulting two equations from (11) are:

$$\frac{1}{2\pi^2}\frac{d^2S_0(z)}{dz^2}-j\frac{2(\varepsilon_0)^{1/2}\cos\theta}{\pi\lambda}\frac{dS_0(z)}{dz}+\frac{\varepsilon_1}{\lambda^2}S_1(z)=0, \quad (42)$$

$$\frac{1}{2\pi^2}\frac{d^2S_1(z)}{dz^2}-j\frac{2}{\pi}\left[\frac{(\varepsilon_0)^{1/2}\cos\theta}{\lambda}-\frac{\cos\phi}{\Lambda}\right]\frac{dS_1(z)}{dz}$$
$$+\frac{2(m-1)}{\Lambda^2}S_1(z)+\frac{\varepsilon_1}{\lambda^2}S_0(z)=0. \quad (43)$$

Kong [6] has presented analytical solutions for the two-wave second-order coupled-wave theory expressed in the form of two transmission and two reflection coefficients for the unslanted-fringe planar slab grating.

8.3. Multiwave Coupled-Wave Theory

Multiwave first-order coupled-wave theory may also be developed directly from the rigorous coupled-wave equations (11). In this approach higher-order waves are retained (hence "multiwave"). The second derivatives of the field amplitudes (and thus boundary effects) are neglected. The resulting multiwave coupled-wave equations from (11) are:

$$-j\frac{2}{\pi}\left[\frac{(\varepsilon_0)^{1/2}\cos\theta}{\lambda}-\frac{i\cos\phi}{\Lambda}\right]\frac{dS_i(z)}{dz}+\frac{2i(m-i)}{\Lambda^2}S_i(z)$$
$$+\frac{\varepsilon_1}{\lambda^2}[S_{i+1}(z)+S_{i-1}(z)]=0. \quad (44)$$

For the case of an unslanted transmission grating ($\phi=\pi/2$) and normal incidence ($\theta=0$, $m=0$), the multiwave coupled-wave equations first appeared in a 1936 paper by Raman and Nath [31] for a sinusoidal (rather than cosinusoidal) grating. This paper was the fourth in a series of five papers by Raman and Nath [31–35] on the diffraction of light by sound waves. The first three papers [32–34] from the basis of the "Raman-Nath theory" described below. This simplified multiwave coupled-wave equation was referred to by Nath [36] as being due to Nath [37]. In this 1936 paper, Nath [37] obtained a very slowly converging series solution for the multi-wave coupled-wave difference-differential equations. An alternative series solution was later presented by Berry [38]. This series solution is in terms of Bessel functions and is also very slowly converging. Numerical solutions of the multiwave coupled-wave equations (also for acousto-optic interaction studies) have been obtained by Klein and Cook [3].

The multiwave coupled-wave equations have been generalized to include loss and gratings of arbitrary nonsinusoidal profile by Magnusson and Gaylord [7]. In that paper, numerical solutions were obtained for unslanted transmission gratings using a Runge-Kutta algorithm to solve the first-order system of coupled-wave equations. Diffraction efficiency results for sinusoidal, square-wave, and sawtooth phase gratings at first, second, and third Bragg incidence are presented there.

Comparison of diffraction efficiency results from multiwave first-order coupled-wave theory with exact rigorous theory has shown that this theory without second derivatives gives good results in transmission gratings (forward-diffraction waves dominate) when the grating modulation is small. Comparison data are shown in [Ref. 8, Figs. 7 and 8].

8.4. Two-Wave First-Order Coupled-Wave Theory

If higher-order waves ($i\neq 0, 1$) and second derivatives of field amplitudes (and thus boundary effects) are *both* neglected, the rigorous coupled-wave equations(11) reduce to two-wave first-order coupled-wave theory. For general slanted gratings at arbitrary incidence the two governing equations are:

$$\cos\theta \frac{dS_0(z)}{dz} + j\frac{\pi\varepsilon_1}{2(\varepsilon_0)^{1/2}\lambda} S_1(z) = 0, \tag{45}$$

$$\left(\cos\theta - \frac{\lambda}{\Lambda(\varepsilon_0)^{1/2}}\right)\frac{dS_1(z)}{dz} + j\frac{\pi\lambda(m-1)}{\Lambda^2(\varepsilon_0)^{1/2}} S_1(z) + j\frac{\pi\varepsilon_1}{2(\varepsilon_0)^{1/2}\lambda} S_0(z) = 0. \tag{46}$$

Two-wave first-order coupled-wave theory was applied to acousto-optics by Phariseau [2]. It was first applied to holography by Kogelnik [4]. His thorough 1969 paper [4] is now very widely referenced. As a result, this theory is commonly called "Kogelnik theory" and this is noted in Fig. 3. The substantial recognition received by Kogelnik's paper [4] is due in part to the comprehensive coverage of 1) phase, absorption, and mixed gratings; 2) on-Bragg and off-Bragg incidence; 3) pure transmission ($\phi=\pi/2$), pure reflection ($\phi=0$), and general slanted fringe gratings; and 4) both H-mode and E-mode polarization.

