Chapter 8
Introduction to Yasuura's Method of Modal Expansion with Application to Grating Problems

Akira Matsushima, Toyonori Matsuda and Yoichi Okuno

Abstract In this chapter we introduce the theory of the Yasuura's method based on modal expansion and explain the methods of numerical computation in detail for several grating problems. After a sample problem we discuss the methods for solving two types of problems that require additional knowledge and steps, that is, scattering by a dielectric cylinder and diffraction by a grating. Some numerical results are shown to give an evidence of an experimental rule for the number of linear equations in formulating the least-squares problem that determines the modal coefficients. After confirming the rule we show a couple of examples of practical interest, i.e., scattering by a relatively deep metal grating, plasmon surface waves on a metal grating placed in conical mounting, scattering by a metal surface modulated in two directions, and scattering by periodically located dielectric spheres. To provide supplementary explanations of particular problems, four appendices are given; H-wave scattering from a cylinder, the normal equation and related topics, conical diffraction by a dielectric grating, and comparison of modal functions and the algorithm of the smoothing procedures.

A. Matsushima (✉)
Kumamoto University, Kumamoto 860-8555, Japan
e-mail: matsua@cs.kumamoto-u.ac.jp

T. Matsuda
National Institute of Technology, Kumamoto College,
Kumamoto 861-1102, Japan
e-mail: tmatsu@kumamoto-nct.ac.jp

Y. Okuno
South China Normal University, Guangzhou 510006, China
e-mail: okuno@gpo.kumamoto-u.ac.jp

T. Wriedt and Y. Eremin (eds.), *The Generalized Multipole Technique for Light Scattering*, Springer Series on Atomic, Optical, and Plasma Physics 99,
https://doi.org/10.1007/978-3-319-74890-0_8

8.1 Introduction

In Chap. 6 of the last edition [27] we have introduced Yasuura's method of modal expansion from two main points of view: one was the relation with the fictitious or equivalent source methods; and another was the employment of smoothing procedures (SP's) [10, 24, 25, 31, 41, 42] to obtain rapidly converging solutions. We needed the first point to have the method recognized as one of the modal expansion methods having firm theoretical foundations and a wide range of application. While in the second point we tried to explain our tool to cope with the problem of slow convergence. Because we had been working with the separated solutions as modal functions, we were often troubled by their poor approximation power. Accordingly, Yasuura et al. hit upon an idea of the SP, which works to accelerate the convergence of solutions by reducing the higher-order oscillations on the boundary. The SP, hence, is an important step in solving a 2-D problem[1] where the cross section of the obstacle is strongly deformed from the coordinate curves of a separable system of coordinates.

In the last edition we have included: (**1**) the theory and the method of numerical execution of the original form of Yasuua's method, which we call the conventional Yasuura's method (CYM) today; (**2**) Yauura's method with a smoothing or a singular-smoothing procedure (YMSP or YMSSP); and (**3**) numerical examples obtained mainly by the YMSP and YMSSP. In the present chapter, however, we decided to omit a greater part of SP-related topics in view of the recent trend in computational electromagnetics. That is, the methods for 3-D as well as 2-D analysis of structures made of a dielectric are required in various areas. Instead of removing the SP, we include a detailed explanation on the solution process by the CYM. We hope this helps those who are interested in solving their problems by using Yasuura's methods, CYM, YMSP, and YMSSP. Because the process with the SP's are almost in common with that of the CYM, the detailed introduction of the CYM execution process would be useful not only for the CYM users but also for those who intend to employ the YMSP or YMSSP.

The contents of this chapter are as follows: In Sect. 8.2 we first introduce the theory of the CYM briefly and explain the method of numerical computation in detail taking a sample problem. Then, we move on to the methods for solving two types of problems that require additional knowledge and steps: (**1**) scattering by a dielectric cylinder; and (**2**) diffraction by a grating. In Sect. 8.3 we show some numerical results. The aim of Sect. 8.3.1 is to give an evidence of an experimental rule for the number of sampling points or, in general, the number of linear equations in formulating the least-squares problem that determines the modal coefficients. Computational results show the number should be twice as many as the number of unknown modal coefficients. After confirming the rule we show a couple of examples of practical interest in Sects. 8.3.2–8.3.5: Scattering by a relatively deep metal grating, Plasmon surface waves on a metal grating placed in conical mounting, Scattering by a metal surface modulated in two directions, and Scattering by periodically located dielectric

[1] The reason why we set a limit "2-D" is that the SP, in the present form, is available only in 2-D problems. This is because we employ an indefinite integral to realize a low-pass spatial filter.

spheres. Section 8.4 is a conclusion where we state some additional remarks. Finally, four appendices follow mainly providing supplementary explanations of particular problems: **1** H-wave scattering from a cylinder; **2** The normal equation and related topics; **3** Conical diffraction by a dielectric grating; and **4** Comparison between two types of modal functions and a brief introduction to the algorithm with the SP.

8.2 Yasuura's Method of Modal Expansion

In this section we introduce the foundations of Yasuura's method of modal expansion. We start by formulating a sample problem: plane wave scattering by a perfectly-conducting (PC) cylinder, the problem from which we can learn the essential part of the method together with important concepts and ideas in Yasuura's method.

8.2.1 Scattering by a Perfectly-Conducting Cylinder

The geometry of the sample problem is shown in Fig. 8.1. The closed curve C is the cross section and S_e is the exterior infinite region of C. We denote a point in S_e by $\mathbf{r}\,(r, \theta)$; and one on C by an arc-length s along C measured counterclockwise from a fixed point s_0. S_{e0} is an arbitrary closed region that is entirely inside S_e. Let the incident plane wave be polarized in z and

$$\mathbf{E}^{\mathrm{i}}(\mathbf{r}) = \mathbf{u}_z F(\mathbf{r}) = \mathbf{u}_z \exp[-\mathrm{i}kr\cos(\theta - \iota)], \tag{8.1}$$

where $\mathbf{u}_z$ is a unit vector in z-direction, ι is the angle of incidence shown in Fig. 8.1, and $k = 2\pi/\lambda = \omega/c$ is the wavenumber of the incident field. The $e^{\mathrm{i}\omega t}$ time dependence is assumed. This case of polarization is called E-wave,[2] which is one of the two basic polarizations. We will deal with an E-wave problem in this section and summarize important results of an H-wave case, which is another basic polarization, in Appendix 1.

In the present problem a surface current flows in the z-direction exciting a scattered wave polarized in z again:

$$\mathbf{E}^{\mathrm{s}}(\mathbf{r}) = \mathbf{u}_z \Psi(\mathbf{r}). \tag{8.2}$$

Other non-zero components of the scattered wave, $\mathbf{H}^{\mathrm{s}}(\mathbf{r}) = \mathbf{u}_x H_x^{\mathrm{s}}(\mathbf{r}) + \mathbf{u}_y H_y^{\mathrm{s}}(\mathbf{r})$, can be obtained by[3]

[2] It is also termed Transverse-Electric (TE) wave, which means the electric field is orthogonal to the xy-plane. While in the H-wave (or TM-wave) the magnetic field has the z-component alone.

[3] The component is called a leading field if it gives other nonzero components as in (8.3). Note that the derivation of $\mathbf{H}^{\mathrm{s}}$ by (8.3) is a proper procedure because the sequence of our approximate

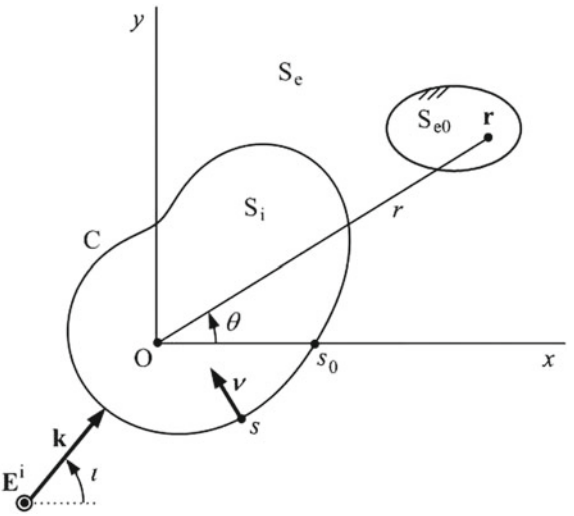

Fig. 8.1 Geometry of the sample problem. C is a closed curve of length C representing the cross section of a cylindrical obstacle. S_e is the exterior infinite region of C; and S_i is the interior region, which we need in Sects. 8.2.2.1 and 8.2.4

$$\mathbf{H}^s(\mathbf{r}) = \frac{\mathrm{i}}{\omega\mu_0}\nabla\Psi(\mathbf{r}) \times \mathbf{u}_z, \tag{8.3}$$

where $\nabla = (\partial/\partial x, \partial/\partial y, 0)$ is the 2-D nabla operator. Hence, our target is $\Psi(\mathbf{r})$ and we can state our sample problem as:

Problem 1 E-wave, PC. Find the scattered electric field $\Psi(\mathbf{r})$ that satisfies:

(**D1**) The 2-D Helmholtz equation in S_e

$$\nabla^2\Psi(\mathbf{r}) + k^2\Psi(\mathbf{r}) = 0 \quad (\mathbf{r} \in S_e), \tag{8.4}$$

(**D2**) The 2-D radiation condition at infinity

$$\sqrt{r}\left(\frac{\partial\Psi(\mathbf{r})}{\partial r} + ik\Psi(\mathbf{r})\right) \to 0 \quad (r \to \infty), \tag{8.5}$$

(**D3**) The boundary condition

$$\Psi(s) = f(s) \equiv -F(s) \quad (s \in \mathrm{C}, \text{ i.e., } 0 \le s \le C). \tag{8.6}$$

Here, $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ denotes the 2-D Laplacian. The condition given by (8.6) is called Dirichlet's or the first-kind boundary condition.

---

solutions converges to the true solution uniformly in wider sense in the exterior region S_e as we will see later.

8.2.2 *Modal Functions, Approximate Solution, and Least-Squares Boundary Matching*

Here we introduce the analytical part of Yasuura's method [38–40]. Because it is one of the modal expansion methods, we need: (**i**) definition of a set of modal functions; (**ii**) a method to construct an approximate solution; and (**iii**) the sense in which the solution approximates the boundary condition. Let us see these points below.

8.2.2.1 Definition of the Set of Modal Functions

Modal functions for the sample problem are solutions of Helmholtz's equation (8.4) satisfying some additional requirements. Here, we define a set of modal functions $\{\varphi_m(\mathbf{r}) : m = 1, 2, \ldots\}$ as a countable set that satisfy the following three requirements:

(**M1**) Each $\varphi_m(\mathbf{r})$ satisfies the Helmholtz equation in S_e;
(**M2**) Each $\varphi_m(\mathbf{r})$ meets the 2-D radiation condition;
(**M3**) Both the set of boundary values $\{\varphi_m(s) : m = 1, 2, \ldots\}$ and the set of normal derivatives $\{\partial\varphi_m(s)/\partial\nu : m = 1, 2, \ldots\}$ are complete (or total)[4] in the function space $\mathbf{H} = L^2(\mathrm{C})$ consisting of all the square-integrable functions defined on the boundary C.

The first two requirements are natural and easy to understand; but the third is rather complicated and needs explanation. Here, we would like to call readers' attention to the fact that (**M3**) is a little different from the original requirement given in [39], which seems to be lacking in concreteness than the statement above. We have modified the original statement to require completeness of the boundary values.

Now, let us see a couple of examples first to facilitate the understanding. Then, we will give additional explanations for this issue throughout this section.

Example 1 The set of radiative separated solutions

$$\varphi_m(\mathbf{r}) = H_m^{(2)}(kr)\exp(\mathrm{i}m\theta) \quad (m = 0, \pm 1, \pm 2, \ldots), \tag{8.7}$$

where $H_m^{(2)}(kr)$ is the second kind Hankel function of order m and the coordinate origin should be inside S_i, the complimentary region of S_e.[5]

[4] We hope the readers consult a treatise on *Functional Analysis*, e.g. [14], in case of need.

[5] If C is a circle centered at the origin, it is apparent that the sets of boundary values and normal derivatives are both complete because the members of each set are nothing other than the Fourier bases. Even in case if C is not a circle, the sets are still complete because of Example 2: let L be a circle centered at the origin and take the Fourier bases for $f_m(t)$ in (8.8), then we get a set of separated solutions.

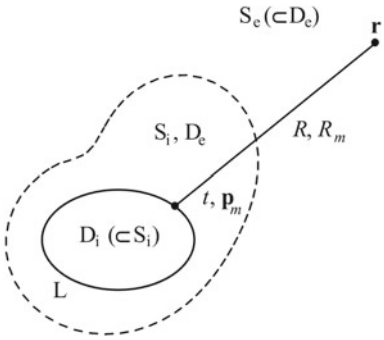

Fig. 8.2 Relevant to definition of the set of modal functions. The cross section C is shown by a dashed curve; L is another closed curve inside C

Example 2 Let L be a smooth closed curve that is entirely inside S_i and D_e be an exterior infinite region of L. As shown in Fig. 8.2, S_e is a subregion of D_e; and D_i, the complementary region of D_e, is a subregion of S_i. Now, let an enumerable set of functions $\{f_m(s) : m = 1, 2, \ldots\}$ be complete in the function space $L^2(\mathrm{L})$. Then, the set of potential functions defined in D_e with $f_m(t)$'s as double-layer density functions on L

$$\varphi_m(\mathbf{r}) = -\int_{\mathrm{L}} f_m(t)\,\frac{\partial \psi(kR)}{\partial \nu_t}\,dt \quad (\mathbf{r} \in \mathrm{D_e};\ R = \overline{t\mathbf{r}};\ m = 1, 2, \ldots) \tag{8.8}$$

is a set of modal functions in D_e provided that k does not coincide with a member of $\{k_{\mathrm{H}}(\mathrm{D_i})\}$, the set of eigenvalues of the homogeneous H-wave (Neumann) problem in D_i.[6] Here, R is the distance between t and $\mathbf{r}$, $\psi(kR) = H_0^{(2)}(kR)/4\mathrm{i}$ is the free-space Green's function, and $\partial/\partial\nu_t$ denotes normal derivative at t. Note that the ensemble of single-layer potentials can also be the set of modal functions provided $k \notin \{k_{\mathrm{E}}(\mathrm{D_i})\}$, the set of eigenvalues of homogeneous E-wave (Dirichlet) problem in D_i.

Example 3 Monopole fields whose poles $\mathbf{p}_m$ are located on L

$$\varphi_m(\mathbf{r}) = H_0^{(2)}(kR_m) \quad (\mathbf{r} \in \mathrm{D_e};\ R_m = \overline{\mathbf{p}_m\mathbf{r}};\ m = 1, 2, \ldots, M) \tag{8.9}$$

form a set of modal functions in D_e when we let $M \to \infty$ while letting $\overline{\mathbf{p}_m\mathbf{p}_{m+1}} \to 0$ provided there is no internal resonance in D_i [26, 30].

Example 4 The set of multiple-multipole fields whose poles $\mathbf{p}_m$ are on L

$$\begin{aligned}\varphi_{mn}(\mathbf{r}) &= H_n^{(2)}(kR_m)\exp(\mathrm{i}n\theta_m)\\ &\quad (\mathbf{r} \in \mathrm{D_e};\ R_m = \overline{\mathbf{p}_m\mathbf{r}};\ m = 1, 2, \ldots, M;\ n = 0, \pm 1, \pm 2, \ldots)\end{aligned} \tag{8.10}$$

is also an example of modal functions.

[6] This is not a strong exception because we can modify the contour L (and hence D_i) slightly to avoid the coincidence. Example 2 is a key theorem of generation of complete sets, which has been proven by Yasuura and Itakura [39] as an analogy of Runge's (or Runge-Walsh's) theorem known in *Theory of Complex Functions*.

8.2.2.2 Construction of an Approximate Solution

To define an approximate solution, we first choose a set of modal functions from among possible candidates. Let us take the set of separated solutions in the following analysis. This is because the set of separated solutions is one of the most familiar functions and each member has physical meaning.[7] Thus we can define an approximate solution as a *finite summation* of the outgoing separated solutions (8.7) with unknown coefficients:

$$\Psi_N(\mathbf{r}) = \sum_{m=-N}^{N} A_m(M)\,\varphi_m(\mathbf{r}). \tag{8.11}$$

Here, $A_m(M)$ means that the A_m coefficient depends on $M = 2N + 1$, the number of modal functions employed.[8] Because of the definition of modal functions, the approximate solution already satisfies the requirements (**D1**) and (**D2**). The A_m coefficients, hence, should be determined so that the solution meets the boundary condition in a sense of approximation. Let us call this procedure *boundary matching* and keep in mind that the sense of approximation in boundary matching determines a method of solution.

We employ the least-squares approximation in Yasuura's method, i.e., minimization of mean-squares boundary residual. We will see in Sect. 8.2.2.3 that this is a promising way in boundary matching provided the completeness of the set of boundary values (**M3**) is guaranteed.

