On the use of the energy balance criterion as a check of validity of computations in grating theory
R. Petit and G. Tayeb
Laboratoire d'Optique Electromagnétique, U.A. CNRS n° 843, Faculté des Sciences, Centre de St-Jérôme, 13397 Marseille Cedex 13, France

ABSTRACT

We consider rigorous theories of diffraction gratings in which the electromagnetic field can be considered as infinite series. Numerical implementation of these theories needs truncation of the series. We show that in many circumstances, some properties of the exact solution (conservation of energy, reciprocity relations) are also verified by the truncated solution. Consequently these properties can not be systematically used as a check of validity of the truncated solution.

1. DEFINITION OF THE GRATING PROBLEM

We deal with time harmonic fields represented by complex vectors taking into account a time dependence in \( \exp(-i\omega t) \). We use a rectangular coordinate system Oxyz and denote by \( \hat{e}_x, \hat{e}_y, \hat{e}_z \) the unit vectors of x, y, z axes. The permeability is equal to \( \mu_0 \) everywhere. Referring to fig.1, we consider a structure composed of three regions. The superstrate \((y > a)\) and the substrate \((y < 0)\) are homogeneous regions of relative permittivities \( \varepsilon_1 \) and \( \varepsilon_3 \) (\( \varepsilon_1 \) is supposed to be real). In the region \( 0 < y < a \), the relative permittivity \( \varepsilon \) is a function of \( x \) and \( y \) which is, for fixed \( y \), periodic with respect to \( x \) (period \( d \)). Some particular cases of this periodic structure are for instance the coated grating (fig.2) and the "slanted" grating. 1 The grating is illuminated by a plane wave and we denote the total field by \( E \) and \( H \). We assume for the sake of simplicity that the incident wave vector lies in the \( xy \) plane and that consequently the fields are \( z \) independent. We put \( k_0 = \omega / \varepsilon_0 \mu_0 \). In all the paper, \( \bar{Z} \) denotes the complex conjugate of \( Z \), and \( \delta_{nm} \) is the kronecker symbol.

\[ E_x(x,y) = \sum_{n=-\infty}^{\infty} E_{x_n}(y) \exp(i\alpha_n x), \]

with \( \alpha_n = \alpha_0 + nX \), \( \alpha_0 = \sqrt{\varepsilon_1} k_0 \sin \theta \), \( k_0 = \omega / \sqrt{\varepsilon_0 \mu_0} = 2\pi/\lambda \) and \( X = 2\pi/d \).

The right member of (1) can be interpreted as a generalized Fourier series, and consequently we will say that the \( E_{x_n}(y) \) are the Fourier coefficients of the field \( E_x(x,y) \). Noticing that \( E_x(x+d, y) = \exp(i\alpha_0 d) E_x(x,y) \), we say that \( \exp(i\alpha_0 d) \) is the pseudo periodicity coefficient.

of the field. The Fourier coefficients are taken as the unknowns of the problem. We will subsequently denote by \( |E_n| \) the infinite column matrix whose elements are the \( E_n \). This matrix is a function of \( y \). The same notation will be used for the other field's components.

Elementary calculations detailed in appendix A show that:

1) The two components \( E_y \) and \( H_z \) can be deduced from the four others (eq.(A1) and (A2)); we will take interest only in \( E_x, E_z, H_x, H_z \), and more precisely in the column matrices \( |E_x|, |E_z|, |H_x| \) and \( |H_z| \).

2) Maxwell equations can be written in the form:

\[
\begin{align*}
\frac{d}{dy} |E_x| &= A(y) |H_z|, \\
\frac{d}{dy} |E_z| &= B(y) |H_x|, \\
\frac{d}{dy} |H_x| &= C(y) |E_z|, \\
\frac{d}{dy} |H_z| &= D(y) |E_x|,
\end{align*}
\]

where \( A, B, C, D \) are "infinite square" matrices.

It is worth noting that the equations (2) to (5) are valid in the sense of distributions, which means that the continuity relations on the surfaces where \( \epsilon \) is discontinuous (i.e. the boundary conditions) are automatically taken into account. In general, matrices \( A, B, C, D \) are \( y \) dependent, and eq.(2) to (5) have to be integrated between \( y = 0 \) and \( y = a \) using a numerical algorithm.