From (8), if $k_2\cos\theta - K\cos\phi$ is positive, the single diffracted wave in this analysis is forward-diffracted and the grating is called a transmission grating. If $k_2\cos\theta - K\cos\phi$ is negative, the single diffracted wave is backward-diffracted and the grating is called a reflection grating. For the forward-diffracted case, the boundary condition used is $S_1(0)=0$. In the backward-diffracted case, the boundary condition used is $S_1(d)=0$. Due to the first-order nature of this theory, some phase matched waves will be neglected. In the transmission grating case, for example, the two backward-traveling waves (that are phase matched to the zero-order transmitted wave and the fundamental diffracted wave) are neglected.

For the special case of a phase grating with unslanted fringes ($\phi=\pi/2$) and incidence at the first Bragg angle ($m=1$), the first-order diffracted amplitude from (45) and (46) is given by [2, 4]

$$S_1(z) = -j\sin\left(\frac{\pi\varepsilon_1 z}{2(\varepsilon_0)^{1/2}\lambda\cos\theta}\right), \tag{47}$$

where z is the distance into the grating at which the amplitude is determined. This well-known expression predicts a diffraction efficiency [DE $=S_i(d)S_i^*(d)$ for this case] that is sinusoidal in modulation and has a maximum value of 100%. Although the two-wave first-order coupled-wave theory neglects higher-order diffracted waves and second derivatives of field amplitudes (and thus also boundary effects), it nevertheless contains many of the basic features of the diffraction process in an extended grating. This theory has been sucessfully extended to numerous other cases including finite beams [39, 40], finite and nonplanar gratings [41–43], and attenuated gratings [29, 44–46]. When grating diffraction is described by the two-wave result, (47), it is often referred to as "Bragg regime" diffraction. Incidence at the Bragg angle is essential in "Bragg regime" diffraction whereas in "Raman Nath regime" diffraction described below it is not. Criteria for "Bragg regime" behavior are given in [47].

For the basic case of a uniform grating and plane wave, a comparison of diffraction efficiency results from two-wave first-order coupled-wave theory with exact rigorous theory is presented in [8] for a series of grating slant angles. When transmission grating behavior dominates, the error in two-wave coupled-wave theory. is due primarily to the neglect of higher-order waves in the theory. Conversely, when reflection grating behavior dominates, the error is primarily due to the neglect of second derivatives and boundary effects.

8.5. Raman-Nath Theory

The theory of Raman and Nath [32–35] may also be obtained directly from the rigorous coupled-wave equations. If second derivatives of the field amplitudes and dephasing from the Bragg condition are both neglected, the rigorous coupled-wave equations (11) reduce to the Raman-Nath diffraction equations:

$$-j\frac{2}{\pi}\left[\frac{(\varepsilon_0)^{1/2}\cos\theta}{\lambda} - \frac{i\cos\phi}{\Lambda}\right]\frac{dS_i(z)}{dz} + \frac{\varepsilon_1}{\lambda^2}[S_{i+1}(z) + S_{i-1}(z)] = 0, \tag{48}$$

where a general angle of incidence and grating slant angle have been retained. The S_i term in (11) has been neglected. For the i-th diffracted order, this term is zero for the m-th Bragg incidence, (12), when $i=m$. For an arbitrary angle of incidence, each diffracted order will be dephased by varying amounts from their corresponding Bragg conditions. This in turn reduces the synchronism between the input wave and that diffracted order. The result is less coupling from the input to that order. The Raman-Nath theory therefore, treats all diffracted orders as though the Bragg conditions for all them were simultaneously satisfield.

For the important case of an unslanted fringe transmission grating ($\phi=\pi/2$), (48) takes the form of a recurrence relation satisfied by Bessel functions. The solution is

$$S_i(z) = (-j)^i J_i\left(\frac{\pi\varepsilon_1 z}{(\varepsilon_0)^{1/2}\lambda\cos\theta}\right) \tag{49}$$

for boundary conditions $S_0(0)=1$ and $S_i(0)=0$ $(i\neq 0)$ where J_i is an integer-order Bessel function of the first kind. Equation (49) is the famous Bessel function expression of Raman and Nath. It predicts maximum diffraction efficiencies of $DE_{\pm 1}=33.8\%$, $DE_{\pm 2}=23.6\%$, $DE_{\pm 3}=18.8\%$, and so forth. When grating diffraction behavior may be approximated by (49), it is referred to as "Raman-Nath regime" diffraction. This result, (49), has been extensively used to predict the light intensities diffracted by sound waves [38, 48]. Criteria for "Raman-Nath regime" diffraction are given in [49]. Raman-Nath theory has been extended to describe nonsinusoidal phase gratings [50–52].

