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,[1,2] 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[3] 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[4] and Granet and Guizal[5] 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 Σ 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,[1,2] one solves the coupled first-order differential system,

$$\frac{1}{i}\frac{\mathrm{d}H_{zn}}{\mathrm{d}y} = -k_0 \sum_m \epsilon_{n-m} E_{xm}, \tag{1a}$$

$$\frac{1}{i}\frac{\mathrm{d}E_{xn}}{\mathrm{d}y} = -k_0\mu_0 H_{zn} + \frac{\alpha_n}{k_0}\sum_m \left(\frac{1}{\epsilon}\right)_{n-m} \alpha_m H_{zm}, \tag{1b}$$

or better yet, the equivalent second-order system

$$\frac{\mathrm{d}^2 H_{zn}}{\mathrm{d}y^2} = \sum_m \epsilon_{n-m} \sum_p \left[\alpha_m \left(\frac{1}{\epsilon}\right)_{m-p} \alpha_p - \mu_0 {k_0}^2 \delta_{mp}\right] H_{zp}. \tag{2}$$

Here, k_0 is the vacuum wave number; $\mu_0 = 1$; δ_{mp} is the Kronecker symbol; ϵ_n and $(1/\epsilon)_n$ are the Fourier coefficients of the permittivity and the reciprocal permittivity functions, respectively; E_{xn} and H_{zn} are the y-dependent Fourier coefficients of the fields; and $\alpha_n = \alpha_0 + nK$, with $K = 2\pi/d$ and α_0 being the Floquet exponent. In the new formulation,[4,5] one solves the coupled first-order system,

$$\frac{1}{i}\frac{\mathrm{d}H_{zn}}{\mathrm{d}y} = -k_0 \sum_m \left[\!\left[\frac{1}{\epsilon}\right]\!\right]^{-1}_{nm} E_{xm}, \tag{3a}$$

$$\frac{1}{i}\frac{\mathrm{d}E_{xn}}{\mathrm{d}y} = -k_0\mu_0 H_{zn} + \frac{\alpha_n}{k_0}\sum_m [\![\epsilon]\!]^{-1}_{nm} \alpha_m H_{zm}, \tag{3b}$$

or the second-order system,

$$\frac{\mathrm{d}^2 H_{zn}}{\mathrm{d}y^2} = \sum_m \left[\!\left[\frac{1}{\epsilon}\right]\!\right]^{-1}_{nm} \sum_p (\alpha_m [\![\epsilon]\!]^{-1}_{mp} \alpha_p - \mu_0 {k_0}^2 \delta_{mp}) H_{zp}, \tag{4}$$

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.*,[6] in which the matrix $[\![1/\epsilon]\!]$ in Eq. (2) is replaced by $[\![\epsilon]\!]^{-1}$:

0740-3232/96/0901870-07$10.00

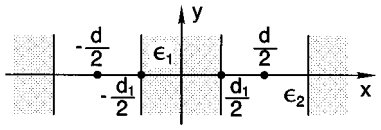

Fig. 1. Periodic, piecewise-constant medium. The periodicity of the permittivity is d, and its discontinuities are located at $x = \pm d_1/2$.

$$\frac{\mathrm{d}^2 H_{zn}}{\mathrm{d}y^2} = \sum_m \epsilon_{n-m} \sum_p (\alpha_m [\![\epsilon]\!]^{-1}_{mp} \alpha_p - \mu_0 k_0{}^2 \delta_{mp}) H_{zp}. \tag{5}$$

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 **P** be the set of piecewise-continuous, piecewise-smooth, bounded, periodic functions of x with period 2π. For every $f(x) \in \mathbf{P}$ and $g(x) \in \mathbf{P}$,

$$h(x) = f(x)g(x) \tag{6}$$

is obviously also in **P**. Let

$$U_f = \{x_j | f(x_j + 0) \neq f(x_j - 0), \qquad j = 1, 2, \ldots\} \tag{7}$$

be the set of the abscissas of the discontinuities of $f(x)$, and let U_g be similarly defined for $g(x)$. Then,

$$U_{fg} = U_f \cap U_g \tag{8}$$

is the set of the abscissas of the concurrent discontinuities of $f(x)$ and $g(x)$. If $h(x)$ is such that