8.2.2.3 Least-Squares Boundary Matching

We employ integral representations of the solutions to explain the method of solution including convergence of the approximate solutions. For this purpose let us define the Green's function of our problem, $G(\mathbf{r}, \mathbf{r}')$, satisfying Helmholtz's equation with a unit source at $\mathbf{r}$, radiation condition with respect to $\mathbf{r}'$, and a homogeneous boundary condition[9]

---

[7] Unfortunately, separated solutions are not very efficient in a problem where C is strongly modulated from a circle (or, in general, a coordinate surface of the system of coordinates employed). As an example, we show a comparison between types of modal functions: the separated solutions (8.7) and monopole fields (8.9) in Appendix 4.

[8] This dependence is natural because the boundary values of modal functions, in general, do not form an orthogonal set in **H**. This type of summation is usually called a flexible summation. Note that the approximate solution is defined in a finite summation of modal functions. By considering a sequence of finite-sum solutions, we can avoid the constraint of the convergence area of an infinite series solution. Yasuura's original papers [38–40] has been written from this point of view. Reference [9] includes an interpretation of the difference between series and sequence solutions.

[9] $G(\mathbf{r}, \mathbf{r}')$ is a total electric field observed at $\mathbf{r}'$ ($\neq \mathbf{r}$) when a unit line source is placed at $\mathbf{r}$ in Fig. 8.1. Note that employment of the Green function satisfying (8.12) is for convenience and is not essential: The whole theory has been established in [38–40], where the free-space Green function alone was used.

$$G(\mathbf{r}, s) = 0 \quad (\mathbf{r} \in \mathrm{S_e};\ s \in \mathrm{C}). \tag{8.12}$$

Using the Green's formula to $\Psi(\mathbf{r}')$ and $G(\mathbf{r}, \mathbf{r}')$, we have

$$\Psi(\mathbf{r}) = -\int_{s=0}^{C} \partial_\nu G(\mathbf{r}, s)\Psi(s)\, ds = -\int_{s=0}^{C} \partial_\nu G(\mathbf{r}, s) f(s)\, ds \quad (\mathbf{r} \in \mathrm{S_e}). \tag{8.13}$$

Here, ∂_ν denotes the normal derivative at s and the second equality comes from (**D3**). Besides, we get a similar representation for the approximate solution $\Psi_N(\mathbf{r})$. Subtracting (8.13) from the representation of $\Psi_N(\mathbf{r})$ side by side, we have

$$\Psi_N(\mathbf{r}) - \Psi(\mathbf{r}) = -\int_{s=0}^{C} \partial_\nu G(\mathbf{r}, s)[\Psi_N(s) - f(s)]\, ds \quad (\mathbf{r} \in \mathrm{S_e}). \tag{8.14}$$

Although (8.14) is a formal representation, we can deduce useful results starting from it.

Let the observation point $\mathbf{r}$ be inside the closed region $\mathrm{S_{e0}}$ in Fig. 8.1. Then, $\partial_\nu G(\mathbf{r}, s)$ is a continuous function of s because there is a non-zero distance between s and $\mathbf{r}$. Taking the absolute value of both sides of (8.14) and applying Cauchy–Schwarz's inequality to the right-hand side, we obtain

$$\left|\Psi_N(\mathbf{r}) - \Psi(\mathbf{r})\right| \le \sqrt{\int_{s=0}^{C} |\partial_\nu G(\mathbf{r}, s)|^2\, ds}\, \left\|\Psi_N - f\right\| \quad (\mathbf{r} \in \mathrm{S_e}). \tag{8.15}$$

Here, $\|f\|$ stands for the Euclidean norm of a function $f(s)$ defined by

$$\|f\| = \left[\int_{s=0}^{C} |f(s)|^2\, ds\right]^{1/2}. \tag{8.16}$$

Because the integrand on the right of (8.15) is a continuous function of $\mathbf{r}$, the integral, as a function of $\mathbf{r}$, has a maximum inside the closed region $\mathrm{S_{e0}}$:

$$G(\mathrm{S_{e0}}) = \max_{\mathbf{r} \in \mathrm{S_{e0}}} \sqrt{\int_{s=0}^{C} |\partial_\nu G(\mathbf{r}, s)|^2\, ds} \quad (\mathrm{S_{e0}} \subset \mathrm{S}). \tag{8.17}$$

Thus we have an estimation

$$|\Psi_N(\mathbf{r}) - \Psi(\mathbf{r})| \le G(\mathrm{S_{e0}})\left\|\Psi_N - f\right\| \quad (\mathbf{r} \in \mathrm{S_{e0}} \subset \mathrm{S_e}), \tag{8.18}$$

which means that the maximal absolute error in $\mathrm{S_{e0}}$ cannot exceed the product of the mean-squares boundary residual and a factor of proportionality $G(\mathrm{S_{e0}})$. Note that the latter depends on the region $\mathrm{S_{e0}}$ but does not depend on $\mathbf{r}$.

Now, let us remember the completeness (**M3**) of the set of boundary values of modal functions. Because the given boundary value $f(s)$ is a member of $\mathbf{H} = L^2(\mathrm{C})$, for given any positive number ε, there is a positive integer N_0 such that

$$\left\| \Psi_N - f \right\| < \varepsilon \quad (N > N_0). \tag{8.19}$$

That is, there exists a sequence of boundary values of the approximate solutions $\{\Psi_0(s), \Psi_1(s), \Psi_2(s), \ldots\}$ that converges to the true boundary value $f(s)$ in the mean-squares sense:

$$\left\| \Psi_N - f \right\| \to 0 \quad (N \to \infty). \tag{8.20}$$

Referring to (8.18), we can conclude that the corresponding sequence of approximate solutions $\{\Psi_0(\mathbf{r}), \Psi_1(\mathbf{r}), \Psi_2(\mathbf{r}), \ldots\}$ converges to $\Psi(\mathbf{r})$ uniformly in the closed region $\mathrm{S_{e0}}$[10]: for given any positive number ε, there is a positive integer $N_0(\mathrm{S_{e0}}, \varepsilon)$ such that

$$|\Psi_N(\mathbf{r}) - \Psi(\mathbf{r})| < \varepsilon \quad (\mathbf{r} \in \mathrm{S_{e0}} \subset \mathrm{S_e};\ N > N_0(\mathrm{S_{e0}}, \varepsilon)). \tag{8.21}$$

We can get such a sequence by solving repeatedly the following least-squares problem (LSP) stated in the function space **H**.

LSP 1: E-wave, PC. Find the coefficients $A_m(M)$ $(m = 0, \pm 1, \ldots, \pm N;\ M = 2N + 1)$ that minimize the normalized mean-squares boundary residual

$$E_N = \frac{\left\| \Psi_N - f \right\|^2}{\|f\|^2} = \frac{1}{\|f\|^2} \left\| \sum_{m=-N}^{N} A_m(M)\varphi_m - f \right\|^2. \tag{8.22}$$

Note that the least-squares boundary matching means a relaxation of the boundary condition because (8.6) implies $\|\Psi - f\| = 0$; but the converse is not always true. The smoothing procedure (SP), which we mentioned in Introduction, is an extension of the relaxation idea: we minimize $\| \int (\Psi - f)\, ds \|$ instead of $\|\Psi - f\|$; and extinction of the latter is stronger than vanishing of the former [10, 24, 31, 41, 42]. Although the Yasuura's method with the SP is a strong tool for 2-D problems, we shall not get deeply in this subject.

8.2.3 Method of Numerical Solution

Because computers cannot handle continuous functions, we need (**i**) method of discretization of **LSP 1** and (**ii**) method of solution to the discretized problem.[11]

[10] This kind of convergence is called uniform convergence in wider sense in $\mathrm{S_e}$.

[11] Until the middle of 80s we employed normal equations (NE) in solving **LSP 1**. Now we solve the problem using the method in Sect. 8.2.3.2. We state the reason why we stopped using the NE and attach some comments in Appendix 2.

8.2.3.1 Method of Discretization

To discretize the problem we first locate J ($\geq M$) sampling points on C, assign a set of integers from 0 through J to them, and get a numbered set of sampling points $\{s_0, s_1, \ldots, s_J\}$. Because C is a closed curve, we give two numbers, 0 and J, to the point $s = 0$. Two methods are usually used in locating the points:

1. An equal division of the boundary C: In most applications we can recommend to use the points given by

$$s_j = \frac{jC}{J} \quad (j = 0, 1, 2, \ldots, J) \tag{8.23}$$

   without any reservation in theory. If we take this method, we may have to solve an (transcendental) (8.23) in locating the points.
2. An equal division with respect to a coordinate variable: For example, if C is represented as $r = r(\theta)$, it must be convenient to use the discretization

$$\theta_j = \frac{j2\pi}{J} \quad (j = 0, 1, 2, \ldots, J). \tag{8.24}$$

   This choice, however, means a variable transformation in (8.13) and in other integrals on C and will lead us solving a weighted least-squares problem unexpectedly. Users should notice this and be careful in applying this method of location in a problem where the boundary C is strongly deformed from a circle.[12]

Having located the sampling points on C, we can define discretized forms of the functions $f(s)$, $\varphi_m(s)$, and so on:

$$\mathbf{f} = [f(s_1)\ f(s_2)\ \cdots\ f(s_J)]^{\mathrm{T}} \tag{8.25}$$

and

$$\boldsymbol{\varphi}_m = [\varphi_m(s_1)\ \varphi_m(s_2)\ \cdots\ \varphi_m(s_J)]^{\mathrm{T}}. \tag{8.26}$$

Here, the superscript T denotes a transposed vector or matrix and the discretized forms are J-dimensional complex-valued column vectors. Next, we define a $J \times M$ matrix by

$$\Phi = \left[\boldsymbol{\varphi}_{-N}\ \boldsymbol{\varphi}_{-N+1}\ \cdots\ \boldsymbol{\varphi}_N\right] = \begin{bmatrix} \varphi_{-N}(s_1) & \varphi_{-N+1}(s_1) & \cdots & \varphi_N(s_1) \\ \varphi_{-N}(s_2) & \varphi_{-N+1}(s_2) & \cdots & \varphi_N(s_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{-N}(s_J) & \varphi_{-N+1}(s_J) & \cdots & \varphi_N(s_J) \end{bmatrix} \tag{8.27}$$

[12] On the other hand, there is a possibility to make possible use of the weighting function accompanying the variable transformation. For example, a Schwarz–Christoffel-type transformation works to remove the singularity of Green's function in a problem of an edged cross section [23].

which is usually termed a Jacobian matrix. Finally, defining an M-dimensional solution vector

$$\mathbf{A} = \left[A_{-N}(M)\ A_{-N+1}(M)\ \cdots\ A_N(M)\right]^{\mathrm{T}}, \tag{8.28}$$

we can represent a discretized form of an approximate solution on C in vector-matrix notation

$$\boldsymbol{\Psi}_N = \sum_{m=-N}^{N} A_m(M)\boldsymbol{\varphi}_m = \Phi\mathbf{A}. \tag{8.29}$$

Thus, we have an approximation to the mean-squares boundary residual in (8.22):

$$E_{NJ} = \frac{\|\Phi\mathbf{A} - \mathbf{f}\|^2}{\|\mathbf{f}\|^2}. \tag{8.30}$$

Here, $\|\mathbf{f}\|$ denotes a Euclidean norm of a J-dimensional complex-valued vector $\mathbf{f}$. Because C is a closed curve and $f(s_0) = f(s_J)$, etc., (8.30) can be understood as a trapezoidal-rule approximation of (8.22). Now, we can state a discretized form of **LSP 1** as follows:

DLSP 1: E-wave, PC. Find the solution vector $\mathbf{A}$ that minimizes the numerator of (8.30).

Here arises an important issue of the number of sampling points[13]: How many J do we need? If we answer to this question in generality, we should say: It depends. However, employing the results of examination in Sect. 8.3.1, we can state an experimental rule:

$$J \doteq 2(2N+1) = 2M. \tag{8.31}$$

Here, the symbol $\doteq$ means that the number on the right-hand side is usually *sufficient in finding the scattered field*. This might be considerably smaller than what uninitiates expect because **DLSP 1** with the number J of (8.31) does not seem to be a good approximation of **LSP 1**. This is because an inner product $(C/J)\mathbf{f}^{\dagger}\mathbf{g}$ implicitly included in the norm on the right of (8.30) cannot be a precise approximation of (f, g) in (8.22) if either $f(\mathbf{r})$ or $g(\mathbf{r})$ is a higher-order space harmonic. Nevertheless, **DLSP 1** with (8.31) gives an approximate solution having converged with respect to J. We know this welcome nature of **DLSP 1** since we started solving the scattering problem on a computer in early 70s. At that time the method in Sect. 8.2.3.2 was not known widely and we solved the problem using a normal equation (NE; see Appendix 2). We found (8.31) was effective even in using the NE where the inner products $(C/J)\mathbf{f}^{\dagger}\mathbf{g}$ appeared explicitly as the matrix elements. In the method of solution that we introduce next, we do not have to calculate these inner products. That is one of the advantages of the method.

[13]It is more reasonable to ask "How many linear equations do we need?" This is because (i) we get two equations at one sampling point in a 2-media problem (see Sect. 8.2.4); and (ii) we should understand (8.31) as a relation between the numbers of equations J and unknowns M.

8.2.3.2 Solution Method to the Discretized Problem

To solve the least-squares problem in the J-dimensional vector space, we employ orthogonal decomposition of the Jacobian matrix: the singular-value decomposition (SVD) and the QR decomposition (QRD) [15]. They have the following features:

- The SVD informs us of the character of the Jacobian matrix through singular values. This is helpful in designing and testing a process of numerical solution, in particular, choice of modal functions, number and location of sampling points, etc. Instead, the computational complexity, in both memory and time, is bigger than that of the QRD.
- The QRD needs less computation than the SVD and solves the problem provided no rank deficiency occurs.[14]

Hence, we recommend the use of the SVD for designing and testing the discretized least-squares problem. After the problem is established, application of the QRD is appropriate. Let us see how to use these decompositions in examining and solving **DLSP 1**.

Utilization of the SVD

Applying the SVD, we get a decomposition of the Jacobian matrix in the form

$$\Phi = \mathrm{U}\Sigma\mathrm{V}^{\dagger}, \tag{8.32}$$

where U ($J \times J$) and V ($M \times M$) are unitary matrices, and † denotes Hermitian conjugation: $\mathrm{V}^{\dagger} = \bar{\mathrm{V}}^{\mathrm{T}}$. Σ is a stack of an $M \times M$ diagonal matrix and a $(J - M) \times M$ zero matrix. The diagonal elements of Σ, σ_m, are non-negative and are called the singular values of Φ. Arranging the M singular values in the order of decreasing magnitude, we have $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M$ ($M = 2N + 1$). Let us call σ_1 and σ_M by $\sigma_{\max}$ and $\sigma_{\min}$ because this order of σ_m does not necessarily agree with the order of modal functions. The following items are widely known and accepted:

- The singular values are non-negative square roots of the eigenvalues of a positive semidefinite Hermitian matrix $\Phi^{\dagger}\Phi$: $\sigma_m(\Phi) = \sqrt{\lambda_m(\Phi^{\dagger}\Phi)}$. And, vanish of the smallest singular value, $\sigma_{\min} = 0$, means $\det \Phi^{\dagger}\Phi = 0$. Because $\Phi^{\dagger}\Phi$ is the coefficient matrix of the NE (8.106) in Appendix 2, this is a serious problem: the least-squares problem does not have a unique solution. Although $\sigma_{\min} = 0$ in strict sense seldom occurs in practice, very tiny $\sigma_{\min}$ is not rare and causes substantial rank deficiency.
- The ratio of the maximum singular value to the minimum

$$\mathrm{cond}(\Phi) = \frac{\sigma_{\max}}{\sigma_{\min}} \tag{8.33}$$

[14] In addition, the solution by a QRD program, usually, is not inferior in accuracy to one by an SVD program. This may be because of the greater computational complexity of the SVD.

defines the condition number of Φ, which shows the degree of numerical difficulty in solving the least-squares problem with the Jacobian matrix Φ. In general, a problem with a small cond(Φ) is easy to solve and is termed well-conditioned; while one with a huge cond(Φ) is difficult and called ill-conditioned. In this connection an empirical rule is known: if the reciprocal of cond(Φ) is of the same order as or smaller than the machine epsilon[15] of the system of floating-point numbers, effective rank of Φ might be less than M and **DLSP 1** may not be solved properly.

Although our main purpose to employ the SVD is to check the nature of Φ, we can solve **DLSP 1** in the following way:

(a) Modifying $\|\Phi\mathbf{A}-\mathbf{f}\|^2$ by insertion of (8.32), we have

$$\|\Phi\mathbf{A}-\mathbf{f}\|^2=\|\mathrm{U}^\dagger(\Phi\mathbf{A}-\mathbf{f})\|^2=\|\Sigma\mathrm{V}^\dagger\mathbf{A}-\mathrm{U}^\dagger\mathbf{f}\|^2=\|\Sigma\mathbf{B}-\mathbf{d}\|^2. \quad (8.34)$$

Here, we have used that the matrices U and V are unitary and that a unitary transformation does not change the norm of a vector. Also, note that the last equal sign defines the vectors **B** and **d**.