8.6. Amplitude Transmittance Theory

For gratings, the amplitude transmittance approach is closely related to Raman-Nath diffraction theory. The amplitude transmittance approach is widely used in optics [53–54] and may be applied to slabs, lenses, apertures, and general two-dimensional objects as well as gratings. The amplitude transmittance is defined as the ratio of the field amplitude over the output plane to the field amplitude incident on the input plane. The amplitude transmittance function in general is complex. It may be applied to gratings with unslanted fringes. Both amplitude gratings [53–57] and phase gratings [51–56] have been treated in the literature using the amplitude transmittance approach.

For a phase grating with periodicity in the x direction, the amplitude transmittance function is

$$\tau(x,z)=\exp\left(-j\frac{2\pi[\varepsilon(x)]^{1/2}z}{\lambda\cos\theta}\right), \tag{50}$$

where z is the grating thickness and $n(x)=[\varepsilon(x)]^{1/2}$ is the periodic refractive index. Since the transmittance function is also periodic in x, it may be expanded in a complex Fourier series. Further, because the exponentials in this series are in the form of an expansion of the diffracted plane waves, then the Fourier coefficients are the diffracted wave amplitudes. The Fourier series expansion may thus be written

$$\tau(x,z)=\sum_i S_i(z)\exp(\mathrm{j}iKx), \tag{51}$$

where S_i represents the amplitude of the i-th diffracted order. By definition, the coefficients of the Fourier series may be calculated from

$$S_i(z)=\frac{1}{\Lambda}\int_0^{\Lambda}\exp\left(-\mathrm{j}\frac{2\pi[\varepsilon(x)]^{1/2}z}{\lambda\cos\theta}\right)\exp(-\mathrm{j}iKx)\,dx. \tag{52}$$

Thus the diffracted amplitudes may be determined directly knowing $\varepsilon(x)$ by integrating (52). Results for sinusoidal, square-wave, sawtooth, triangular, and rectangular refractive-index profiles are given in [51].

For the unslanted-fringe (co)sinusoidal-permittivity transmission grating, the corresponding index of refraction is

$$n(x)\equiv[\varepsilon(x)]^{1/2}=(\varepsilon_0+\varepsilon_1\cos Kx)^{1/2} \tag{53}$$

which may be expanded in a Fourier cosine series as

$$[\varepsilon(x)]^{1/2}=[\varepsilon(x)]_0^{1/2}+\sum_{h=1}^{\infty}[\varepsilon(x)]_h^{1/2}\cos(hKx) \tag{54}$$

with Fourier harmonic amplitudes given by

$$[\varepsilon(x)]_h^{1/2}=\frac{2}{\Lambda}\int_0^{\Lambda}(\varepsilon_0+\varepsilon_1\cos Kx)^{1/2}\cos(hKx)\,dx. \tag{55}$$

The average value of the refractive index may be expressed concisely as

$$n_0(x)\equiv[\varepsilon(x)]_0^{1/2}=(2/\pi)(\varepsilon_0+\varepsilon_1)^{1/2}E(\zeta,\pi/2), \tag{56}$$

where $E(\zeta,\pi/2)$ is the complete elliptic integral of the second kind and $\zeta\equiv 2\varepsilon_1/(\varepsilon_0+\varepsilon_1)$. Clearly, the case of a sinusoidal permittivity (or dielectric constant) being treated throughout this paper is not the same as a sinusoidal refractive-index grating. The index of refraction corresponding to sinusoidal permittivity has higher spatial frequency harmonics ($h>1$) in addition to a fundamental sinusoidal component ($h=1$) as represented by (54). However, for the case of sufficiently small modulation, a sinusoidal permittivity produces nearly a sinusoidal index of refraction. In the limit of small modulation (ε_1 approaches zero), (55) and (56) yield

$$[\varepsilon(x)]_0^{1/2}\simeq\varepsilon_0, \tag{57}$$

$$[\varepsilon(x)]_1^{1/2}\simeq\varepsilon_1/2(\varepsilon_0)^{1/2}, \tag{58}$$

$$[\varepsilon(x)]_2^{1/2},\quad [\varepsilon(x)]_3^{1/2},\ldots\simeq 0. \tag{59}$$