3. EXPRESSION OF THE LORENTZ RECIPROCITY THEOREM

Let us consider two solutions \( (E, H) \) and \( (E', H') \) of the harmonic Maxwell equations in a volume \( V \) bounded by the closed surface \( S \). If there is no current distribution in \( V \), and if \( \hat{n} \) denotes the unit normal of \( S \), the Lorentz reciprocity theorem takes the form:

\[
\int_S (\hat{E} \wedge \hat{H}' - \hat{E}' \wedge \hat{H}).\hat{n} \, dS = 0. \tag{6}
\]

If \( S \) is the surface depicted in fig.3, then, since the fields are \( z \) independent:

\[
\int_{A_1B_1B_2A_2} (\hat{E} \wedge \hat{H}' - \hat{E}' \wedge \hat{H}).\hat{n} \, dl = 0 \tag{7}
\]

Let us suppose now that \( (E, H) \) is a pseudo-periodic field which can be written as in (1), and that \( (E', H') \) is another pseudo-periodic field whose components can be written as:

\[
E'_x(x,y) = \sum_{n=-\infty}^{+\infty} E_{Xn}(y) \exp(-i\alpha_n x). \tag{8}
\]

Let us note that \( (E', H') \) has a pseudo periodicity coefficient inverse of that of \( (E, H) \). It is also worth noting that, if \( (E, H) \) represents for instance the electromagnetic field corresponding to the propagating plane waves depicted fig.4a, \( (E', H') \) can represent the electromagnetic field of another problem where the incident plane wave falls on the grating with an incidence \( \theta_p \). Fig. 4b illustrates this case, supposing \( p = 1 \).

As shown in appendix B, (7) implies that the integral:

\[
I = \int_0^d (\hat{E} \wedge \hat{H}') - \hat{E}' \wedge \hat{H}).\hat{e}_y \, dx, \quad \text{is } y \text{ independent } \quad \frac{dy(y)}{dy} = 0. \tag{9}
\]
Figures 4. Only the plane waves propagating in the superstrate are represented. \( \lambda \) is the wavelength in the superstrate. The radius of the circle is equal to unity. The incidence angles are measured anticlockwise, while the diffracted angles are measured clockwise.

We want to emphasize that the property (9), which is true for the exact solution of the problem, still holds true (see appendix B) for a truncated solution which verifies the truncated equations (2) to (5). We call truncated solution a solution obtained assuming that a Fourier series such as (1) or (8) can be replaced by a finite sum, an assumption which is obviously necessary for the numerical implementation.

4. CONSERVATION OF ENERGY

In this section, we deal with unlossy materials, consequently \( \varepsilon(x,y) \) is a real function. In eq.(9), \( (E', H') \) represents an electromagnetic field solution of the harmonic Maxwell equations, which has a coefficient of pseudo periodicity inverse of that of \( (E, H) \). It is easy to show (Appendix C) that \( (E' = E, H' = -H) \) verifies these conditions. We make this choice in this section, and from (9), we can claim that:

\[
\int_0^d (E \cdot H + E' \cdot H') \cdot e_y \, dx = 2 \int_0^d \text{Re} (E \cdot H) \cdot e_y \, dx, \quad \text{is independent of } y. \tag{10}
\]

We recognize here an expression of the Poynting theorem *, which as it is well known, leads to the famous energy balance criterion **. Therefore, as a consequence of considerations developed in section 3, it appears that this criterion is verified not only by the exact solution, but also by any truncated solution verifying the equations deduced from (2) to (5) by truncature. To lay stress on this rather amazing result, consider for instance a grating problem in which the incident wave gives rise to \( N \) propagating diffracted waves in the superstrate. If the approximate numerical solution is obtained by retaining only \( P \) Fourier coefficients for the fields \( (P \) being possibly less than \( N \), and even equal to 1), the energy balance criterion will be verified whatever \( P \), provided that the integration of eq.(2) to (5) between \( y = 0 \) and \( y = a \) is accurately performed. Indeed, as claimed by Nevière et al., and at least for certain integration algorithms, it seems that the energy balance criterion is verified whatever the integration step. Anyway, in all the computations we have done using the differential method, this criterion has been verified with a precision of the order of \( 10^{-5} \) (even when using unreasonable truncatures and very large integration steps; cf table 1.)