(b) We get the solution to **DLSP 1** from

$$B_m=\frac{d_m}{\sigma_m}\quad (m=1,2,\ldots,M\ (=2N+1)) \quad (8.35)$$

and the squared norm by

$$\|\Phi\mathbf{A}-\mathbf{f}\|^2=\sum_{j=M+1}^{J}|d_j|^2. \quad (8.36)$$

Utilization of the QRD

Employment of the QRD leads us to a decomposition of the form

$$\Phi=\mathrm{Q}\tilde{\mathrm{R}}=\mathrm{Q}\begin{bmatrix}\mathrm{R}\\0\end{bmatrix}, \quad (8.37)$$

where Q is a $J\times J$ unitary matrix and $\tilde{\mathrm{R}}$ is a stack of $M\times M$ upper triangular matrix R and a $(J-M)\times M$ zero matrix. Having the decomposition (8.37), we can solve **DLSP 1** by the following procedure:

(a) Inserting (8.37) into $\|\Phi\mathbf{A}-\mathbf{f}\|^2$, we have

[15]The machine epsilon, EPS, is the minimum positive number that satisfies $1+\mathrm{EPS}>1$ in the floating-point system employed.

$$\|\Phi\mathbf{A}-\mathbf{f}\|^2=\|\mathrm{Q}^\dagger(\Phi\mathbf{A}-\mathbf{f})\|^2=\|\tilde{\mathrm{R}}\mathbf{A}-\mathrm{Q}^\dagger\mathbf{f}\|^2\equiv\left\|\begin{bmatrix}\mathrm{R}\mathbf{A}\\0\end{bmatrix}-\begin{bmatrix}\mathbf{d}\\\mathbf{z}\end{bmatrix}\right\|^2, \tag{8.38}$$

where the last equality defines the vectors $\mathbf{d}$ and $\mathbf{z}$.

(b) We obtain the solution by solving

$$\mathrm{R}\mathbf{A}=\mathbf{d}. \tag{8.39}$$

Because R is triangular, we need only back substitution to solve (8.39). The residual norm is given by

$$\|\Phi\mathbf{A}-\mathbf{f}\|^2=\|\mathbf{z}\|^2. \tag{8.40}$$

8.2.4 Application to Dielectric or Metal Obstacles

This section introduces the Yasuura's method applied to problems with dielectric or metal obstacles [35, 43–45]. Although metals have unique nature, we here regard a metal as a dielectric with a complex permittivity depending on the frequency. Therefore we consider a material whose permittivity and refractive index are given by complex numbers ε and $n=\sqrt{\varepsilon/\varepsilon_0}$. Usually the material is penetrable and there is a non-zero transmitted field in $\mathrm{S_i}$, the complementary region of $\mathrm{S_e}$. Thus we have two unknown functions $\Psi_\mathrm{i}(\mathbf{r})$ $(\mathbf{r}\in\mathrm{S_i})$ and $\Psi_\mathrm{e}(\mathbf{r})$ $(\mathbf{r}\in\mathrm{S_e})$; and we need two boundary conditions to determine the two unknown functions. The continuity of tangential components of the electric and magnetic field satisfies the necessity.

8.2.4.1 E-Wave Scattering by a Cylindrical Obstacle Made of a Dielectric

Let us assume that the obstacle in Fig. 8.1 is made of a dielectric and that an E-wave is incident. The electric field in $\mathrm{S_e}$ is a sum of the incident and the scattered wave: $\mathbf{u}_z(F+\Psi_\mathrm{e})(\mathbf{r})$; while the field in the interior region $\mathrm{S_i}$ is the transmitted field $\mathbf{u}_z\Psi_\mathrm{i}(\mathbf{r})$. They are the solutions of Helmholtz's equation in each region:

$$\begin{cases}\left(\nabla^2+k^2\right)\Psi_\mathrm{e}(\mathbf{r})=0 & (\mathbf{r}\in\mathrm{S_e}),\\ \left(\nabla^2+(nk)^2\right)\Psi_\mathrm{i}(\mathbf{r})=0 & (\mathbf{r}\in\mathrm{S_i}).\end{cases} \tag{8.41}$$

The exterior solution, in addition, should meet the radiation condition (8.5). The continuity of tangential components of electric and magnetic fields requires the boundary conditions[16]

[16]To get the latter we set $\mathbf{u}_\nu\times(\mathbf{H}_\mathrm{e}-\mathbf{H}_\mathrm{i})=0$. Insertion of $\mathbf{H}_\mathrm{e}=(\mathrm{i}/\omega\mu)\nabla E_\mathrm{e}\times\mathbf{u}_z$ etc. finds the desired relation. Here, $E_\mathrm{e}=F+\Psi_\mathrm{e}$ stands for the total electric field in $\mathrm{S_e}$.

$$\begin{cases} \Psi_{\mathrm{e}}(s) - \Psi_{\mathrm{i}}(s) = f(s) \equiv -F(s), \\ \dfrac{\partial \Psi_{\mathrm{e}}(s)}{\partial \nu} - \dfrac{\partial \Psi_{\mathrm{i}}(s)}{\partial \nu} = g(s) \equiv -\dfrac{\partial F(s)}{\partial \nu}. \end{cases} \tag{8.42}$$

Thus we get a boundary-value problem for $\Psi_{\mathrm{e}}(\mathbf{r})$ and $\Psi_{\mathrm{i}}(\mathbf{r})$:

Problem 2 E-wave, dielectric. Find the electric fields $\Psi_{\mathrm{e}}(\mathbf{r})$ and $\Psi_{\mathrm{i}}(\mathbf{r})$ that satisfy (8.41), (8.5), and (8.42).

Note that in dealing with an H-wave problem, the second line of (8.42) should include the refractive index n or permittivity ε (see (8.100) in Appendix 1).

8.2.4.2 Modal Functions and Approximate Solutions

We need two sets of modal functions to solve the problem, one is for $\Psi_{\mathrm{e}}(\mathbf{r})$ and another is for $\Psi_{\mathrm{i}}(\mathbf{r})$. Let us call them the exterior and interior modal functions and represent them as $\{\varphi_{\mathrm{e}m}(\mathbf{r})\}$ and $\{\varphi_{\mathrm{i}m}(\mathbf{r})\}$. They should satisfy the requirements below, which are almost in common with the conditions from (**M1**) through (**M3**) given in Sect. 8.2.2.1

(**MD1**) Each member of the set of exterior modal functions satisfies the Helmholtz equation in $\mathrm{S_e}$ and meets the radiation condition at infinity.

(**MD2**) Each member of the set of interior modal functions satisfies the Helmholtz equation in $\mathrm{S_i}$.

(**MD3**) The sets of boundary values $\{\varphi_{\mathrm{e}m}(s)\}$ and $\{\varphi_{\mathrm{i}m}(s)\}$, and the sets of normal derivatives $\{\partial\varphi_{\mathrm{e}m}(s)/\partial\nu\}$ and $\{\partial\varphi_{\mathrm{i}m}(s)/\partial\nu\}$ are all complete in the function space **H**.

Here, we take the sets of separated solutions again because they are familiar to many people working with boundary-value problems. Then, the exterior and interior modal functions are:

$$\begin{cases} \varphi_{\mathrm{e}m}(\mathbf{r}) = H_m^{(2)}(kr)\exp(\mathrm{i}m\theta), \quad \varphi_{\mathrm{i}m}(\mathbf{r}) = J_m(nkr)\exp(\mathrm{i}m\theta) \\ \qquad (m = 0, \pm 1, \pm 2, \ldots). \end{cases} \tag{8.43}$$

Here, $J_m(nkr)$ stands for the Bessel function of order m. Then, we can define approximate solutions in $\mathrm{S_e}$ and $\mathrm{S_i}$ as

$$\begin{cases} \Psi_{\mathrm{e}N}(\mathbf{r}) = \displaystyle\sum_{m=-N}^{N} A_{\mathrm{e}m}(M)\varphi_{\mathrm{e}m}(s) \quad (\mathbf{r} \in \mathrm{S_e}), \\ \Psi_{\mathrm{i}N}(\mathbf{r}) = \displaystyle\sum_{m=-N}^{N} A_{\mathrm{i}m}(M)\varphi_{\mathrm{i}m}(s) \quad (\mathbf{r} \in \mathrm{S_i}). \end{cases} \tag{8.44}$$

They satisfy the Helmholtz equation in each region and $\Psi_{\mathrm{e}N}$ meets the radiation condition.

8.2.4.3 Error Estimation and Least-Squares Boundary Matching

After some analytical work we get error estimations similar to (8.18)[17]:

$$\begin{aligned} |\Psi_{eN}(\mathbf{r}) - \Psi_e(\mathbf{r})| \;&\le G_{e1}(S_{e0}) \left\| \frac{\partial \Psi_{eN}}{\partial \nu} - \frac{\partial \Psi_{iN}}{\partial \nu} - g \right\| \\ &\quad + G_{e2}(S_{e0}) \, \| \Psi_{eN} - \Psi_{iN} - f \| \quad (\mathbf{r} \in S_{e0} \subset S_e) \end{aligned} \tag{8.45}$$

and

$$\begin{aligned} |\Psi_{iN}(\mathbf{r}) - \Psi_i(\mathbf{r})| \;&\le G_{i1}(S_{i0}) \left\| \frac{\partial \Psi_{eN}}{\partial \nu} - \frac{\partial \Psi_{iN}}{\partial \nu} - g \right\| \\ &\quad + G_{i2}(S_{i0}) \, \| \Psi_{eN} - \Psi_{iN} - f \| \quad (\mathbf{r} \in S_{i0} \subset S_i). \end{aligned} \tag{8.46}$$

Here, S_{e0} and S_{i0} are arbitrary closed regions in S_e and S_i, and G_{pq} (p = e, i; q = 1, 2) are positive constants depending on S_{e0} and S_{i0}.

We can prove that: provided the sets of modal functions satisfy the requirement (**MD3**), there exists a sequence of pairs of approximate solutions

$$\begin{bmatrix} \Psi_{e0}(\mathbf{r}) \\ \Psi_{i0}(\mathbf{r}) \end{bmatrix}, \begin{bmatrix} \Psi_{e1}(\mathbf{r}) \\ \Psi_{i1}(\mathbf{r}) \end{bmatrix}, \ \cdots, \begin{bmatrix} \Psi_{eN}(\mathbf{r}) \\ \Psi_{iN}(\mathbf{r}) \end{bmatrix}, \ \cdots \tag{8.47}$$

whose boundary values and normal derivatives satisfy

$$E_N \equiv \frac{\left\| \Psi_{eN} - \Psi_{iN} - f \right\|^2}{\| f \|^2} + \frac{\left\| \partial \Psi_{eN}/\partial \nu - \partial \Psi_{iN}/\partial \nu - g \right\|}{\| g \|^2} \to 0 \quad (N \to \infty). \tag{8.48}$$

The sequence (8.47), hence, converges to the true solutions of the problem uniformly in wider sense in S_e and S_i:

$$\Psi_{pN}(\mathbf{r}) \to \Psi_p(\mathbf{r}) \quad (N \to \infty; \ \mathrm{p} = 1, 2; \ \text{uniformly in } S_{p0}). \tag{8.49}$$

Members of such a sequence can be found by solving the least-squares problem:

LSP 2: E-wave, dielectric. Find the modal coefficients $\{A_{pm}(M) : m = 0, \pm 1, \ldots, \pm N\}$ (p = e, i) that minimize the normalized mean-square error E_N defined in (8.48).

It is worth to note the following matters: Because the convergence in (8.48) is a consequence of completeness of the four sets of boundary functions in a product space $\mathbf{H} \times \mathbf{H}$, the choice of denominators in (8.48), $\| f \|^2$ and $\| g \|^2$, is no more than a

[17] We cannot include the derivation of equations from (8.45) through (8.49) because it takes much space. Interested readers can find the details in [35, 43–45]. The paper by Petit and Cadilhac [33] is also helpful.

convention to get non-dimensional quantities or to unify the units.[18] Speaking from a computational point of view, however, the ratio $\|f\|^2/\|g\|^2$ may have an effect on the condition number of **LSP 2** and, sometimes it is effective to introduce a parameter γ $(0 < \gamma < 1)$ to modify the definition of E_N as

$$E_N \equiv \gamma \frac{\left\|\Psi_{eN} - \Psi_{iN} - f\right\|^2}{\|f\|^2} + (1-\gamma)\frac{\left\|\partial\Psi_{eN}/\partial\nu - \partial\Psi_{iN}/\partial\nu - g\right\|^2}{\|g\|^2}. \quad (8.50)$$

The parameter should be determined by optimization to get a permissible condition number.

8.2.4.4 Notes on the Method of Numerical Computation

In above formulation we have $2M = 2(2N+1)$ unknowns. If we apply the rule in Sect. 8.3.1 (and also in Sect. 8.2.3.1), we need $2 \times 2M = 4M$ linear equations. The number of sampling points required for the $4M$ equations, however, is $2M$ again. This is because we have two equations at each sampling point: the first and the second equation of (8.42).

Let us follow the method of discretization in Sect. 8.2.3.1. Locating J $(= M = 2(2N+1))$ sampling points on C, we define J-dimensional vectors

$$\begin{cases} \mathbf{f} = [f(s_1)\ f(s_2)\ \cdots\ f(s_J)]^{\mathrm{T}}, \\ \mathbf{g} = [g(s_1)\ g(s_2)\ \cdots\ g(s_J)]^{\mathrm{T}}, \end{cases} \quad (8.51)$$

$$\begin{cases} \boldsymbol{\varphi}_{em} = [\varphi_{em}(s_1)\ \varphi_{em}(s_2)\ \cdots\ \varphi_{em}(s_J)]^{\mathrm{T}}, \\ \boldsymbol{\varphi}_{im} = [\varphi_{im}(s_1)\ \varphi_{im}(s_2)\ \cdots\ \varphi_{im}(s_J)]^{\mathrm{T}} \end{cases} \quad (8.52)$$

and

$$\begin{cases} \partial_\nu\boldsymbol{\varphi}_{em} = [\partial_\nu\varphi_{em}(s_1)\ \partial_\nu\varphi_{em}(s_2)\ \cdots\ \partial_\nu\varphi_{em}(s_J)]^{\mathrm{T}}, \\ \partial_\nu\boldsymbol{\varphi}_{im} = [\partial_\nu\varphi_{im}(s_1)\ \partial_\nu\varphi_{im}(s_2)\ \cdots\ \partial_\nu\varphi_{im}(s_J)]^{\mathrm{T}}, \end{cases} \quad (8.53)$$

where the mode-number m runs from $-N$ to N.

Next, we construct four $J \times M$ matrices

$$\begin{cases} \Phi_{11} = \left[\boldsymbol{\varphi}_{e,-N}\ \boldsymbol{\varphi}_{e,-N+1}\ \cdots\ \boldsymbol{\varphi}_{e,N}\right], \\ \Phi_{12} = \left[\boldsymbol{\varphi}_{i,-N}\ \boldsymbol{\varphi}_{i,-N+1}\ \cdots\ \boldsymbol{\varphi}_{i,N}\right] \end{cases} \quad (8.54)$$

and

$$\begin{cases} \Phi_{21} = \left[\partial_\nu\boldsymbol{\varphi}_{e,-N}\ \partial_\nu\boldsymbol{\varphi}_{e,-N+1}\ \cdots\ \partial_\nu\boldsymbol{\varphi}_{e,N}\right]. \\ \Phi_{22} = \left[\partial_\nu\boldsymbol{\varphi}_{i,-N}\ \partial_\nu\boldsymbol{\varphi}_{i,-N+1}\ \cdots\ \partial_\nu\boldsymbol{\varphi}_{i,N}\right]. \end{cases} \quad (8.55)$$

[18] The use of intrinsic impedance is also possible and is widely employed. That is: find the coefficients by minimization of $|\text{error in } \mathbf{E}|^2 + Z_0^2|\text{error in } \mathbf{H}|^2$. Here, Z_0 is the intrinsic impedance of vacuum or surrounding material. We use this formulation in Appendix 3.

Arranging the four matrices, we get a $2J \times 2M$ Jacobian matrix

$$\Phi = \begin{bmatrix} p\Phi_{11} & p\Phi_{12} \\ q\Phi_{21} & q\Phi_{22} \end{bmatrix}. \tag{8.56}$$

Here,

$$p = \frac{\gamma}{\mathbf{f}^\dagger \mathbf{f}}, \quad q = \frac{1-\gamma}{\mathbf{g}^\dagger \mathbf{g}} \tag{8.57}$$

are normalizing constants with the parameter γ appeared in (8.50).[19] Finally, defining a $2M$-dimensional solution vector

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_\mathrm{e} \\ \mathbf{A}_\mathrm{i} \end{bmatrix}, \tag{8.58}$$

where

$$\begin{cases} \mathbf{A}_\mathrm{e} = \left[A_{\mathrm{e},-N}(M) \; A_{\mathrm{e},-N+1}(M) \; \cdots \; A_{\mathrm{e},N}(M) \right]^\mathrm{T}, \\ \mathbf{A}_\mathrm{i} = \left[A_{\mathrm{i},-N}(M) \; A_{\mathrm{i},-N+1}(M) \; \cdots \; A_{\mathrm{i},N}(M) \right]^\mathrm{T} \end{cases} \tag{8.59}$$

are M $(= 2N + 1)$ dimensional column vectors. Thus, we can state a discretized problem as:

DLSP 2: E-wave, dielectric. Find the solution vector $\mathbf{A}$ that minimizes the discretized form of normalized boundary residual

$$E_{NJ} = \left\| \Phi \mathbf{A} - \begin{bmatrix} p\,\mathbf{f} \\ q\,\mathbf{g} \end{bmatrix} \right\|^2 = \left\| \begin{matrix} p\Phi_{11}\mathbf{A}_\mathrm{e} + p\Phi_{12}\mathbf{A}_\mathrm{i} - p\mathbf{f} \\ q\Phi_{21}\mathbf{A}_\mathrm{e} + q\Phi_{22}\mathbf{A}_\mathrm{i} - q\mathbf{g} \end{matrix} \right\|^2. \tag{8.60}$$

8.2.5 Application to Gratings

Here we consider the problem of plane-wave diffraction by a grating and state the points of difference from scattering by a cylindrical obstacle. The book edited by Petit [32] includes a nice introduction to Yasuura's method applied to grating problems as of late 70s.