This analysis is important in that it now allows the Raman-Nath theory and amplitude transmittance theory to be interrelated. The result is that although (52) was obtained using the amplitude transmittance approach, it is *also* a solution of the Raman-Nath difference-differential equation (48) for unslanted gratings in the limit of small modulation. This may be shown by direct substitution of S_i as given by (52) into the Raman-Nath diffraction equation (48). Thus, for a cosinusoidal refractive-index profile, the integral in (52) when evaluated gives the Bessel function result (49). This may be accomplished using the identity

$$\exp(-\mathrm{j}b\cos\alpha)\equiv\sum_{i=-\infty}^{+\infty}(-\mathrm{j})^i J_i(b)\exp(\mathrm{j}i\alpha) \tag{60}$$

and the orthogonality relationship

$$\frac{1}{\Lambda}\int_0^{\Lambda}\exp(\mathrm{j}lKx)\exp(-\mathrm{j}iKx)\,dx=\delta_{li}, \tag{61}$$

where δ_{li} is the Kronecker delta. Therefore, as depicted in Fig. 3, it has been shown that Raman-Nath theory and amplitude transmittance theory are equivalent in the limit of small grating modulation. This is true in general for unslanted gratings regardless of the grating profile (squarewave, sawtooth, etc.).

9. Discussion and Conclusions

Theories describing the diffraction of a plane electromagnetic wave incident with arbitrary wavelength and angle of incidence upon a slanted fringe planar sinusoidal permittivity grating have been reviewed. For simplicity the analysis has been restricted to H-mode polarization (electric field perpendicular to plane of incidence). Exact formulations without approximations (rigorous coupled-wave analysis and rigorous modal analysis) have been developed and shown to be mathematically equivalent. Rigorous coupled-wave equations have been developed in alternative forms, (11) and (14), and their usefulness discussed. The solution of the rigorous coupled-wave equations (11) in terms of state variables has been presented in detail along with the phase matching and boundary conditions necessary to determine the diffracted amplitudes outside of the grating.

These rigorous theories have been shown to reduce to the approximate theories: 1) two-wave modal theory, 2) two-wave second-order coupled-wave theory, 3) multiwave first-order coupled-wave theory, 4) two-wave first-order coupled-wave theory (Kogelnik theory), 5) Raman-Nath theory, and 6) amplitude transmittance theory in the appropriate limits. The assumptions associated with each of these approximate theories have been explicitly presented.

The rigorous theories presented in this paper (and thus their approximate versions) are based on the Floquet theorem. As such, they require a truly periodic grating (an infinite number of periods). These theories may be applied in the angular limit as the slanted fringes of a general grating approach being parallel to the surface (ϕ approaches zero). However, for exact parallelism with the surface ($\phi=0$), the grating is no longer strictly periodic and a continuum of solutions is possible depending on the number of periods, the starting conditions, and the ending conditions of the grating. This pure reflection grating case can be analyzed without approximation using a rigorous chain-matrix method of analysis. This is discussed in [58].

Acknowledgements. This work was sponsored by the National Science Foundation and by the Joint Services Electronics Program.