* Let us remark that (10) involves a spatially averaged (on the grating spacing \( d \)) Poynting vector.

** i.e. the sum of all the diffracted efficiencies is equal to unity.
Table 1. Relative to a sinusoidal grating described fig.2, with: 

e = 0, d = 1 \mu m, h = 0.2 \mu m, \lambda = 0.6 \mu m, \theta = 20^\circ, \text{superstrate is vacuum, } \varepsilon = 2.25 \text{ for the substrate. } P \text{ is the number of Fourier coefficients used to represent the fields.}

Polarization: \vec{E} \text{ is parallel to Oz (TE case). Integration between } y = 0 \text{ and } y = a \text{ is performed using 38 steps with Runge-Kutta (4th order) algorithm.}

5. RECIPROCITY

In both the superstrate (y > a) and the substrate (y < 0), each Fourier coefficient of the fields can be written as the product of a constant (usually called the Rayleigh coefficient) by a function exponentially dependent on y.

Figure 5. In case 1, the grating G is illuminated under the incidence \theta and we turn our attention to the pth diffracted wave (angle \theta_p). In case 2, G is illuminated under the incidence \theta' = -\theta. The reciprocity theorem claims that the pth diffracted wave in case 2 and the incident wave in case 1 propagate in opposite directions, and that the efficiencies in the pth order are the same (say e_p) in both cases. The theorem holds also for a transmitted order, when dealing with lossless substrates.

Table 2. Relative to a sinusoidal grating described fig.2, with: e = 0, d = 1 \mu m, h = 0.2 \mu m, \lambda = 0.6 \mu m, \theta = 20^\circ, \text{superstrate is vacuum, } \varepsilon = 2.25 \text{ for the substrate. The notations are those of fig.5, for the transmitted wave in the first order (p = 1). In case 1, } \theta = 20^\circ. \text{ In case 2, the incident field comes from the substrate, with } \theta' = 38.904^\circ. \text{ P is the number of Fourier coefficients used to represent the fields. We denote by TE (resp. TM) the situation where } \vec{E} \text{ (resp. } \vec{H} \text{) is parallel to Oz. The table gives the values of } e_{p1}. \text{ Integration between } y = 0 \text{ and } y = a \text{ is performed using 40 steps with Adams-Moulton (initiated by Runge-Kutta) algorithm.}
For \( y > a \) (resp. \( y < 0 \)) (see figure 1), it turns out that the integral appearing in (9) can be written as a finite sum. Each term of the sum contains the Rayleigh coefficients corresponding to the only propagating waves of the incident and reflected (resp. of the transmitted) fields. Using for \((\mathbf{E}', \mathbf{H}')\) a field associated to an incident plane wave which falls on the grating with the same direction as one of the diffracted waves of \((\mathbf{E}, \mathbf{H})\) (as explained in section 2 and fig.4), it is easy to deduce the well known reciprocity theorem (fig.5).

These considerations are developed in detail by D. Maystre for a simple case of polarization in Progress in Optics, and by A. Roger in ref. 6.

Let us now consider truncated fields, described in the superstrate and in the substrate by truncated Rayleigh developments. From section 3 we know that (9) still holds true. Consequently, provided that the truncated developments contain the orders represented fig.5.1 and fig.5.2, the reciprocity theorem still holds true. We may confess that our numerical computations does not perfectly agree with this prediction (cf table 2). Up to now, we have not been able to explain this discrepancy satisfactorily (it could perhaps be due to the numerical process).

6. THE SLANTED GRATING

In this section we show that the problem of a slanted grating described fig.6, even when solved using an eigenvalues problem, is a particular case of the differential method, and that the results of sections 3, 4, and 5 still hold true.