$$h(x_p - 0) = h(x_p + 0) \quad (x_p \in U_{fg}), \tag{9}$$

$f(x)$ 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 $\hat{f}_j$,

$$\hat{f}_j = f(x_j + 0) - f(x_j - 0), \tag{10}$$

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 **P**, the Fourier coefficients of $h(x)$ can be obtained from the Fourier coefficients of $f(x)$ and $g(x)$ by Laurent's rule:[8]

$$h_n = \sum_{m=-\infty}^{+\infty} f_{n-m} g_m. \tag{11}$$

The Fourier factorization of $h(x)$ is then given by

$$h(x) = \sum_{n=-\infty}^{+\infty} h_n \exp(inx) = \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} f_{n-m} g_m \exp(inx). \tag{12}$$

To be more precise, Eq. (12) should be understood in the following sense:

$$h(x) = \lim_{N\to\infty} \sum_{n=-N}^{N} \left(\lim_{M\to\infty} \sum_{m=-M}^{M} f_{n-m} g_m \right) \exp(inx). \tag{13}$$

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.[9]

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

$$\text{Laurent's rule:} \quad h_n^{(M)} = \sum_{m=-M}^{M} f_{n-m} g_m, \tag{14}$$

$$h^{(M)}(x) = \sum_{n=-M}^{M} h_n^{(M)} \exp(inx), \tag{15}$$

$$h_M(x) = \sum_{n=-M}^{M} h_n \exp(inx). \tag{16}$$

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 $M \to \infty$ and that

$$h^{(\infty)}(x) = h(x). \tag{17}$$

Although the mathematical theory on the multiplication of Fourier series is well developed,[9] 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

$$h^{(M)}(x) = h_M(x) - \sum_{x_p \in U_{fg}} \frac{\hat{f}_p \hat{g}_p}{2\pi^2} \Phi_M(x - x_p) + o(1), \tag{18}$$

where the term $o(1)$ uniformly tends to zero, and

$$\Phi_M(x) = \sum_{n=1}^{M} \frac{\cos nx}{n} \sum_{|m|>M} \frac{1}{m-n}. \tag{19}$$

Furthermore,

$$\lim_{M\to\infty} \Phi_M(x) = 0 \qquad (x \neq 0), \tag{20}$$

but

$$\lim_{M\to\infty} \Phi_M(0) = \frac{\pi^2}{4}. \tag{21}$$

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 $[\![1/f]\!]^{(M)}$ 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) Re $[1/f]$ does not change sign in $[0, 2\pi)$, Re $[1/f] \neq 0$ in S, and Im $[1/f]$ does not change sign in $\bar{S}$; (b) Im $[1/f]$ does not change sign in $[0, 2\pi)$, Im $[1/f] \neq 0$ in S, and Re $[1/f]$ does not change sign in $\bar{S}$—then Eq. (17) is valid provided that, instead of Eq. (14), the inverse rule

$$\text{Inverse Rule:} \quad h_n^{(M)} = \sum_{m=-M}^{M} \left[\!\!\left[\frac{1}{f} \right]\!\!\right]^{(M)-1}_{nm} g_m \tag{22}$$

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 $\Phi_\infty(x)$ 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 $M \to \infty$ 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 ± 1, $\pm 2, \ldots$, 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-