8.2.5.1 Diffraction by a PC Grating

Figure 8.3 shows the cross section of a grating, an incident wave, and the system of coordinates. The cross section C is periodic in X with a period d and the surface is

[19] If the compensation by γ is not necessary, we can set $p = 1/\mathbf{f}^\dagger \mathbf{f}$ and $q = 1/\mathbf{g}^\dagger \mathbf{g}$ or use the intrinsic impedance.

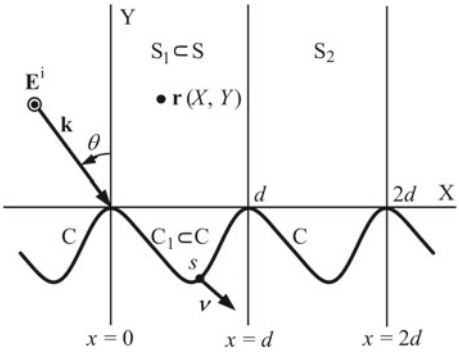

Fig. 8.3 Diffraction by a PC grating

uniform in Z. The semi-infinite region S over C is a vacuum and the region below C is occupied by a PC. We assume C is represented by a single-valued smooth function

$$\mathrm{C}: \ y = \eta(x), \tag{8.61}$$

where $\eta(x)$ is periodic in x, $\eta(x + d) = \eta(x)$, and (x, y) denotes a point on C.

Let an electromagnetic wave having an electric field

$$\mathbf{u}_Z F(\mathbf{r}) = \mathbf{u}_Z \exp(-\mathrm{i}kX \sin\theta + \mathrm{i}kY \cos\theta) \tag{8.62}$$

is incident on the grating. This case of polarization is termed E-wave, TE wave, or s-polarization.[20] Here, $\mathbf{r} = (X, Y)$ is a point in S, $\mathbf{u}_Z$ is a unit vector in Z, and θ is the angle of incidence shown in Fig. 8.3. The diffracted electric field has only a Z-component, which we describe by $\Psi(\mathbf{r})$. $\Psi(\mathbf{r})$ is the solution of the following problem.

Problem 3 E-wave, PC grating. Find $\Psi(\mathbf{r})$ that satisfies the conditions below:

(**GD1**) The 2-D Helmholtz equation in S;

(**GD2**) A radiation condition in Y that $\Psi(\mathbf{r})$ propagates or attenuates in positive Y;

(**GD3**) A periodicity condition

$$\Psi(X + d, Y) = \exp(-\mathrm{i}kd \sin\theta)\, \Psi(X, Y); \tag{8.63}$$

(**GD4**) The boundary condition

$$\Psi(x, \eta(x)) = f(x) \equiv -F(x, \eta(x)). \tag{8.64}$$

[20] s stands for senkrecht (German), which means the electric field is perpendicular to the plane of incidence, the plane spanned by $\mathbf{u}_Y$ (grating normal) and the incident wavevector.

Conditions (**GD1**) and (**GD4**) are common to the case of cylindrical obstacle in Sect. 8.2.1, while (**GD2**) is quite different from (**D2**) and (**GD3**) is a new requirement. These differences come from the pseudo-periodic nature of the problem. Because a grating has a periodic structure, and because we have assumed a plane-wave incidence, the phenomena at (X, Y) and $(X + d, Y)$ are almost the same; the only discrepancy can be seen in the phase difference (8.63). Hence, if we divide S by vertical lines $X = 0, \pm d, \pm 2d, \ldots$ as shown in Fig. 8.3, the diffracted fields in neighboring strip regions are the same except for the phase shift. This is a characteristic feature of a grating problem called quasi- or pseudo-periodicity and explains why the 1-D radiation condition appears in (**GD2**). In solving a grating problem, hence, we can assume that the observation point $\mathbf{r} = (X, Y)$ is inside the first strip region S_1 $(0 < X \leq d;\ Y \geq \eta(X))$ shown in Fig. 8.3.

8.2.5.2 Modal Functions, Approximate Solution, and Key Points in the Solution Method

Here again we choose separated solutions as modal functions. The separated solutions satisfying the periodicity are known as Floquet modes. We take the Floquet modes satisfying the radiation condition (**GD2**)

$$\varphi_m(\mathbf{r}) = \exp(-\mathrm{i}\alpha_m X - \mathrm{i}\beta_m Y) \quad (m = 0, \pm 1, \pm 2, \ldots) \tag{8.65}$$

as the set of modal functions, where

$$\alpha_m = k \sin\theta + \frac{2m\pi}{d}, \quad \beta_m = \sqrt{k^2 - \alpha_m^2} \quad (\mathrm{Re}\,\beta_m \geq 0,\ \mathrm{Im}\,\beta_m \leq 0). \tag{8.66}$$

The term $k \sin\theta$ in α_m is for the periodicity, the definition of β_m implies the Helmholtz equation, and the sign of β_m (positive or negative imaginary) is for the radiation condition.

We construct an approximate solution following the way we took in Sect. 8.2.2.2:

$$\Psi_N(\mathbf{r}) = \sum_{m=-N}^{N} A_m^{\mathrm{E}}(M)\, \varphi_m(\mathbf{r}). \tag{8.67}$$

This solution satisfies conditions (**GD1**), (**GD2**), and (**GD3**). Hence, the A_m^{E} coefficients[21] should be determined so that the solution satisfies the boundary condition (**GD4**) approximately. Let us see briefly the least-squares boundary matching works to yield a sequence of solutions converging to the true solution.

Some analysis starting from an assumption that $\mathbf{r}$ is inside a closed region S_{10} $(\subset S_1)$ leads us to an estimation

[21] The superscript E denotes that the coefficients concern the E-wave. Later we will also use the superscripts H, TE, and TM in accordance with polarizations.

$$|\Psi_N(\mathbf{r}) - \Psi(\mathbf{r})| \leq G(\mathrm{S}_{10})\|\tilde{\Psi}_N - \tilde{f}\| \quad (\mathbf{r} \in \mathrm{S}_{10} \subset \mathrm{S}_1). \tag{8.68}$$

Here, G is a positive constant depending on the closed region S_{10} and the quantities with tildes, e.g. $\tilde{f}$, mean periodic functions derived from the pseudo-periodic functions:

$$\tilde{f}(s) = \exp(\mathrm{i}\alpha_0 x) f(x, y) = -\exp(-\mathrm{i}\beta_0 y), \tag{8.69}$$

$$\tilde{\Psi}_N(s) = \exp(\mathrm{i}\alpha_0 x)\Psi_N(x, y) = \sum_{m=-N}^{N} A_m^{\mathrm{E}}(M)\,\tilde{\varphi}_m(x, y), \tag{8.70}$$

and

$$\tilde{\varphi}_m(x, y) = \exp(\mathrm{i}\alpha_0 x)\varphi_m(x, y) = \exp\left(-\frac{2m\pi \mathrm{i} x}{d} - \mathrm{i}\beta_m y\right). \tag{8.71}$$

The norm of a function $g(s)$ defined on C_1, the first period of C, is defined by

$$\|g\| = \left[\int_{s=0}^{C} |g(s)|^2 ds\right]^{1/2}, \tag{8.72}$$

where C denotes the length of C_1. Thus we have a least-squares problem:

LSP 3: E-wave, PC grating. Find the A_m^{E} coefficients that minimize the numerator of the normalized mean-square error

$$E_N = \frac{\left\|\tilde{\Psi}_N - \tilde{f}\right\|^2}{\left\|\tilde{f}\right\|^2}. \tag{8.73}$$

The modification of the boundary values to define the periodic functions is the key point in the solution of grating problems. Introducing the modification, we can establish a correspondence between one period of the grating surface C_1 and the cross section of a cylindrical obstacle C in Sect. 8.2.1.[22] The method of numerical solution for **LSP 3** is similar to that in Sect. 8.2.3. To solve the problem of diffraction by a grating made of dielectric or metal we can combine the method in this section with that in Sect. 8.2.4. Guidance to the problem of conical diffraction can be found in Appendix 3.

[22]If we employ the SP, this correspondence is essentially important because we need periodicity of the functions defined on the boundary. In using Yasuura's method without the SP, we can say the following points: (**1**) If we get the solution through the NE, this modification is not necessary because it is done automatically in calculating the inner products; (**2**) While if we employ the QRD or SVD: (**2.i**) The modification may accelerate the convergence of the solutions because the target function and the modal functions are periodically continuous after modification; (**2.ii**) And, a quadrature by parts (or rectangular-rule) approximation is equivalent to a trapezoidal-rule in numerical integrations.

8.3 Numerical Examples

In this section we show some results of numerical computations obtained by the methods in the last section. First, we examine the nature of the Jacobian matrices taking grating problems as examples to show the validity of the experimental rule (8.31). Meanwhile we add some comments that are useful in applying the method. Then we give the results of four problems of practical interest.

8.3.1 Rule on the Number of Sampling Points

We have solved the problem of diffraction by a grating made of PC and by one made of BK7 optical glass varying the number of sampling points or of linear equations. The results support our experimental rule. In addition, we have made a comparison between the two methods of locating the sampling points, (8.23) and (8.24), introduced in Sect. 8.2.3.1 and found little difference in the rage $J \geq 2M$ for the problem parameters employed in numerical analysis.

8.3.1.1 A PC Grating

We consider the grating shown in Fig. 8.3 and assume that the cross section C is given by[23]

$$\mathrm{C} : y = H\left(\cos\frac{2\pi x}{d} - 1\right). \tag{8.74}$$

We assume also that an E- or H-polarized plane wave is incident at $\theta = 0$ (normal incidence). Other physical parameters are: $d = 556\,\mathrm{nm}$, $H/d = 0.15$, and $\lambda = 500\,\mathrm{nm}$.[24] The computational parameters are: the number of truncation $N = 20$; the total number of modal functions $M = 41$; and the number of sampling points J is in the range $M \leq J \leq 4M$. This means that the number of unknown coefficients is M and the number of linear equations is between M and $4M$.

The first example, Fig. 8.4, shows the convergence of the solution and related parameters in the E-wave. The curves in Fig. 8.4a includes the maximum and minimum singular value, the condition number $\mathrm{cond}(\Phi)$, $E_{20\,J}$ of (8.30), and an error on the power balance

$$e_{NJ} = 1 - \sum_{\mathrm{prop}} \rho_m = 1 - \sum_{\beta_m > 0} \frac{\beta_m}{\beta_0} |A_m^{\mathrm{E}}(M, J)|^2, \tag{8.75}$$

[23] Note that the bias setting (we used -1 here) has an effect on the accuracy of numerical computation when the grating is deep.

[24] Although the use of normalization by wavelength (i.e., $kd = 2\pi d/\lambda$ etc.) is convenient in handling a problem with a PC obstacle, we employ real length here.

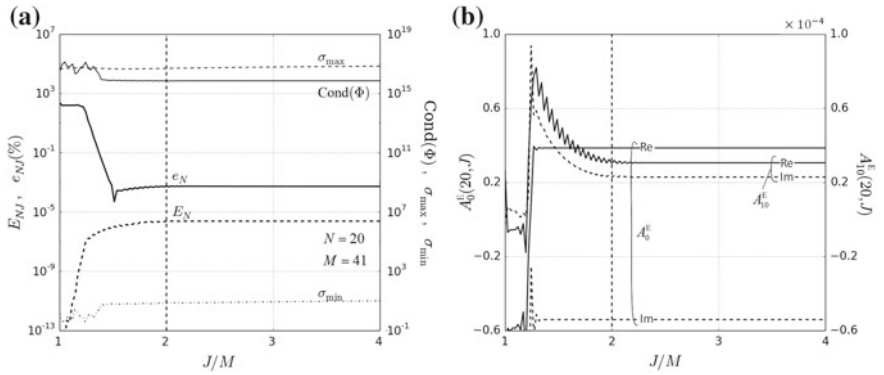

Fig. 8.4 Convergence of the solutions with respect to J for a PC grating in the E-wave. $\theta = 0$, $d = 556\,\text{nm}$, $H/d = 0.15$, and $\lambda = 500\,\text{nm}$: **a** Cond(Φ) and errors; **b** $A_0^{\text{E}}(20, J)$ and $A_{10}^{\text{E}}(20, J)$

where $\sum_{\text{prop}}$ and $\sum_{\beta_m > 0}$ mean the summation in respect to the propagating orders.[25] We observe these quantities are approaching final values with increasing J; and have converged for $J \geq 2M$.[26] Figure 8.4b illustrates the convergence of $A_0^{\text{E}}(20, J)$ and $A_{10}^{\text{E}}(20, J)$ coefficients. The former has converged before reaching $J = 2M$; while the latter is with small ripples until $J = 2.2M$. We, however, can neglect this oscillation in finding the diffracted wave because the mode with $m = 10$ is evanescent and cannot be observed at a point apart from the grating surface.

The second set of figures, Fig. 8.5, displays the same thing for the H-wave. The curves in Fig. 8.5a show the max and min singular value, cond(Φ), $E_{20\,J}$, and $e_{20\,J}$. While in Fig. 8.5b we show the convergence of $A_0^{\text{H}}(20, J)$ and $A_{10}^{\text{H}}(20, J)$. We observe all the quantities have converged substantially in the range $J \geq 2M$.

The third example, Fig. 8.6, shows the convergence of solutions: N dependence of the normalized mean-square error E_N and energy error e_N of E- and H-wave solutions. The rule $J = 2M$ is applied. Because the surface modulation is moderate in the problem, we get precise solutions with 10^{-6} or 10^{-4} percent energy error easily for both E- and H-wave problem.[27] It is worth to mention that a modal coefficient—e.g. $A_m^{\text{E}}(M)$, as a function of M, converges to a final value: $A_m^{\text{E}}(M) \to A_m^{\text{E}}$ $(M \to \infty)$. This convergence, however, is not uniform with respect to m.

[25] ρ_m is referred to as the (reflection) efficiency of the mth order.

[26] When the number of truncation is small (e.g., $N \leq 10$), we sometimes observe a phenomenon that the condition number continues to decrease slightly beyond $J = 2M$ due to tiny increment of $\sigma_{\min}$.

[27] We should notice, however, that the accuracy of an H-wave solution is lower than that of an E-wave solution by one or two digits. This is observed generally; and was Yasuura's motivation of introducing the SP. His idea came from the fact that a Neumann problem for an electrostatic potential is equivalent to a Dirichlet problem for a stream function. The prototype of the SP, hence, was called *an algorithm using the stream function in a wave field.*

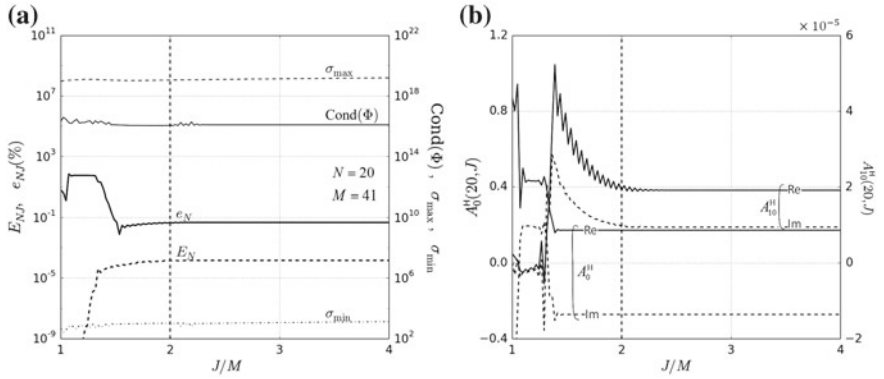

Fig. 8.5 Convergence of the solutions with respect to J for a PC grating in the H-wave. Optical parameters are the same as those in Fig. 8.4: **a** Cond(Φ) and errors; **b** $A_0^H(20, J)$ and $A_{10}^H(20, J)$

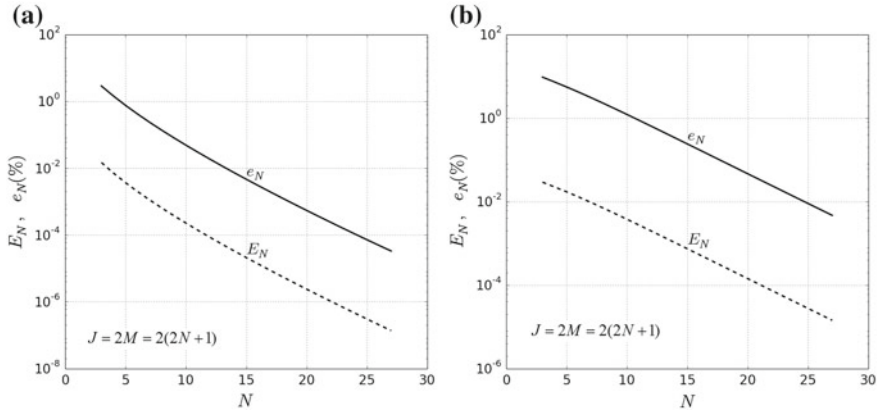

Fig. 8.6 Convergence of the solutions with respect to N for a PC grating. Optical parameters are the same as those in Fig. 8.4: **a** E-wave; **b** H-wave

8.3.1.2 A BK7 Optical Glass Grating

Here we examine the case of a dielectric grating made of an optical glass BK7 whose refractive index is 1.5139 [1]. Other parameters are the same as in Sect. 8.3.1.1. In the present problem we have transmitted fields ($\mathbf{E}^t$ and $\mathbf{H}^t$) in the region V_2 below the grating surface in addition to the reflected fields ($\mathbf{E}^r$ and $\mathbf{H}^r$) over the grating V_1. We, hence, define approximate solutions following (8.44) in Sect. 8.2.4. That is, we employ Floquet modes in V_1 and V_2 and construct approximations of leading fields in each region in the form of finite linear combinations of the Floquet modes. Let the number of truncation be N. Then, we have $2(2N + 1) = 2M$ unknown coefficients in total.