In the region \( 0 < y < a \), the permittivity can be written as:

\[
\varepsilon(x,y) = \sum_n \varepsilon_n(y) \exp(inKx) = \sum_n f(L \cdot x + L_y \cdot y),
\]

where \( f \) is a periodic function of period \((L_x, d)\). It is easy to show that, in these circumstances, the Fourier coefficients of \( \varepsilon \) take the form:

\[
\varepsilon_n(y) = c_n \exp(in\mu y),
\]

where

\[
c_n = \frac{L_y \cdot d}{L_x \cdot d} \int_0^{L_x \cdot d} f(t) \exp(-in \frac{K}{L_x} t) \, dt,
\]

is independent of \( y \), and \( \mu = K L_y / L_x \).

For the sake of simplicity, let us suppose that the field is polarized \( \mathbf{E} / \mathbf{OZ} \), and put

\[
u(x,y) = \mathbf{E}_x(x,y) = \sum_n u_n(y) \exp(inx).
\]

Then, system (A1) to (Ag) is equivalent to:

\[
\Delta u(x,y) + k_s^2 \varepsilon(x,y) u(x,y) = 0
\]

which implies, for any \( n \):

\[
u_n''(y) - c_n^2 u_n(y) + k_s^2 \sum_m \varepsilon_{n-m}(y) u_m(y) = 0.
\]

Putting now \( u_n(y) = w_n(y) \exp(in\mu y) \), we obtain:

\[
w_n'' + 2in\mu w_n' - (n^2\mu^2 + a_n^2) w_n + k_s^2 \sum_m c_{n-m} w_m = 0,
\]

which can be written in matrix form as:

\[
|w''| = R |w'| + S |w|,
\]

(11)

where \( R \) and \( S \) are \( y \) independent infinite matrices whose elements are:

\[
R_{nm} = -2in\mu \delta_{nm}; \quad S_{nm} = (n^2\mu^2 + a_n^2) \delta_{nm} - k_s^2 c_{n-m}.
\]

It is important to note that (11) is strictly equivalent to the system (2) to (5).

We can also write (11) in the form:

\[
Comparing (12) to its equivalent expression (2) to (5), it is worth noting that now the square matrices are $y$ independent. This remark allows us to reduce the solving of (12) to an eigenvalues problem, avoiding so a numerical integration. Anyway, except for their numerical treatment, the problems are strictly equivalent. The results of sections 3, 4, and 5 are still valid, and we agree with the paper by Russell. 7

7. A LAST EXAMPLE

Obviously, the previous considerations are reminiscent with a question which we refer to during the last SPIE Symposium. In a paper by R. Petit, J.L. Suratteau and M. Cadilhac, 8 we had to write the continuity of two functions on a bounded interval, which is equivalent to express the vanishing of two other functions (i.e. their jumps on this interval). This was done by projecting on a convenient basis in order to get an algebraic system. We can use either the same basis or two different bases. In the first case, the energy balance is not automatically satisfied after truncature. On the other hand, when using two different basis as done by the Australian group of Sidney University, 9 we are led to the opposite conclusion. 10 In our opinion, this is an important point since the energy balance is often used as a check of validity of the numerical results obtained after truncature.

APPENDIX A

$\mathbf{E}$ and $\mathbf{H}$ verify the Maxwell equations:

$$\text{curl } \mathbf{E} = i \omega \mu_0 \mathbf{H} ; \quad \text{curl } \mathbf{H} = -i \omega \varepsilon_0 \mathbf{E}.$$ 

Putting $Z_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}}$, we get:

$$E_y = -i \frac{Z_0}{k_0} \frac{1}{\varepsilon} \frac{\partial E_z}{\partial x} \quad (A1)$$

$$H_y = \frac{1}{k_0^2 Z_0} \frac{\partial E_z}{\partial x} \quad (A2)$$

$$\frac{\partial E_x}{\partial y} = -i \frac{Z_0}{k_0} \frac{\partial}{\partial x} \left[ \frac{1}{\varepsilon} \frac{\partial E_z}{\partial x} \right] - i k_0 Z_0 H_z \quad (A3)$$

$$\frac{\partial E_z}{\partial y} = i k_0 Z_0 H_x \quad (A4)$$

$$\frac{\partial H_x}{\partial y} = \frac{1}{k_0^2 Z_0} \frac{\partial^2 E_z}{\partial x^2} + 1 \frac{k_0}{Z_0} \varepsilon E_z \quad (A5)$$

$$\frac{\partial H_z}{\partial y} = -i \frac{k_0}{Z_0} \varepsilon E_x . \quad (A6)$$