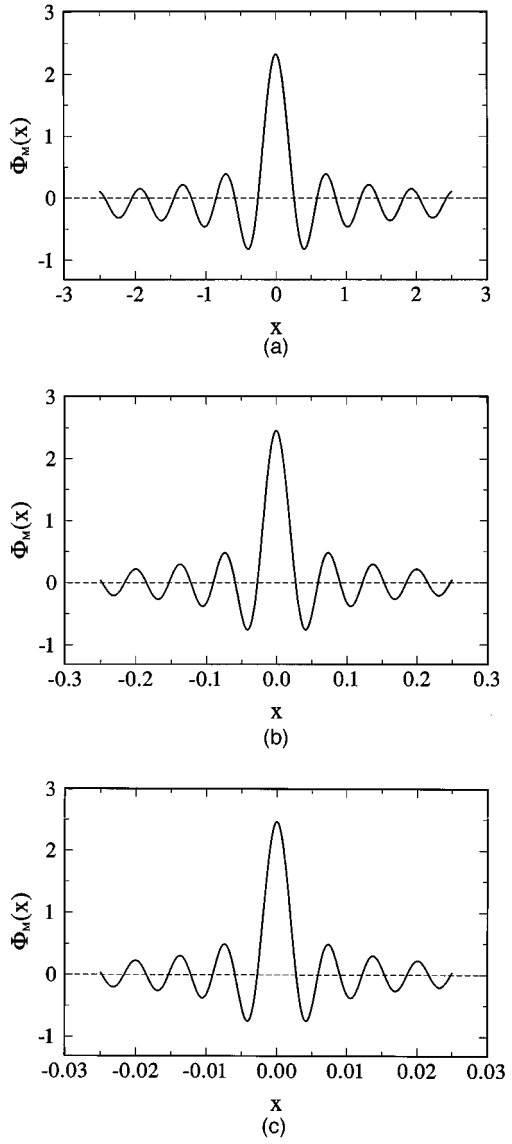

Fig. 2. Graphs of $\Phi_M(x)$ in the neighborhood of $x = 0$ for (a) $M = 10$, (b) $M = 100$, and (c) $M = 1000$. Note the change of scale for the horizontal axes.

stant but its position tends to $x = 0$ as $M \to \infty$. 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 $M \to \infty$. 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

$$\hat{f}_p\hat{g}_p = -h(x_p)\,\frac{(1-\alpha)^2}{\alpha}. \tag{23}$$

Thus the magnitude of the overshoot can be arbitrarily large as $\alpha \to 0$ or $\alpha \to \pm\infty$. As illustrations of what has just been said, let us consider two graphical examples.

In the first example, we choose

$$f(x) = \begin{cases} a & |x| < \dfrac{\pi}{2} \\ \quad , & \qquad , \\ \dfrac{a}{2} & \dfrac{\pi}{2} < |x| \leqslant \pi \end{cases} \qquad (a \neq 0), \tag{24}$$

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

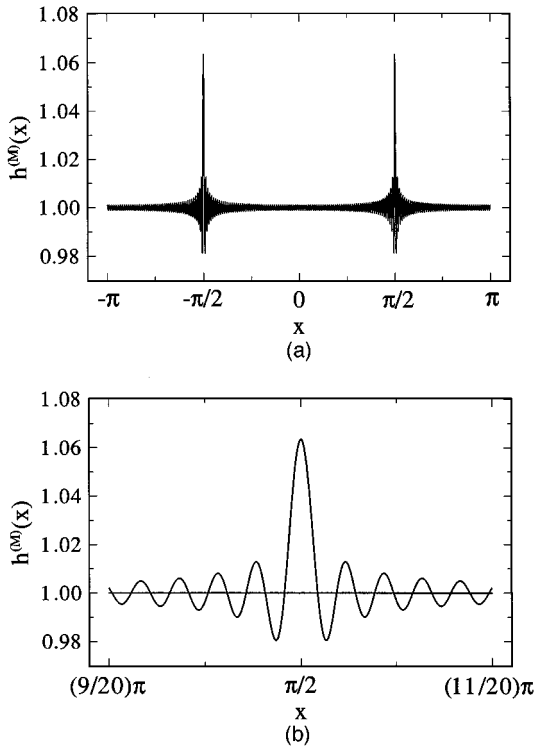

Fig. 3. (a) Graph of $h^{(M)}(x)$ that is Fourier factorized by the finite Laurent rule, with $f(x)$ given by Eq. (24), $g(x) = 1/f(x)$, and $M = 200$. (b) Enlarged view of Fig. 3(a) in the neighborhood of $x = \pi/2$. The straight horizontal line in Fig. 3(b) is obtained by the inverse rule.