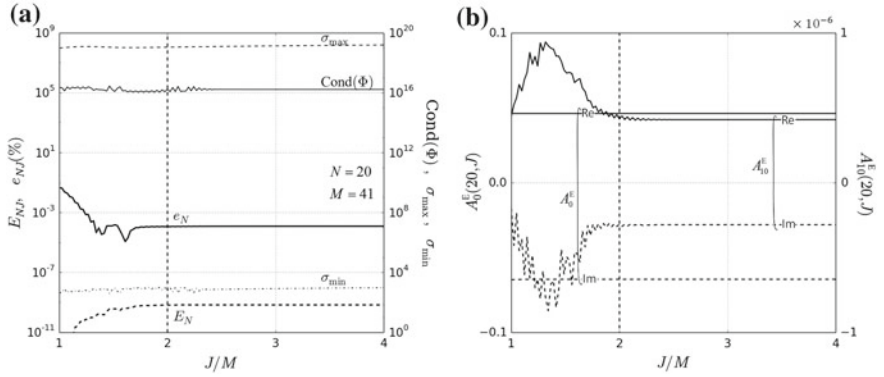

Fig. 8.7 Convergence of the solutions with respect to J for a BK7 grating in the E-wave. $\theta = 0$, $d = 556\,\text{nm}$, $H/d = 0.15$, and $\lambda = 500\,\text{nm}$: **a** Cond(Φ) and errors; **b** $A_0^{\text{E}}(20, J)$ and $A_{10}^{\text{E}}(20, J)$

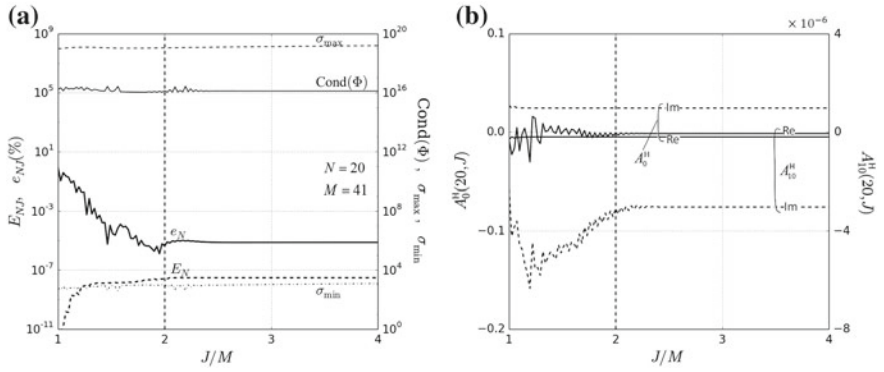

Fig. 8.8 Convergence of the solutions with respect to J for a BK7 grating in the H-wave. Optical parameters are the same as those in Fig. 8.7: **a** Cond(Φ) and errors; **b** $A_0^{\text{H}}(20, J)$ and $A_{10}^{\text{H}}(20, J)$

Figure 8.7 shows the convergence of the solution and related parameters as functions of the number of sampling points J in the E-wave and $N = 20$. We observe that all the errors and the parameters have converged in the range $J \geq 2M = 2(2N + 1)$ except for small ripples. Figure 8.8 shows the same thing in the H-wave. This means that the number of linear equations in the least-squares problem can be twice as many as the number of unknowns ($2 \times 2M = 4(2N + 1)$; see Sect. 8.2.4.4).

Figure 8.9 illustrates the N dependence of the errors of the solutions. We get precise solutions with 10^{-5} percent energy error easily on a personal computer.

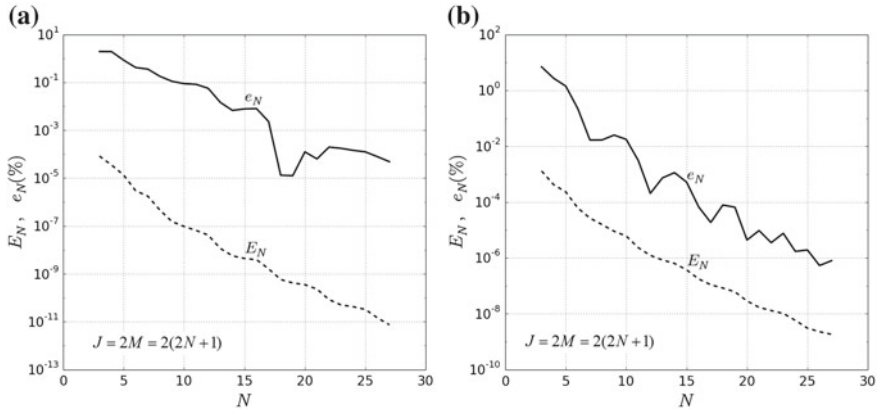

Fig. 8.9 Convergence of the solutions with respect to N for a BK7 grating. Optical parameters are the same as those in Fig. 8.7: **a** E-wave; **b** H-wave

8.3.2 Scattering by Relatively Deep Gratings

Yasuura's method, when combined with the partition of the groove region, can solve the problem of diffraction from a deep grating with a depth-to-period ratio beyond unity. In the conventional Yasuura's method without partition, this ratio is said to be about 0.5 and a little less than 0.4 in the E- and H-wave cases, respectively. In the present subsection, some numerical results are given for the scattering by relatively deep gratings using a combination of up-and down-going Floquet modal functions [22].

The period and height of the sinusoidal profile are d and $2H$, respectively, as shown in Fig. 8.10. At first we deal with a perfectly conducting grating as a fundamental problem where the electromagnetic fields exist only in the vacuum region. The semi-infinite region over the grating surface is divided into an upper half plane U_0 and a groove region a fictitious boundary (a horizontal line). The latter is further divided into shallow horizontal layers $\mathrm{U}_1, \mathrm{U}_2, \ldots, \mathrm{U}_Q$ again by fictitious boundaries.

An approximate solution in U_0, that is $\Psi_{0N}(\mathbf{r})$, is defined in a usual manner as (8.67), while the solutions in U_q $(q = 1, 2, \ldots, Q)$ include not only the up-going but also the down-going modal functions as

$$\Psi_{qN}(\mathbf{r}) = \sum_{m=-N}^{N} \left[A_{qm}^{+}(N)\, \varphi_m^{+}(\mathbf{r} - \mathbf{u}_Y y_q) + A_{qm}^{-}(N)\, \varphi_m^{-}(\mathbf{r} - \mathbf{u}_Y y_{q-1}) \right], \quad (8.76)$$

where $\varphi_m^{\pm}(\mathbf{r}) = \exp(\mathrm{i}\alpha_m X \pm \mathrm{i}\beta_m Y)$, and the plane $Y = y_q$ is the boundary between U_q and U_{q+1}. Thus the total number of unknown coefficients is $(2N + 1)(2Q + 1)$. These coefficients should be determined in order that the solutions meet the boundary condition (**GD4**) and an additional set of boundary conditions on the Q fictitious boundaries:

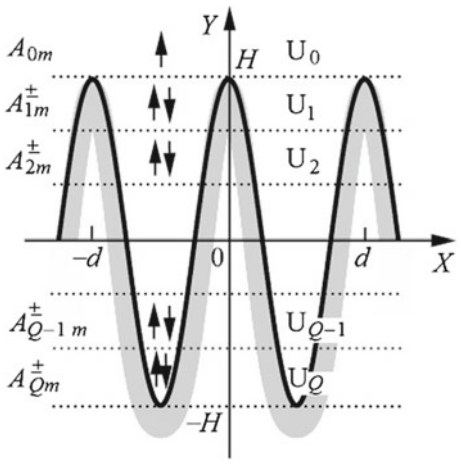

Fig. 8.10 Cross section of a perfectly conducting (PC) or dielectric sinusoidal grating. In the case of PC grating, partition into $Q+1$ layers is done only in the vacuum region

$$\begin{cases} \left(F\delta_{q0}+\Psi_q\right)\big|_{Y=y_q+0} = \Psi_{q+1}\big|_{Y=y_q-0}, \\ \dfrac{\partial\left(F\delta_{q0}+\Psi_q\right)}{\partial Y}\bigg|_{Y=y_q-0} = \dfrac{\partial\Psi_{q+1}}{\partial Y}\bigg|_{Y=y_q-0}, \end{cases} \tag{8.77}$$

where δ_{q0} is Kronecker's delta. The mean-square error is defined in the same form as (8.73), but the integration range in the norm (8.72) must include not only the grating surface but also the fictitious boundaries.

Let us check the convergence of the results obtained by the present method. Figure 8.11 shows the variation of the normalized mean-square error and the energy error as functions of the number of truncation N for both E- and H-wave incidence. As is observed in these figures, the mean-square error decreases as N increases. An approximate solution with 0.1 percent energy error is accomplished at $N = 14$ for an E-wave. In the H-wave case convergence of solutions is not so fast as in the E-wave case. We attain to one percent energy error at $N = 23$ in that case of polarization.

Figure 8.12 shows comparison of reflection efficiency for a perfectly conducting grating as functions of the incident angle at E-wave incidence. The numbers (N, Q) are (15, 4), (15, 5), and (30, 20) as $H/d = 0.31$, 0.4, and 1.066, respectively. The curves and symbols represent the present results and the results by the integral equation method [46]. We find good agreement between the results. For dielectric gratings, partition must be made not only in the vacuum region but also in the dielectric one. As a result, numbers of unknown modal coefficients and boundary conditions become doubled compared with the previous case.

Figure 8.13 shows comparison of transmission efficiency for a dielectric grating as functions of the incident angle at H-wave incidence. The numbers (N, Q) are (11,4). The curves and symbols represent the present results and the results by the finite element method [20]. We find that the results agree with each other except for the grazing limit.

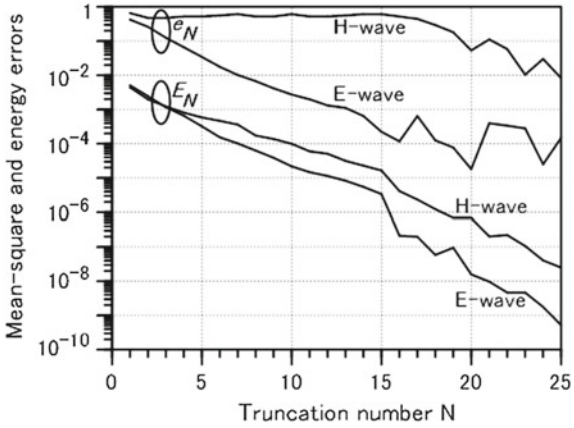

Fig. 8.11 Normalized mean-square error E_N and energy error e_N as functions of the truncation number N for a PC grating with $2H/d = 1$, $d/\lambda = 0.5$, and $\theta = 30°$

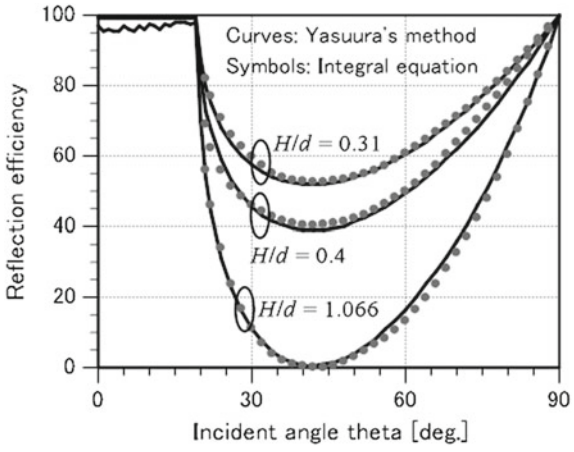

Fig. 8.12 Reflection efficiency in percent for a PC grating with $d/\lambda = 0.75$ at E-wave incidence. Comparison with the integral equation method [46]

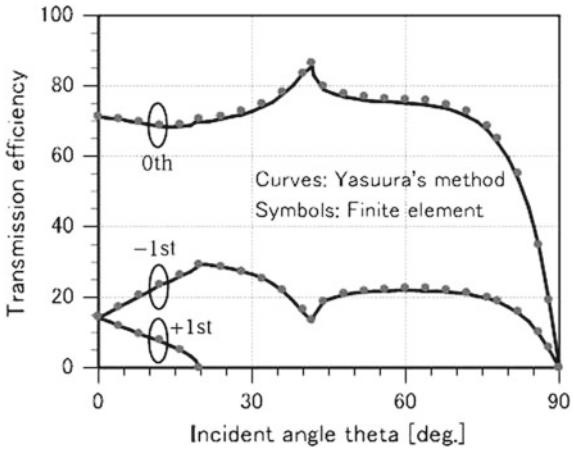

Fig. 8.13 Transmission efficiency in percent for a dielectric grating with $\varepsilon_r = 4$, $2H/d = 1$, and $d/\lambda = 0.6$ at H-wave incidence. Comparison with the finite element method [20]

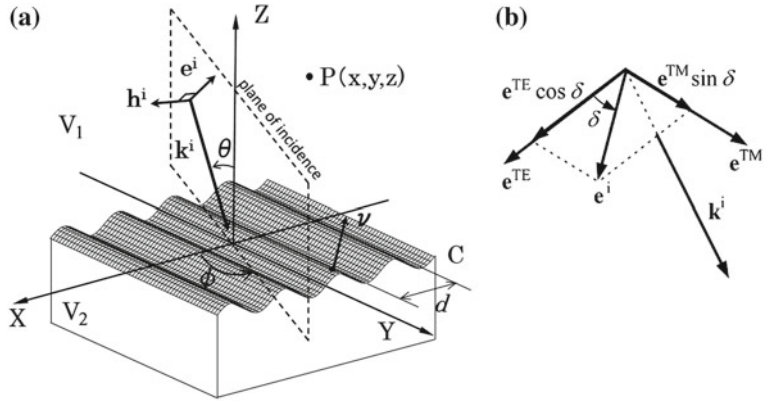

Fig. 8.14 Diffraction by a dielectric grating: **a** Conical mounting; **b** Definition of δ

Although there are a couple of methods that are capable of solving the problems of extremely deep gratings, the present results make sense because they show a limit of a conventional modal-expansion approach when using the Floquet modes as basis functions.

8.3.3 Plasmon Surface Waves Excited on a Metal Grating Placed in Conical Mounting

We show some numerical results in regard to plasmon surface waves excitation on a metal grating placed in conical mounting [29]. Conical mounting is an optical arrangement in which the plane of incidence is not perpendicular to grooves of a grating as shown in Fig. 8.14. Readers can find detailed description of problem in Appendix 3. We here illustrate the results obtained by the method explained there.

We deal with a sinusoidal silver grating whose surface profile is given by $z = H\sin(2\pi x/d)$. The upper region V_1 over the grating surface is assumed to be vacuum with a refractive index $n_1 = 1$ and the grating is made of silver with a complex refractive index n_2. As an incident light we consider an electromagnetic plane wave, which is specified by the wavenumber in vacuum ($k = 2\pi/\lambda$), the polar angle (θ) between the wavevector and the grating normal, and the azimuthal angle (ϕ) between the X axis and the plane of incidence.

The diffracted fields in the conical mounting are decomposed into a TE and a TM component which mean that the relevant electric and magnetic field are perpendicular to the plane of incidence. The efficiency of the mth-order diffracted mode in V_1, hence, is represented as $\rho_m = \rho_m^{\mathrm{TE}} + \rho_m^{\mathrm{TM}}$. Here, ρ_m^{TE} or ρ_m^{TM} is the efficiency of the

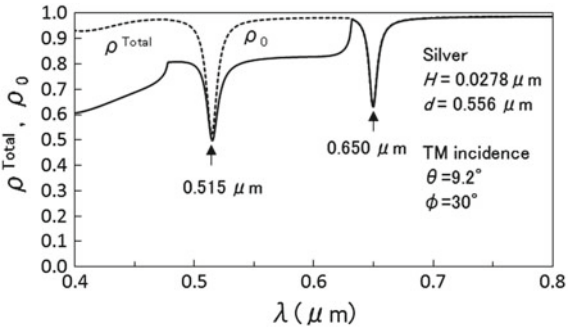

Fig. 8.15 ρ_0 and ρ^{Total} as functions of wavelength λ

TE- or TM-component of the mth-order diffracted mode.[28] In the numerical examples below we deal with a shallow grating made of silver with a period $d = 0.556\,\mu\text{m}$ and an amplitude $H = 0.0278\,\mu\text{m}$. Yasuura's method provides sufficiently reliable results for the problem of such a grating at the truncation number of the approximate solutions $N = 10$.