Taking into account the periodicity of $\varepsilon$ and $1/\varepsilon$, we write their Fourier series:

$$\varepsilon(x,y) = \sum_{n=-\infty}^{+\infty} \varepsilon_n (y) \exp(i n K x) , \quad (A7)$$

$$\left(\frac{1}{\varepsilon}\right)(x,y) = \sum_{n=-\infty}^{+\infty} \left[\frac{1}{\varepsilon_n}\right] (y) \exp(i n K x) . \quad (A8)$$

Replacing in (A3) to (A6) the field's components by their expressions (1), $\varepsilon$ and $1/\varepsilon$ by
their expressions (A7) and (A8), and with a projection on the $\exp(i\alpha_n x)$ basis, we obtain for any value of $n$:

$$
\frac{dE_{zn}}_{\text{dy}} = i \frac{k_0}{z_0} H_{xn}
$$

(A10)

$$
\frac{dH_{zn}}{dy} = \frac{i}{\frac{1}{m=-\infty}} \left[ \frac{k_0}{z_0} H_{xn} \right]
$$

(A11)

$$
\frac{dH_{zn}}{dy} = \frac{i k_0}{\frac{1}{m=-\infty}} \left[ -\frac{k_0}{z_0} H_{xn} \right]
$$

(A12)

which can be written in matrix form:

$$
\begin{align*}
\frac{d}{dy} [E_x] &= A(y) [H_z] \\
\frac{d}{dy} [E_z] &= B [H_x] \\
\frac{d}{dy} [H_x] &= C(y) [E_z] \\
\frac{d}{dy} [H_z] &= D(y) [E_x]
\end{align*}
$$

(A13) (A14) (A15) (A16)

where $A, B, C, D$ are infinite matrices whose elements are given by:

$$
A_{nm}(y) = i \frac{k}{k_0} a_n a_m \left[ \frac{1}{c} \right]_{n-m} - i k_0 \frac{z_0}{c} m
$$

(A17)

$$
B_{nm} = i k_0 \frac{z}{c} m
$$

(A18)

$$
C_{nm}(y) = -\frac{i k_0}{k_0} a_n a_m \left[ \frac{1}{c} \right]_{n-m} + i k_0 \frac{z}{c} m
$$

(A19)

$$
D_{nm}(y) = -\frac{i k_0}{k_0} a_n a_m \left[ \frac{1}{c} \right]_{n-m}
$$

(A20)

APPENDIX B

In this section we propose to calculate the quantity

$$
J = \int \left[ \mathbf{E} \cdot \mathbf{H}' - \mathbf{E}' \cdot \mathbf{H} \right] \cdot dy 
$$

in terms of the Fourier coefficients of the fields.

Let $u$ be a component of $(\mathbf{E}, \mathbf{H})$ and $u'$ a component of $(\mathbf{E}', \mathbf{H}')$; eqs. (1) and (8) show that:

$$
\begin{align*}
u(x + d, y) &= \exp(i \alpha_d d) u(x,y) , \\
u'(x + d, y) &= \exp(-i \alpha_d d) u'(x,y) ,
\end{align*}
$$

(B1) (B2)

which means that $\mathbf{E} \cdot \mathbf{H}'$ and $\mathbf{E}' \cdot \mathbf{H}$ are $d$ periodic with respect to $x$. This implies that the contributions of segments $(B_1 B_2)$ and $(A_2 A_1)$ cancel each other, and $J$ takes the form:

$$
J = \int_{A_1 B_1} \left[ \mathbf{E} \cdot \mathbf{H}' - \mathbf{E}' \cdot \mathbf{H} \right] \cdot dy - \int_{A_2 B_2} \left[ \mathbf{E} \cdot \mathbf{H}' - \mathbf{E}' \cdot \mathbf{H} \right] \cdot dy = I(y_1) - I(y_2)
$$