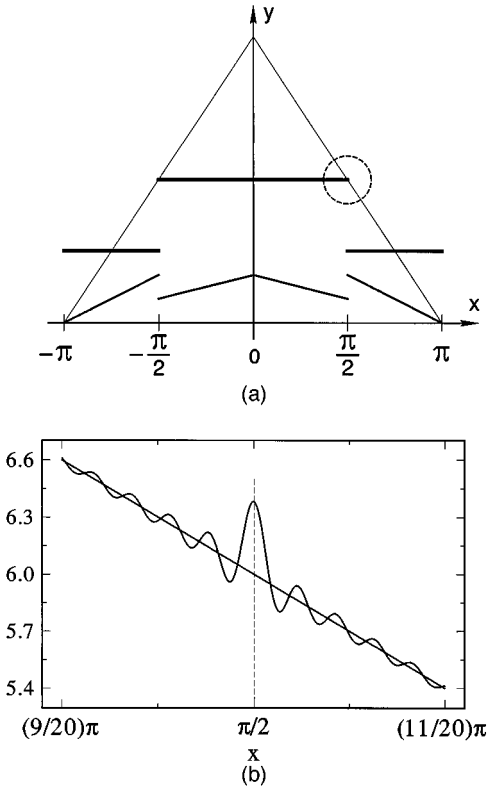

Fig. 4. (a) Schematic representations of functions $f(x)$ and $g(x)$ in Eqs. (24) and (25) and their product $h(x)$ in order of decreasing line thickness. Here $a = 6$ and $b = 2$. (b) Function $h^{(M)}(x)$, with $M = 200$, in the neighborhood of $x = \pi/2$. The oscillatory curve is obtained by Laurent's 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

$$g(x) = \begin{cases} b\left(1 - \dfrac{|x|}{\pi}\right) & |x| < \dfrac{\pi}{2} \\ , \qquad\qquad , & (b \neq 0). \\ 2b\left(1 - \dfrac{|x|}{\pi}\right) & \dfrac{\pi}{2} < |x| \leqslant \pi \end{cases} \tag{25}$$

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 ϵE_x in Eq. (26a) should be continuous in x. Since ϵ is discontinuous at $x = \pm d_1/2$, ϵ 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 ϵE_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 ϵE_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 ϵ 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:

$$-\frac{\partial^2 E_z}{\partial y^2} = \frac{\partial^2 E_z}{\partial x^2} + \mu_0 k_0{}^2 \epsilon E_z . \tag{28}$$

Here the product ϵE_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

$$k_z{}^2 E_x - \frac{\partial^2 E_x}{\partial y^2} = \frac{\partial}{\partial x}\left[\frac{1}{\epsilon}\frac{\partial}{\partial x}(\epsilon E_x)\right] + \mu_0 k_0{}^2 \epsilon E_x , \tag{29}$$

where k_z is the z component of the incident wave vector. Based on either the physics or a mathematical analysis, the products ϵE_x and $(1/\epsilon)[\partial(\epsilon E_x)/\partial x]$ must be continuous. Therefore by the inverse rule, Eq. (29) becomes

$$\frac{\partial^2 E_{xn}}{\partial y^2} = k_z{}^2 E_{xn} + \sum_m \left(\alpha_n [\![\epsilon]\!]^{-1}_{nm} \alpha_m - \mu_0 k_0{}^2 \delta_{nm}\right) \sum_p \left[\!\left[\frac{1}{\epsilon}\right]\!\right]^{-1}_{mp} E_{xp} . \tag{30}$$

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.[10] 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, ϵE_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.*[11] in the case in which the grating profiles have sharp edges. This result will be presented in a separate publication.[12]

From Eq. (11), it follows that if $\epsilon(x) \neq 0$, then

$$\sum_{l=-M}^{M} \epsilon_{m-l}\left(\frac{1}{\epsilon}\right)_{l-n} = \delta_{mn} + \Delta_{mn} , \tag{31}$$

where

$$\Delta_{mn} = \sum_{|l|>M} \epsilon_{m-l}\left(\frac{1}{\epsilon}\right)_{l-n} . \tag{32}$$

Thus, for discontinuous $\epsilon(x)$, $\Delta_{mn} \to 0$ only if m and n are such that $(M \pm n) \to \infty$ and $(M \pm m) \to \infty$ as $M \to \infty$. In other words, the matrix elements of Δ_{mn} in the vicinity of the two ends of the main diagonal remain finite as $M \to \infty$. Therefore

$$[\![\epsilon]\!]^{(M)-1} \neq \left[\!\!\left[\frac{1}{\epsilon} \right]\!\!\right]^{(M)}, \quad M \to \infty. \tag{33}$$

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[13] 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[4] and by Granet and Guizal[5] 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* (Birkhäuser Verlag, Basel, Switzerland, 1984), Chap. 1, pp. 14–33, and the references therein.