Figure 8.15 shows the efficiency of the 0th-order diffracted mode ρ_0 and the total diffraction efficiency ρ^{Total} as functions of wavelength λ.[29] The incident light is in the TM incidence—a polarization angle $\delta = \pi/2$ in (8.114) of Appendix 3—where the magnetic field is perpendicular to the plane of incidence. The polar angle and the azimuthal angle are chosen as $\theta = 9.2°$ and $\phi = 30°$. As a complex refractive index of silver n we take the interpolated values for the experimental data in the literature [8]. In the figure we observe partial absorption of incident light at $\lambda = 0.515\,\mu\text{m}$ and $\lambda = 0.650\,\mu\text{m}$ as dips in the total efficiency curve.[30] As we will see later, the dips are associated with plasmon resonance absorption, which is caused by coupling of surface plasmons with an evanescent mode diffracted by the grating [21, 34].

Figure 8.16 shows the 0th-order efficiency ρ_0 and the TE and TM component ρ_0^{TM} and ρ_0^{TE} as functions of θ with a fixed azimuthal angle $\phi = 30°$. The wavelength is chosen as $\lambda = 0.650\,\mu\text{m}$ and a refractive index is $n_2 = 0.07 - \text{i}4.2$. Remaining parameters are the same as those of Fig. 8.15. We observe in Fig. 8.16 partial absorption of incident light at $\theta = 9.2°$, we call it a resonance angle, as a dip in the ρ_0 curve. In addition we notice that ρ_0^{TM} takes a minimal value at the resonance angle, but ρ_0^{TE} increases there to the contrary. This illustrates the enhancement of TM-TE mode conversion [5] that a TM component of the incident light is strongly converted into a TE component of the 0th-order diffracted light when plasmon resonance absorption occurs in a metal grating in conical mounting.

---

[28] The efficiencies are given by $\rho_m^{\text{TE}} = (\gamma_{1m}/\gamma_{10})\,|A_{1m}^{\text{TE}}|^2$ and $\rho_m^{\text{TM}} = (\gamma_{1m}/\gamma_{10})\,|A_{1m}^{\text{TM}}|^2$ where γ_{1m} is the propagation constant in the Z-direction of the mth-order propagating mode ($\text{Re}\,(\gamma_{1m}) \geq 0$) concerning the upper region V_1, and A_{1m}^{TE} and A_{1m}^{TM} are the expansion coefficients of the approximate solutions defined in (8.120) of Appendix 3.

[29] ρ^{Total} is a summation of ρ_m over the propagating orders.

[30] $1 - \rho^{\text{Total}}$ represents the ratio of the absorbed light power by a metal grating to the incident light power.

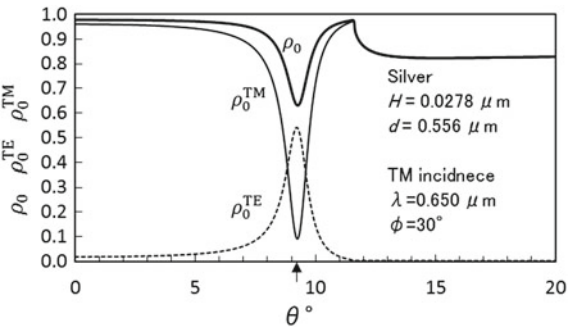

Fig. 8.16 ρ_0, ρ_0^{TE} and ρ_0^{TM} as functions of θ

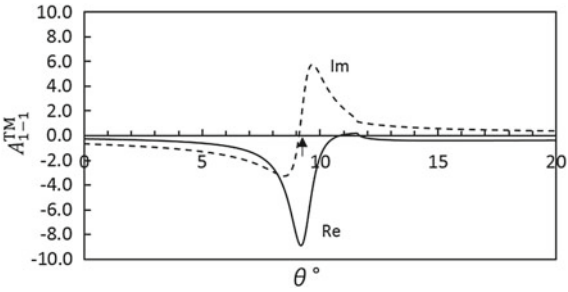

Fig. 8.17 A_{1-1}^{TM} as functions of θ

In Fig. 8.17 we show the expansion coefficient of the -1st-order TM vector modal function A_{1-1}^{TM} defined in (8.120) of Appendix 3 as a function of θ. The parameters in the figure are the same as those of Fig. 8.16 where the -1st-order mode is evanescent. The solid curve in Fig. 8.17 represents the real part of the expansion coefficient and the dashed curve is the imaginary part. From this result we observe the resonance property of the expansion coefficient A_{1-1}^{TM} at the angle of incidence $\theta = 9.2°$ and confirm that the TM component of the -1st-order evanescent mode couples with surface plasmons at the resonance angle. We thus demonstrate that plasmon resonance absorption is associated with coupling of surface plasmons with an evanescent mode diffracted by a metal grating.

We note that the excitation of surface plasmons is largely affected by the azimuthal angle ϕ. Figure 8.18 shows the plasmon resonance absorption for several ϕ's under the same parameters as those of Fig. 8.16. We observe that the resonance angle varies with ϕ as shown in Fig. 8.18. This means direction of propagation depends on ϕ, the direction in which the plasmon surface wave propagates. The azimuthal angle ϕ has also large influence on the enhancement of TM-TE mode conversion through plasmon resonance absorption. For example, a TM component of the 0th-order diffracted mode almost vanishes at the resonance absorption at $\phi = 45°$, but a TE component becomes to be 0.7 there.

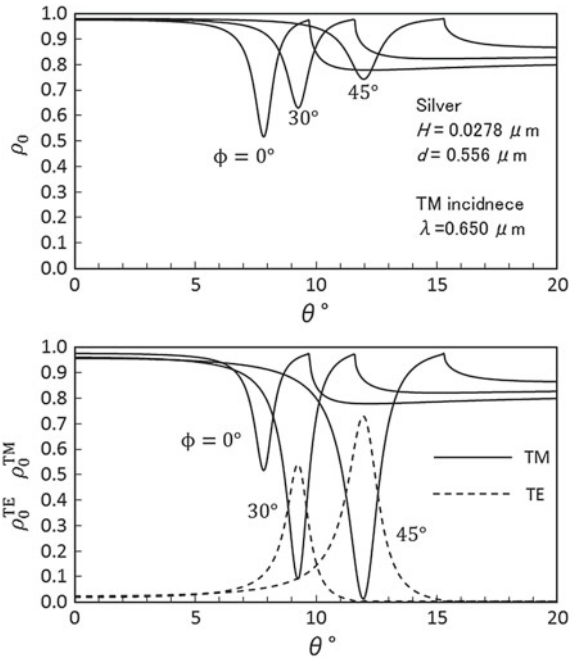

Fig. 8.18 Dependence of TE-TM mode conversion on ϕ

8.3.4 Scattering by a Metal Bigrating

In this subsection we deal with a 3-D problem: diffraction by a metal bigrating whose surface profile is periodically corrugated in two directions. We briefly describe the formulation of Yasuura's method for solving the problem by a metal bigrating and then show numerical results of plasmon resonance absorption in the grating [16].

We consider a bisinusoidal metal grating shown in Fig. 8.19. The surface profile of the grating is given by

$$\eta(x, y) = H\left[\sin\left(\frac{2\pi x}{d}\right) + \sin\left(\frac{2\pi y}{d}\right)\right]. \tag{8.78}$$

The upper region V_1 over the grating surface S_0 is vacuum with a refractive index $n_1 = 1$ and the region V_2 below the grating surface consists of a lossy metal with a complex refractive index n_2. The permeability of the metal is assumed to be μ_0.

The incident light is an electromagnetic plane wave

$$\begin{bmatrix} \mathbf{E}^{\mathrm{i}} \\ \mathbf{H}^{\mathrm{i}} \end{bmatrix}(\mathbf{r}) = \begin{bmatrix} \mathbf{e}^{\mathrm{i}} \\ \mathbf{h}^{\mathrm{i}} \end{bmatrix} \exp(-\mathrm{i}\mathbf{k}^{\mathrm{i}} \cdot \mathbf{r}). \tag{8.79}$$

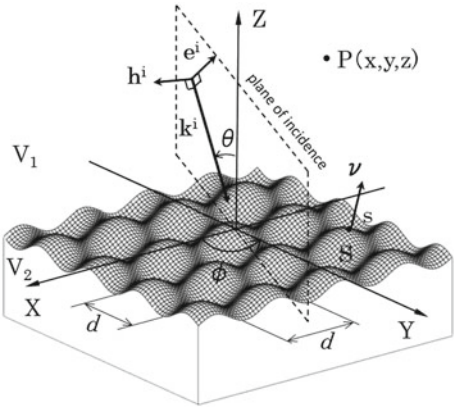

Fig. 8.19 A bisinusoidal grating and coordinate system

Here, $\mathbf{r}$ is the position vector for an observation point, $\mathbf{k}^{\mathrm{i}}$ is the wavevector of the incident wave, and $\mathbf{h}^{\mathrm{i}} = (1/\omega\mu_0)\,\mathbf{k}^{\mathrm{i}} \times \mathbf{e}^{\mathrm{i}}$. The wavevector is given by

$$\mathbf{k}^{\mathrm{i}} = (\alpha, \beta, -\gamma) \tag{8.80}$$

with $\alpha = n_1 k \sin\theta \cos\phi$, $\beta = n_1 k \sin\theta \sin\phi$, and $\gamma = n_1 k \cos\theta$. Here, k $(= 2\pi/\lambda)$ is the wavenumber in vacuum, and θ is the polar angle between the Z axis and the incident wavevector, and ϕ is the azimuthal angle between the X axis and the plane of incidence.

We denote the diffracted electric and magnetic fields by $\mathbf{E}^{\mathrm{d}}_{\ell}(\mathrm{P})$, $\mathbf{H}^{\mathrm{d}}_{\ell}(\mathrm{P})$ in the regions V_{ℓ} $(\ell = 1, 2)$. Here we explain briefly Yasuura's method for finding the diffracted fields. We first introduce TE and TM vector modal functions defined in the region V_{ℓ} $(\ell = 1, 2)$:

$$\begin{cases} \boldsymbol{\varphi}^{\mathrm{TE}}_{\ell mn}(\mathbf{r}) = \mathbf{e}^{\mathrm{TE}}_{\ell mn} \exp(-\mathrm{i}\mathbf{k}_{\ell mn} \cdot \mathbf{r}), \quad \mathbf{e}^{\mathrm{TE}}_{\ell mn} = \dfrac{\mathbf{k}_{\ell mn} \times \mathbf{u}_Z}{|\mathbf{k}_{\ell mn} \times \mathbf{u}_Z|}, \\ \boldsymbol{\varphi}^{\mathrm{TM}}_{\ell mn}(\mathbf{r}) = \mathbf{e}^{\mathrm{TM}}_{\ell mn} \exp(-\mathrm{i}\mathbf{k}_{\ell mn} \cdot \mathbf{r}), \quad \mathbf{e}^{\mathrm{TM}}_{\ell mn} = \dfrac{\mathbf{e}^{\mathrm{TE}}_{\ell mn} \times \mathbf{k}_{\ell mn}}{\left|\mathbf{e}^{\mathrm{TE}}_{\ell mn} \times \mathbf{k}_{\ell mn}\right|} \\ \qquad (m,\ n = 0, \pm 1, \pm 2, \ldots). \end{cases} \tag{8.81}$$

Here, $\mathbf{u}_Z$ is a unit vector in the Z-direction and $\mathbf{k}_{\ell mn}$ $(\ell = 1, 2)$ is the wavevector of the (m, n)th-order diffracted wave:

$$\mathbf{k}_{1mn} = (\alpha_m, \beta_n, \gamma_{1mn}), \quad \mathbf{k}_{2mn} = (\alpha_m, \beta_n, -\gamma_{2mn}) \tag{8.82}$$

with

$$\begin{cases} \alpha_m = \alpha + \dfrac{2m\pi}{d}, \quad \beta_n = \beta + \dfrac{2n\pi}{d}, \quad \gamma_{\ell m} = \sqrt{(n_\ell k)^2 - (\alpha_m^2 + \beta_n^2)} \\ \qquad (\mathrm{Re}\,\gamma_{\ell mn} \geq 0,\ \mathrm{Im}\,\gamma_{\ell mn} \leq 0). \end{cases} \tag{8.83}$$

We form approximate solutions for the diffracted electric and magnetic fields:

$$\begin{bmatrix} \mathbf{E}^{\mathrm{d}}_{\ell N} \\ \mathbf{H}^{\mathrm{d}}_{\ell N} \end{bmatrix}(\mathbf{r}) = \sum_{m,n=-N}^{N} A^{\mathrm{TE}}_{\ell mn} \begin{bmatrix} \boldsymbol{\varphi}^{\mathrm{TE}}_{\ell mn} \\ \boldsymbol{\psi}^{\mathrm{TE}}_{\ell mn} \end{bmatrix}(\mathbf{r}) + \sum_{m,n=-N}^{N} A^{\mathrm{TM}}_{\ell mn} \begin{bmatrix} \boldsymbol{\varphi}^{\mathrm{TM}}_{\ell mn} \\ \boldsymbol{\psi}^{\mathrm{TM}}_{\ell mn} \end{bmatrix}(\mathbf{r}) \quad (\ell = 1, 2) \tag{8.84}$$

with

$$\boldsymbol{\psi}^{\mathrm{q}}_{\ell mn}(\mathbf{r}) = \frac{1}{\omega\mu_0} \mathbf{k}_{\ell \mathrm{mn}} \times \boldsymbol{\varphi}^{\mathrm{q}}_{\ell mn}(\mathbf{r}) \quad (\mathrm{q} = \mathrm{TE}, \mathrm{TM}). \tag{8.85}$$

The expansion coefficients $A^{\mathrm{TE}}_{\ell mn}$, $A^{\mathrm{TM}}_{\ell mn}$ are determined so that the approximate solutions $\mathbf{E}^{\mathrm{d}}_{\ell N}(\mathrm{P}), \mathbf{H}^{\mathrm{d}}_{\ell N}(\mathrm{P})$ satisfy the boundary conditions in a weighted least-squares sense. To do this, we minimize the mean-square error

$$\begin{aligned} E_N = &\int_{\mathrm{S}} \left| \boldsymbol{\nu} \times \left(\mathbf{E}^{\mathrm{d}}_{1N} + \mathbf{E}^{\mathrm{i}} - \mathbf{E}^{\mathrm{d}}_{2N} \right)(s) \right|^2 dS \\ &+ Z_0^2 \int_{\mathrm{S}} \left| \boldsymbol{\nu} \times \left(\mathbf{H}^{\mathrm{d}}_{1N} + \mathbf{H}^{\mathrm{i}} - \mathbf{H}^{\mathrm{d}}_{2N} \right)(s) \right|^2 dS, \end{aligned} \tag{8.86}$$

where S is one period cell of the grating surface S_0, $\boldsymbol{\nu}$ is a unit normal vector to the grating surface, and Z_0 is an intrinsic impedance of the medium of V_1.

The mean-square error E_N is discretized by applying a two-dimensional trapezoidal rule where the number of divisions in the X- and Y-directions is chosen to be $J = 2(2N + 1)$. The discretized LSP with $24(2N + 1)^2 \times 4(2N + 1)^2$ Jacobian is solved by QRD.

The diffraction efficiency ρ_{mn} of the (m, n)th-order mode ($\gamma_{1m} \geq 0$) in V_1 is given by

$$\rho_{mn} = \rho^{\mathrm{TE}}_{mn} + \rho^{\mathrm{TM}}_{mn}, \tag{8.87}$$

where the efficiency of the (m, n)th-order TE or TM mode is given by

$$\rho^{\mathrm{TE}}_{mn} = \frac{\gamma_{1m}}{\gamma} |A^{\mathrm{TE}}_{1mn}|^2, \quad \rho^{\mathrm{TM}}_{mn} = \frac{\gamma_{1m}}{\gamma} |A^{\mathrm{TM}}_{1mn}|^2. \tag{8.88}$$

We show the plasmon resonance absorption in a bisinusoidal grating made of silver [12]. We consider a shallow bisinusoidal grating with a corrugation depth $H = 0.0075\,\mu\mathrm{m}$ and a period $d = 0.556\,\mu\mathrm{m}$. The wavelength of the incident light is chosen as $\lambda = 0.650\,\mu\mathrm{m}$ where only the $(0, 0)$th-order diffracted mode propagates. We take $n_2 = 0.07 - \mathrm{i}4.2$ as the refractive index of silver at this wavelength.