(B3)

with

$$
I(y) = \int_{0}^{d} \left[ \mathbf{E}(x,y) \cdot \mathbf{H}'(x,y) - \mathbf{E}'(x,y) \cdot \mathbf{H}(x,y) \right] \cdot \mathbf{d} y
$$

(B4)
Let us now calculate \( \frac{dI(y)}{dy} \); developing the vector-products, we find:

\[
\frac{dI(y)}{dy} = \int_{0}^{d} \left[ \left( \frac{\partial E_{x}}{\partial y} H'_{x} - \frac{\partial E'_{x}}{\partial y} H_{x} \right) + \left( \frac{\partial E_{y}}{\partial y} H'_{y} - \frac{\partial E'_{y}}{\partial y} H_{y} \right) \right. \\
\left. - \left( \frac{\partial E_{z}}{\partial y} H'_{z} - \frac{\partial E'_{z}}{\partial y} H_{z} \right) \right] \ dx 
\]

(B5)

Let us study for example the quantity \( \frac{dI_{3}(y)}{dy} = \int_{0}^{d} \left( - \frac{\partial E_{x}}{\partial y} H'_{z} + \frac{\partial E'_{z}}{\partial y} H_{z} \right) \ dx \).

Using eq. (1) and (8), we get:

\[
\frac{dI_{3}(y)}{dy} = \int_{0}^{d} \sum_{n,m} \left( - \frac{\partial E_{x}}{\partial y} H'_{z} + \frac{\partial E'_{z}}{\partial y} H_{z} \right) \ e^{i(n-m)Kx} \ dx . 
\]

Noting that \( \int_{0}^{d} e^{i(n-m)Kx} \ dx = d \delta_{nm} \), we obtain:

\[
\frac{dI_{1}(y)}{dy} = d \sum_{n} \left( - \frac{\partial E_{x}}{\partial y} H'_{z} + \frac{\partial E'_{z}}{\partial y} H_{z} \right) . 
\]

(B6)

We have shown in Appendix A that \( \frac{\partial E_{x}}{\partial y} = \sum_{m} A_{nm} H_{zm} \). In the same way, we find that:

\[
\frac{\partial E'_{x}}{\partial y} = \sum_{m} A'_{nm} H_{zm} . 
\]

(a)

Eq. (B6) becomes:

\[
\frac{dI_{1}(y)}{dy} = d \sum_{n,m} A_{nm} H_{zm} H'_{zn} + A'_{nm} H_{zm} H_{zn} \\
= d \sum_{m=\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} \left( - A_{nm} + A'_{nm} \right) H_{zm} H'_{zn} . 
\]

(B8)

It is now a trivial matter to verify that \( A_{nm} = A'_{nm} \) and that consequently each term of the double series in (B8) vanishes. The same can be done for the three other terms in (B5) which also vanish. The important result is the following: the quantity \( I(y) \) is independent of \( y \) even if the developments of the fields in generalized Fourier series are truncated.

**APPENDIX C**

(a) Each component of \( \vec{E}, \vec{H} \), denoted for instance by \( u(x,y) \) expresses as:

\[
u(x,y) = \sum_{n} \tilde{u}_{n}(y) \ \exp(i\alpha_{n}x) ; \text{ and consequently } (\alpha_{n} \text{ is real}) : \\
u(x,y) = \sum_{n} \tilde{u}_{n}(y) \ \exp(-i\alpha_{n}x) . 
\]

(b) If \( \vec{E}, \vec{H} \) verifies the harmonic Maxwell equations: curl \( \vec{E} = i\omega \mu_{0} \vec{H} \);

\( \text{curl } \vec{H} = -i\omega \varepsilon \vec{E} \), it is clear that \( \vec{E}, \vec{H} \) verifies the same equations in the case where \( \varepsilon \) is real.
ACKNOWLEDGMENTS

Thanks are due to Professor M. Cadilhac for valuable discussions and advices.

The computing facilities used in this paper have been given by the Scientific Council of the "Centre de Calcul Vectoriel pour la Recherche".

REFERENCES