References

1. R.R. Aggarwal: Proc. Indian Acad. Sci. **31**, 417–426 (1950)
2. P. Phariseau: Proc. Indiana Acad. Sci. A**44**, 165–170 (1956)
3. W.R. Klein, B.D. Cook: IEEE Trans. SU**14**, 123–134 (1967)
4. H. Kogelnik: Bell Syst. Tech. J. **48**, 2909–2947 (1969)
5. G.L. Fillmore, R.F. Tynan: J. Opt. Soc. Am. **61**, 199–203 (1971)
6. J.A. Kong: J. Opt. Soc. Am. **67**, 825–829 (1977)
7. R. Magnusson, T.K. Gaylord: J. Opt. Soc. Am. **67**, 1165–1170 (1977)
8. M.G. Moharam, T.K. Gaylord: J. Opt. Soc. Am. **71**, 811–818 (1981)
9. T. Tamir, H.C. Wang, A.A. Oliner: IEEE Trans. MTT-**12**, 323–335 (1964)
10. T. Tamir, H.C. Wang: Can. J. Phys. **44**, 2073–2094 (1966)
11. T. Tamir: Can. J. Phys. **44**, 2461–2494 (1966)
12. C.B. Burckhardt: J. Opt. Soc. Am. **56**, 1502–1509 (1966)
13. L. Bergstein, D. Kermisch: Proc. Symp. Mod. Opt. **17**, 655–680 (1967)
14. R.S. Chu, T. Tamir: IEEE Trans. MTT-**18**, 486–504 (1970)
15. R.S. Chu, T. Tamir: Proc. IEE **119**, 797–806 (1972)
16. F.G. Kaspar: J. Opt. Soc. Am. **63**, 37–45 (1973)
17. S.T. Peng, T. Tamir, H. L. Bertoni: IEEE Trans. MTT-**23**, 123–133 (1975)
18. R.S. Chu, J.A. Kong: IEEE Trans. MTT-**25**, 18–24 (1977)
19. P.St.J. Russell: Phys. Rep. **71**, 209–312 (1981)
20. L.Solymar, D.J. Cooke: *Volume Holography and Volume Gratings* (Academic Press, London 1981)
21. R. Magnusson, T.K. Gaylord: J. Opt. Soc. Am. **68**, 1777–1779 (1978)
22. D.A. Watkins: *Topics in Electromagnetic Theory* (Wiley, New York 1958) p. 2
23. R.E. Collin: *Field Theory of Guided Waves* (McGraw-Hill, New York 1960) p. 368
24. E.g.C.L. Liu, J.W.S. Liu: *Linear Systems Analysis* (McGraw-Hill, New York 1975)
25. E.g. program EIGRF from the International Mathematics and Statistics Library (IMSL)
26. F. Lederer, U. Langbein: Opt. Acta **27**, 183–200 (1980)
27. S. Kessler, R. Kowarschik: Opt. Quantum Electron. **7**, 1–4 (1975)
28. R. Kowarschik, S. Kessler: Opt. Quantum Electron. **7**, 399–411 (1975)
29. R. Kowarschik: Opt. Acta **23**, 1039–1051 (1976)
30. T. Jaaskelainen, O. Salminen, R.M.K. Hamalainen: Acta Poly.-Appl. Phys. **126**, 3–12 (1979)
31. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A**3**, 119–125 (1936)
32. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A**2**, 406–412 (1935)
33. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A**2**, 413–420 (1935)
34. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A**3**, 75–84 (1936)
35. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A**3**, 459–465 (1936)
36. N.S.N. Nath: Proc. Indian Acad. Sci. A**8**, 499–503 (1938)
37. N.S.N. Nath: Proc. Indian Acad. Sci. A**4**, 222–242 (1936)
38. M.V. Berry: *The Diffraction of Light by Ultrasound* (Academic Press, London 1966)
39. P.St.J. Russell, L. Solymar, M.P. Jordan: Proc. ICO **11**, 635–638 (1978)
40. M.G. Moharam, T.K. Gaylord, R. Magnusson: J. Opt. Soc. Am. **70**, 300–304 (1980)
41. L. Solymar, M.P. Jordan: Microwave Opt. Acoust. **1**, 89–92 (1977)
42. R.P. Kenan: IEEE J. QE-**14**, 924–930 (1978)
43. M.G. Moharam, T.K. Gaylord, R. Magnusson: J. Opt. Soc. Am. **70**, 437–442 (1980)
44. N. Uchida: J. Opt. Soc. Am. **63**, 280–287 (1973)
45. U. Killat: Opt. Commun. **21**, 110–111 (1977)
46. M.P. Owen, L. Solymar: Opt. Commun. **34**, 321–326 (1980)

47. M.G. Moharam, T.K. Gaylord, R. Magnusson: Opt. Commun. **32**, 19–23 (1980)
48. F.H. Sanders: Can. J. Phys. **14**, 158–171 (1936)
49. M.G. Moharam, T.K. Gaylord, R. Magnusson: Opt. Commun. **32**, 14–18 (1980)
50. K.L. Zankel, E.A. Hiedemann: J. Acoust. Soc. Am. **31**, 44–54 (1959)
51. R. Magnusson, T.K. Gaylord: J. Opt. Soc. Am. **68**, 806–809 (1978)
52. R. Magnusson, T.K. Gaylord: Opt. Commun. **25**, 129–132 (1978).
53. J.W. Goodman: *Introduction to Fourier Optics* (McGraw-Hill, New York 1968)
54. W.T. Cathey: *Optical Information Processing and Holography* (Wiley, New York 1974)
55. H.M. Smith: *Principles of Holography* (Wiley, New York 1969)
56. R.J. Collier, C.B. Burckhardt, L.H. Lin: *Optical Holography* (Academic Press, New York 1971)
57. R. Magnusson, T.K. Gaylord: Opt. Commun. **28**, 1–3 (1979)
58. M.G. Moharam, T.K. Gaylord: J. Opt. Soc. Am. **72**, 187–190 (1982)