Figure 8.20 shows the diffraction efficiency of the $(0, 0)$th-order diffracted mode ρ_{00} as functions of the polar angle θ when the azimuthal angle $\phi = 30^\circ$ is fixed. In the efficiency curve we observe four dips A, B, C, and D at which incident light power is strongly absorbed by the grating. The dips are associated with absorption that is caused by the coupling of surface plasmons with an evanescent mode diffracted by

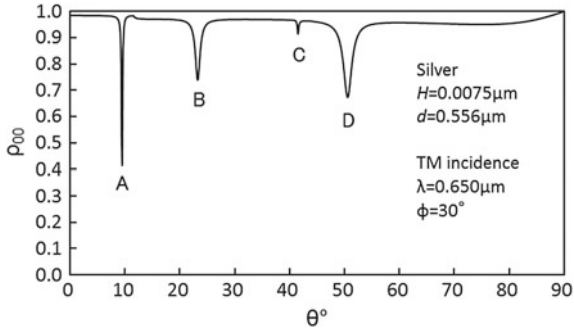

Fig. 8.20 ρ_{00} as functions of θ

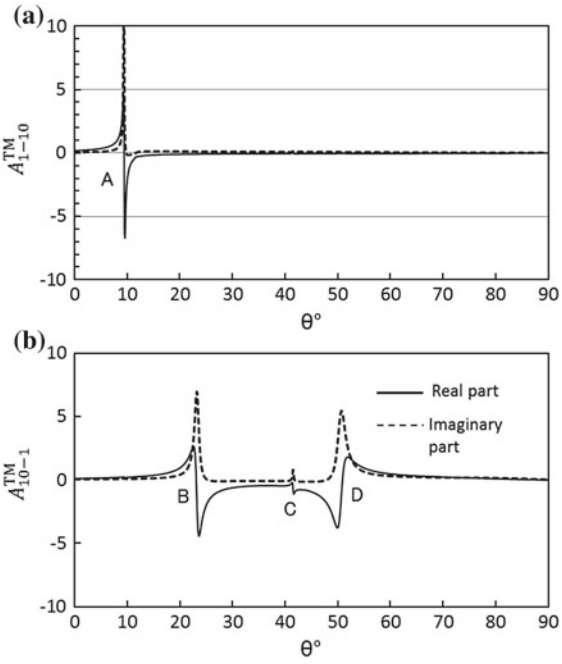

Fig. 8.21 Expansion coefficients as functions of θ: **a** $A^{\mathrm{TM}+}_{1\,-10}$; **b** $A^{\mathrm{TM}+}_{10\,-1}$

a bisinusoidal silver grating. This is confirmed from Fig. 8.21 where the expansion coefficients (a) A^{TM}_{1-10} and (b) A^{TM}_{10-1} are plotted as functions of θ under the same parameters as in Fig. 8.20. The solid curves in Fig. 8.21 represent the real part of the expansion coefficient and the dashed curves are the imaginary part. In Fig. 8.21a, a resonance property of the A^{TM}_{1-10} curve at $\theta = 9.5°$, i.e., a dip A, illustrates that the TM component of the $(-1, 0)$ evanescent mode couples with surface plasmons at a dip A. From the resonant property of the A^{TM}_{10-1} curve in Fig. 8.21b we confirm that dips B and D are associated with the coupling of the $(0, -1)$ evanescent mode with surface plasmons. Similarly, we can show a dip C is caused by coupling of the $(-1, -1)$ evanescent mode.

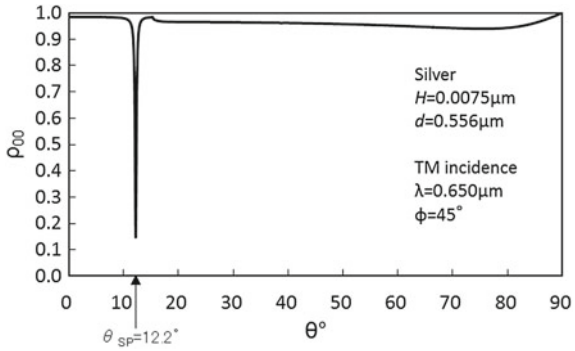

Fig. 8.22 Simultaneous resonance absorption in a bisinusoidal silver grating

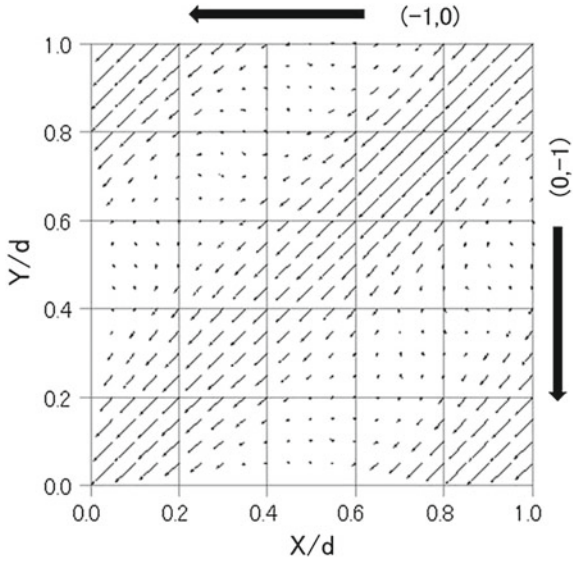

Fig. 8.23 Power flows of the total field when simultaneous resonance absorption occurs

When an incident light with $\phi = 45°$ illuminates a bisinusoidal grating at the specific angle of θ, i.e., the resonance angle, two surface plasmon waves are excited and propagate in directions symmetric with respect to the plane of incidence. The absorption associated with the two surface plasmon waves is called simultaneous resonance absorption [12]. Figure 8.22 shows an example of the simultaneous resonance absorption where the $(-1, 0)$th- and $(0, -1)$st-order evanescent modes couple simultaneously with two surface plasmon waves at the same polar angle $\theta = 12.2°$. The two surface plasmon waves excited simultaneously on the grating surface interact with each other and the interference of the surface plasmon waves causes the standing wave in the vicinity of the grating surface. This is confirmed in Fig. 8.23, where the X and Y components of Poynting's vector $\mathbf{S}$ on the surface $0.01d$ above the one-unit cell of the grating surface are plotted as the vector (S_X, S_Y).

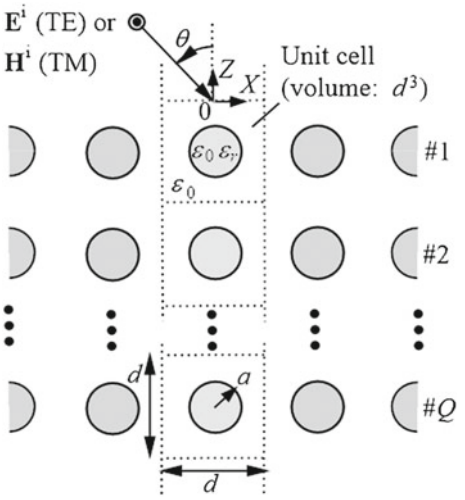

Fig. 8.24 Scattering by dielectric spheres. The structure is periodic both parallel to and perpendicularly to the page

8.3.5 Scattering by Periodically Located Spheres

Some numerical results are given for the scattering by dielectric spheres located periodically in three directions [17]. This kind of structure is a fundamental model of photonic crystals having properties of electromagnetic or optical band gaps.

As shown in Fig. 8.24, the structure is composed by stacking cubic unit cell regions with a volume d^3, each of which includes a sphere with radius a and relative permittivity ε_r. The number of spheres is infinity along the both X and Y axes, and the two-dimensionally infinite periodic structures are stacked to compose finite Q layers in the Z direction. At present we limit ourselves to the case where either electric or magnetic field of the incident plane wave is perpendicular to the page, allowing us to use only one incident angle θ.

In the upper and lower semi-infinite spaces, the approximate wave functions $(\mathbf{E}_{0N}(\mathbf{r}), \mathbf{H}_{0N}(\mathbf{r}))$ and $(\mathbf{E}_{Q+1\,N}(\mathbf{r}), \mathbf{H}_{Q+1\,N}(\mathbf{r}))$ are expressed in terms of modal coefficients $A^{\mathrm{TE,TM}}_{0mn}(N)$ and $A^{\mathrm{TE,TM}}_{Q+1\,mn}(N)$, respectively. The set of modal functions here is the same as that employed in Sect. 8.3.4 for the two-dimensional periodic structures. On the other hand, for the fields in the areas of periodically distributed spheres, a set of vector spherical wave functions $\left\{\mathbf{m}^{\mathrm{e,h}}_{mn}(\mathbf{r}), \mathbf{n}^{\mathrm{e,h}}_{mn}(\mathbf{r})\right\}$ is used to write the approximate wave functions. In the cube region of the layer $\sharp q$, they are expressed by

$$\left\{ \begin{array}{l} \left[\begin{array}{c} \mathbf{E}_{qN}(\mathbf{r}) \\ Z_0\mathbf{H}_{qN}(\mathbf{r}) \end{array} \right] = \displaystyle\sum_{n=1}^{3N} \sum_{m=-n}^{n} \left[\begin{array}{cc} \mathbf{m}^{\mathrm{e}}_{mn}(\mathbf{r}_q) & \mathbf{n}^{\mathrm{e}}_{mn}(\mathbf{r}_q) \\ -\mathrm{i}\mathbf{n}^{\mathrm{h}}_{mn}(\mathbf{r}_q) & -\mathrm{i}\mathbf{m}^{\mathrm{h}}_{mn}(\mathbf{r}_q) \end{array} \right] \left[\begin{array}{c} A^{\mathrm{TE}}_{qmn}(N) \\ A^{\mathrm{TM}}_{qmn}(N) \end{array} \right] \\ \qquad (q = 1, 2, \ldots, Q), \end{array} \right. \tag{8.89}$$

where Z_0 is the intrinsic impedance of vacuum and $\mathbf{r}_q = (r_q, \theta_q, \phi_q)$ is a position vector with its origin placed at the center of the qth sphere on the Z axis. Note that the truncation number is selected as $3N$ in order to maintain the balance with the half spaces from the viewpoint of the degree of approximation. The spherical wave functions $\left\{\mathbf{m}_{mn}^{\mathrm{e,h}}(\mathbf{r}_q), \mathbf{n}_{mn}^{\mathrm{e,h}}(\mathbf{r}_q)\right\}$ are written by combination of the spherical Bessel functions of the nth order, the associated Legendre functions $P_n^{|m|}(\cos\theta_q)$, and the exponential (trigonometric) functions $\exp(\mathrm{i}m\phi_q)$.[31] The functions with respect to $\mathbf{r}_q$ are constructed beforehand so that they automatically satisfy the continuity conditions for E_θ, E_ϕ, H_θ, and H_ϕ over the spherical surfaces $r_q = a$ [17]. As a result, the present problem is reduced to the determination of the modal coefficients such that the remaining boundary conditions on the horizontal planes

$$\begin{cases} \mathbf{u}_Z \times \left(\mathbf{E}_q, \mathbf{H}_q\right) = \mathbf{u}_Z \times \left(\mathbf{E}_{q+1}, \mathbf{H}_{q+1}\right) \\ \qquad \text{(between the layers } \sharp q \text{ and } \sharp q+1;\ q = 0, 1, 2, \ldots, Q) \end{cases} \tag{8.90}$$

and the periodicity conditions on the vertical planes

$$\begin{cases} \mathbf{u}_X \times \left(\mathbf{E}_q, \mathbf{H}_q\right) \exp(\mathrm{i}kd \sin\theta)\big|_{X=-d/2+0} = \mathbf{u}_X \times \left(\mathbf{E}_q, \mathbf{H}_q\right)\big|_{X=d/2-0}, \\ \mathbf{u}_Y \times \left(\mathbf{E}_q, \mathbf{H}_q\right)\big|_{Y=-d/2+0} = \mathbf{u}_Y \times \left(\mathbf{E}_q, \mathbf{H}_q\right)\big|_{Y=d/2-0} \\ \qquad (q = 1, 2, \ldots, Q) \end{cases} \tag{8.91}$$

should be satisfied on the faces of the unit cells in the sense of least-squares. In the boundary conditions (8.91), we count the upper and lower half spaces by the numbers $\sharp 0$ and $\sharp Q + 1$, respectively.

Figure 8.25 shows the normalized mean-square error and energy error as functions of the truncation number N. We find that both errors decrease monotonically when N increases. The period d is 0.8 times as the wavelength of the incident wave λ $(=2\pi/k)$. Since the wavelength in dielectric material is shorter than that in the air, we need large N for big spheres. However, even at $a/d = 0.3$, these errors become less than 1% if $N \geq 4$.

Figure 8.26 is drawn to observe the effect of increasing the layer number on the band of total transmission and total reflection. For the single layer at $Q = 1$, we find two reflection points at $d/\lambda \approx 0.77$ and 0.91. When the layer is increased, these points are changed to reflection bands.

Figure 8.27 presents the reflected power for each mode as a function of incident angle for a 4-layered structure. We observe the power is totally reflected when θ is less than about 40°. This property disappears for larger θ due to the emergence of the $(-1, 0)$th higher order modes having a cutoff angle $\theta = 46°$.

[31] The vector $\mathbf{m}_{mn}^{\mathrm{e,h}}(\mathbf{r}_q)$ is perpendicular to the r_q axis, whereas $\mathbf{n}_{mn}^{\mathrm{e,h}}(\mathbf{r}_q)$ has an r_q component. That is, the superscript TE (TM) in (8.89) means transverse electric (transverse magnetic) with respect to r_q.

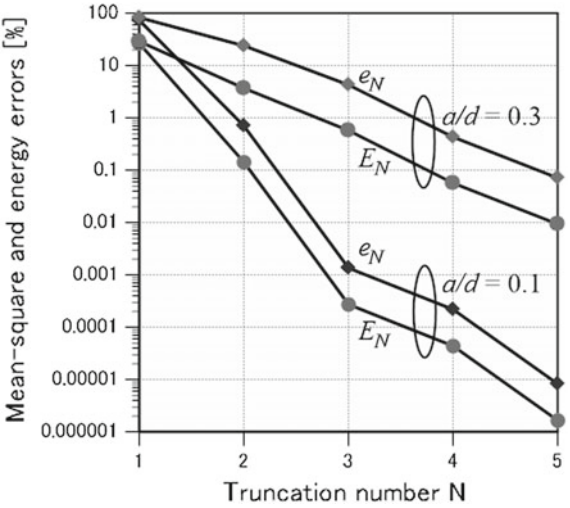

Fig. 8.25 Normalized mean-square error E_N and energy error e_N as functions of truncation number N; $Q = 2$, $\varepsilon_r = 10$, $d/\lambda = 0.8$, and $\theta = 0$

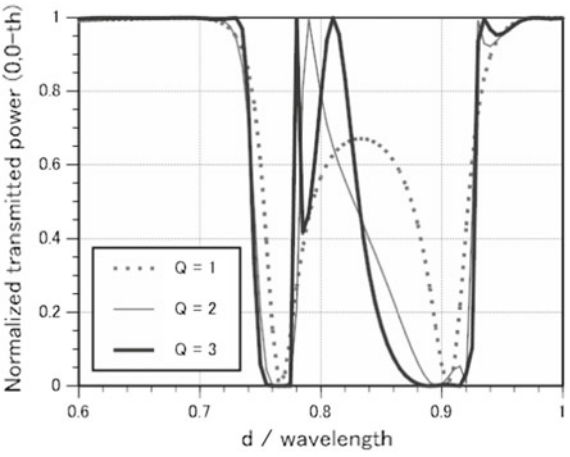

Fig. 8.26 Normalized transmitted powers as functions of wavelength; $a/d = 0.25$, $\varepsilon_r = 12$, and $\theta = 0$

We should note that introduction of sequential accumulation in the process of QR decomposition reduces the computation time from $O(Q^3)$ to $O(Q^1)$ and the memory requirement from $O(Q^2)$ to $O(Q^1)$, with Q being a number of sphere layers. See [17] for the detailed data.

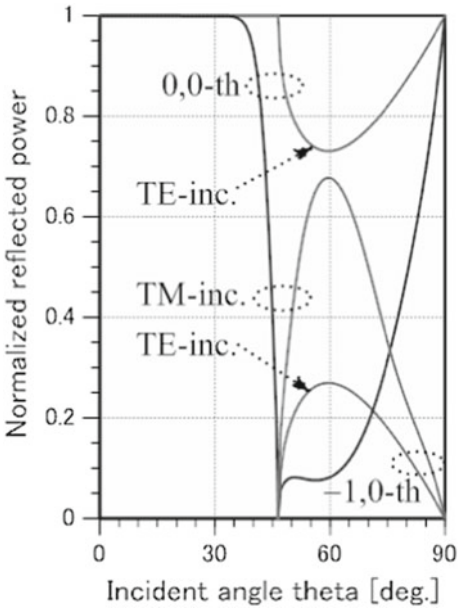

Fig. 8.27 Normalized reflected powers as functions of θ; $Q = 4$, $a/d = 0.25$, $\varepsilon_r = 12$, and $d/\lambda = 0.58$

8.4 Conclusions

Because of the reasons we have stated in Sect. 8.1, we reviewed Yasuura's method of modal expansion attaching importance to the process of solution by the CYM: choice of modal functions; a finite-sum approximate solution; least-squares boundary matching; location and number of sampling points; and solution method for the LSP. In addition, we included guidances for handling dielectric obstacles and gratings placed in planer or conical mounting. Still more, we gave a comparison between separated solutions and monopole fields in approximation power.

As for applications to 3D, we have only two grating problems in Sects. 8.3.4 and 8.3.5. Because we have been working in diffraction gratings, we do not have appropriate examples that show the effectiveness of the CYM in 3D scattering problems. However, our former colleagues have solved the problems using the CYM and published their results [11, 13]. Speaking from a theoretical point of view, they have employed the set of multipole functions as the modal functions whose completeness has been proven by Calderón [6].

We hope that the contents of this chapter would be useful for researchers and engineers who need reliable methods for solving electromagnetic boundary-value problems.

Acknowledgements The authors thank Mr. BenWen Chen and Mr. Rui Gong, Centre for Optical and Electromagnetic Research, South China Academy of Advanced Optoelectronics, South China Normal University for preparing the figures in Sect. 8.3.1 including numerical computations.

One of the authors (A.M.) wish to express his thanks to Japan Society for Promotion of Science (JSPS) for partial support to the work in Sects. 8.3.2 and 8.3.5 under Grant Number JP15K06023 (KAKENHI).

Another one of the authors (Y.O.) is grateful to Prof. S. He, COER-SCNU, COER-ZJU, and JORSEP-KTH for his continuous help and encouragement.

Appendix 1: H-Wave Scattering by a PC Cylinder

Let us consider a problem where an H-wave (TM-wave) is incident to the obstacle shown in Fig. 8.1. That is, the incident wave is polarized in the xy-plane so that the incident magnetic field has only a z-component

$$\mathbf{H}^{\mathrm{i}}(\mathbf{r}) = \mathbf{u}_z F(\mathbf{r}) = \mathbf{u}_z \exp[-\mathrm{i}kr\cos(\theta - \iota)]. \tag{8.92}$$

The scattered magnetic field has only a z-component

$$\mathbf{H}^{\mathrm{s}}(\mathbf{r}) = \mathbf{u}_z \Psi(\mathbf{r}) \tag{8.93}$$

which is a leading field of the problem. Thus, we have

Problem 1': H-wave, PC. Find $\Psi(\mathbf{r})$ that satisfies:

(**N1**) The 2-D Helmholtz equation in $\mathrm{S_e}$;
(**N2**) The 2-D radiation condition at infinity;
(**N3**) The boundary condition

$$\partial_\nu \Psi(s) = g(s) \equiv -\partial_\nu F(s) \quad (s \in \mathrm{C}). \tag{8.94}$$

Here, ∂_ν denotes a normal derivative at s. Equation (8.94) is called Neumann's or the second-kind boundary condition.

Employing the Green's (or Neumann's) function of this boundary-value problem satisfying a homogeneous boundary condition

$$\partial_\nu N(\mathbf{r}, s) = 0 \quad (\mathbf{r} \in \mathrm{S}; s \in \mathrm{C}), \tag{8.95}$$

we get a formal representation similar to (8.13)

$$\Psi_N(\mathbf{r}) - \Psi(\mathbf{r}) = -\int_{s=0}^{C} N(\mathbf{r}, s)\,[\partial_\nu \Psi_N(s) - g(s)]\,ds \quad (\mathbf{r} \in \mathrm{S}). \tag{8.96}$$

Here, Ψ_N denotes an approximate solution defined by

$$\Psi_N(\mathbf{r}) = \sum_{m=-N}^{N} B_m(M)\varphi_m(\mathbf{r}). \tag{8.97}$$

After a discussion similar to that in Sect. 8.2.2.3, we have a least-squares problem for the H-wave problem:

LSM 1': H-wave, PC. Find the coefficients $B_m(M)$ $(m = 0, \pm 1, \ldots, \pm N)$ that minimize the mean-squares boundary residual

$$E_N = \frac{\|\partial_\nu \Psi_N - g\|^2}{\|g\|^2} = \frac{1}{\|g\|^2} \left\| \sum_{m=-N}^{N} B_m(M) \partial_\nu \varphi_m - g \right\|^2. \tag{8.98}$$

We can solve this problem on a computer following the procedure in Sect. 8.2.3. Approximations to other nonzero components can be found by

$$\mathbf{E}_N^{\mathrm{s}}(\mathbf{r}) = \frac{1}{\mathrm{i}\omega\varepsilon_0} \nabla \Psi_N(\mathbf{r}) \times \mathbf{u}_z. \tag{8.99}$$

It is worth noting that in an H-wave scattering from a dielectric obstacle, the boundary condition (8.42) should be altered slightly. Let $\mathbf{H}^{\mathrm{s}}(\mathbf{r}) = \mathbf{u}_z \Psi_{\mathrm{e}}(\mathbf{r})$, and $\mathbf{H}^{\mathrm{t}}(\mathbf{r}) = \mathbf{u}_z \Psi_{\mathrm{i}}(\mathbf{r})$, then we have

$$\begin{cases} \Psi_{\mathrm{e}}(s) - \Psi_{\mathrm{i}}(s) = f(s) \equiv -F(s) \\ \partial_\nu \Psi_{\mathrm{e}} - n^{-2} \partial_\nu \Psi_{\mathrm{i}}(s) = g(s) \equiv -\partial_\nu F(s), \end{cases} \tag{8.100}$$

where the second line means the electric-field continuity and $n^2 = \varepsilon/\varepsilon_0$.

Appendix 2: Solution of LSP 1 by a Normal Equation and Related Topics

Although we do not use a normal equation in numerical analysis, we look over the solution method by the equation because it is an important theoretical tool in working with a least-squares problem. Let us define an inner product between two functions in $\mathbf{H} = L^2(0, C)$ by

$$(f, g) = \int_{s=0}^{C} \overline{f(s)} g(s)\, ds, \tag{8.101}$$

then we find that $\|f\| = \sqrt{(f, f)}$. Employing these relations, we modify (8.22) to obtain

$$E_N = \sum_{m=-N}^{N} \sum_{n=-N}^{N} \overline{A_m}(\varphi_m, \varphi_n) A_n - \sum_{m=-N}^{N} \overline{A_m}(\varphi_m, f) - \sum_{n=-N}^{N} (f, \varphi_n) A_n + \|f\|^2. \tag{8.102}$$

The predictable M is not shown.

Now we define a subspace of **H**, Φ_N, spanned by the boundary values of a finite number of modal functions $\{\varphi_0(s), \varphi_{\pm 1}, \ldots, \varphi_{\pm N}\}$. An element of Φ_N can be represented as

$$\Psi_N(s) = \sum_{n=-N}^{N} A_n \varphi_n(s). \tag{8.103}$$

Apparently, there is a minimum value of E_N, which is a squared distance between $f(s)$ and a point in Φ_N.[32] The minimum is achieved when (8.103) agrees with the foot of a perpendicular line from $f(s)$ to the surface of Φ_N. The necessary and sufficient condition for this is that: The A_m coefficients are the solutions of the set of linear equations

$$\sum_{n=-N}^{N} (\varphi_m, \varphi_n) A_n = (\varphi_m, f) \quad (m = 0, \pm 1, \ldots, \pm N). \tag{8.104}$$

This is referred to as the normal equation (NE) of **LSP 1** and is a formal solution to the problem.[33]

Next, let us consider the minimization from a computational point of view. That is, we try to find the A_m coefficients using the sampled values of boundary functions; and the functions are represented by J-dimensional complex-valued vectors $\mathbf{f}$, $\boldsymbol{\varphi}_m$, and $\boldsymbol{\Psi}_N$ as in Sect. 8.2.3. This leads us to **DLSP 1**. We know the orthogonal decompositions are useful tools for solving the problem. However, setting them aside, we here consider a NE based on **DLSP 1**. Because the Jacobian matrix Φ is $J \times M$ $(J > M)$, the set of linear equations

$$\Phi \mathbf{A} = \mathbf{f} \tag{8.105}$$

is over-determined and does not have a usual solution. However, if we multiply (8.105) by $\Phi^\dagger$ from the left, we have

$$\mathrm{H}\mathbf{A} = \mathbf{b}, \tag{8.106}$$

where

$$\mathrm{H} = \Phi^\dagger \Phi \tag{8.107}$$

is an $M \times M$ positive-definite Hermitian matrix provided Φ is full rank. And,

$$\mathbf{b} = \Phi^\dagger \mathbf{f} \tag{8.108}$$

---

[32] If $f \in \Phi_N$, then $E_N = 0$. This, however, cannot occur in practice: For example, even in the case of scattering from a circular cylinder made of a PC, we need an infinite series to represent a rigorous solution because the boundary value has a form $\exp[-\mathrm{i}ka\cos(\theta_s - \iota)]$. In addition, note that Φ_N is closed.

[33] We get (8.104) by setting $(\varphi_m, \Psi_N - f) = 0$ $(m = 0, \pm 1, \ldots, \pm N)$; or from $\partial E_N / \partial \overline{A}_m = 0$.

is an M-dimensional right-hand side. Usually, (8.106) is referred to as the NE of **DLSP 1** and has been employed as a standard method of solution for a long time.

Obviously, (8.106) is an approximation of (8.104). For example, an (m, n)th element of the coefficient matrix of (8.104) can be represented as

$$(\varphi_m, \varphi_n) = \int_{s=0}^{C} \overline{\varphi_m(s)}\varphi_n(s)\, ds \simeq \frac{C}{J} \sum_{j=1}^{J} \overline{\varphi_m(s_j)}\varphi_n(s_j) = \frac{C}{J}\, \boldsymbol{\varphi}_m^{\dagger}\boldsymbol{\varphi}_n. \tag{8.109}$$

The right-hand side of (8.109) is the (m, n)th entry of H multiplied by the line element C/J. Hence, (8.104) and (8.106) are essentially the same thing, and they have common weak points in numerical computations. Widely-accepted key observations are:

- The NE is rigorous, in principle, and can be employed in theoretical considerations;
- The NE combined with Gaussian elimination (diagonal pivoting assumed) is equivalent to the (modified) Schmidt QRD except for the next two items;
- The NE may lose information in constructing $\mathrm{H} = \boldsymbol{\Phi}^{\dagger}\boldsymbol{\Phi}$, and this process is time consuming usually;
- The NE is dominated by the condition number of H that is square of the original condition number: $\mathrm{cond}(\mathrm{H}) = [\mathrm{cond}(\boldsymbol{\Phi})]^2$.

The last item means (8.104) and (8.106) are more sensitive to computational errors than **LSP 1** and **DLSP 1**. Therefore, the NE's are more difficult to solve on a computer than the original least-squares problems. We, hence, do not recommend the use of (8.104) or (8.106). Even if we are working in the case where the inner products in (8.104) can be calculated analytically, we should not employ (8.104) because of the last item.

Before closing this Appendix, we would like to state a couple of comments on (8.105). Apparently, J cannot be less than M because (8.105) is indeterminate for $J < M$. If we set $J = M$, we have a point-matching method (PMM) or a collocation method. The method is known to be effective if the contour C coincides with a part of a coordinate curve of a system of coordinates in which Helmholtz's equation is separable; and that the modal functions are the separated solutions in that system. Convergence of the PMM solution is related to the validity of the Rayleigh hypothesis [2, 3, 18].

In Yasuura's method we usually set $J = 2M$ as we see in Sect. 8.3.1. That is, we employ $2M$ linear equations to determine M unknown coefficients. This may be understood as a small device or improvement of the PMM. However, this produces good results such as proof of convergence, wide range of application, and so on with little increase of computational complexity as a reasonable cost.

Appendix 3: Conical Diffraction by a Grating

In Sect. 8.2.5 we dealt with diffraction by a grating, where all the field components were functions of two variables (X and Y) and two independent cases of polarization [E-wave (TE, s) and H-wave (TM, p)] existed. In addition, the directions of propagating diffraction-orders were parallel to the plane of incidence. These were possible because: (**1**) the grating surface was uniform in Z; and (**2**) the plane of incidence was in parallel to the direction of periodicity $\mathbf{u}_X$. Here, we concisely examine the problem of a lossless dielectric grating in which the second condition is not satisfied, i.e., the plane of incidence makes a nonzero angle ϕ with the positive X-direction as shown in Fig. 8.14a. We will see that

- The field components are functions of X, Y, and Z, but the dependence on Y—the direction of uniformity—is limited;
- The two cases of polarization are not independent, i.e., both TE and TM diffracted waves exist for TE (or TM) incidence[34];
- The direction of propagating orders lie on the surface of a cone whose vertex agrees with the coordinate origin O; the direction of the zeroth mode is on the plane of incidence at the same time.

Because of the third characteristic, this arrangement ($\phi \neq 0$) is called conical mounting and the term *conical diffraction* is used. In this connection, the arrangement in Sect. 8.2.5 is termed planar mounting.

Let the incident wave be

$$\begin{bmatrix} \mathbf{E}^{\mathrm{i}} \\ \mathbf{H}^{\mathrm{i}} \end{bmatrix} (\mathbf{r}) = \begin{bmatrix} \mathbf{e}^{\mathrm{i}} \\ \mathbf{h}^{\mathrm{i}} \end{bmatrix} \exp(-\mathrm{i}\mathbf{k}^{\mathrm{i}} \cdot \mathbf{r}). \tag{8.110}$$

Here, $\mathbf{e}^{\mathrm{i}}$ and $\mathbf{h}^{\mathrm{i}}$ are electric- and magnetic-field amplitude, which are related by

$$\mathbf{h}^{\mathrm{i}} = \frac{1}{\omega\mu_0} \mathbf{k}^{\mathrm{i}} \times \mathbf{e}^{\mathrm{i}} \tag{8.111}$$

and $\mathbf{k}^{\mathrm{i}}$ is the incident wavevector defined by

$$\mathbf{k}^{\mathrm{i}} = (n_1 k \sin\theta \cos\phi, n_1 k \sin\theta \sin\phi, -n_1 k \cos\theta) \equiv (\alpha, \beta, -\gamma) \tag{8.112}$$

with θ being the polar angle between the wavevector $\mathbf{k}^{\mathrm{i}}$ and the grating normal $\mathbf{u}_Z$. We decompose the incident wave into a TE(s)- and a TM(p)-component, where TE (or TM) means that the electric (or magnetic) field of the relevant incident wave is perpendicular to the plane of incidence. To do this, we define two unit vectors that span a plane orthogonal to the incident wavevector

[34] Assume a PC surface-relief grating with a TE-wave incidence, for simplicity, and imagine the surface current induced. It apparently has a Z-oriented ingredient, which excites a TM-wave component.

$$\mathbf{e}^{\mathrm{TE}} = (\sin\phi, -\cos\phi, 0), \quad \mathbf{e}^{\mathrm{TM}} = (\cos\theta\cos\phi, \cos\theta\sin\phi, \sin\theta). \tag{8.113}$$

They give the directions of the incident electric fields that are in the TE- and TM-polarization.[35] Thus the decomposition is

$$\mathbf{e}^{\mathrm{i}} = \mathbf{e}^{\mathrm{TE}}\cos\delta + \mathbf{e}^{\mathrm{TM}}\sin\delta, \tag{8.114}$$

where δ is a polarization angle shown in Fig. 8.14b. $\delta = 0$ and $\pi/2$ mean TE- and TM-incidence. Hence, an incident wave has three angular parameters: ϕ, θ, and δ.

We consider the problem to seek the diffracted electric and magnetic field in the semi-infinite regions V_1 and V_2 over and below the grating surface S_G.

Problem 4 conical, dielectric grating. Find the solutions that satisfy the following requirements:

(**CD1**) The Helmholtz equation in V_1 and V_2;
(**CD2**) Radiation conditions in the positive and negative Z-direction;
(**CD3**) A periodicity condition that: the relation $f(X+d, Y, Z) = e^{\mathrm{i}\alpha d} f(X, Y, Z)$ holds for any component of the diffracted wave, and the phase constant in Y is β;
(**CD4**) The total tangential component of electric and magnetic field must be continuous across the grating surface S_G.

Dealing with a problem of conical diffraction, we should keep in mind the unique nature of the problem. First, because every field component has a common phase constant β in Y, it is sufficient to match the boundary condition on a cross section between the grating surface and a plane $Y = \text{const}$. The conically-mounted gratings, hence, belong to the class of quasi-3-D structures. Second, because the TE- and TM-wave are not independent, we always need both TE and TM vector modal functions in constructing approximate solutions.

We define the modal functions satisfying (**CD1**)–(**CD3**) by

$$\begin{cases} \boldsymbol{\varphi}^{\mathrm{TE}}_{\ell m}(\mathbf{r}) = \mathbf{e}^{\mathrm{TE}}_{\ell m}\exp(-\mathrm{i}\mathbf{k}_{\ell m}\cdot\mathbf{r}), \quad \boldsymbol{\varphi}^{\mathrm{TM}}_{\ell m}(\mathbf{r}) = \mathbf{e}^{\mathrm{TM}}_{\ell m}\exp(-\mathrm{i}\mathbf{k}_{\ell m}\cdot\mathbf{r}) \\ \qquad (\ell = 1, 2;\ m = 0, \pm1, \pm2, \ldots). \end{cases} \tag{8.115}$$

Here,

$$\mathbf{e}^{\mathrm{TE}}_{\ell m} = \frac{\mathbf{k}_{\ell m}\times\mathbf{u}_Z}{|\mathbf{k}_{\ell m}\times\mathbf{u}_Z|}, \quad \mathbf{e}^{\mathrm{TM}}_{\mathrm{p}m} = \frac{\mathbf{e}^{\mathrm{TE}}_{\ell m}\times\mathbf{k}_{\ell m}}{|\mathbf{e}^{\mathrm{TE}}_{\ell m}\times\mathbf{k}_{\ell m}|} \quad (\ell = 1, 2), \tag{8.116}$$

$$\mathbf{k}_{1m} = (\alpha_m, \beta, \gamma_{1m}), \quad \mathbf{k}_{2m} = (\alpha_m, \beta, -\gamma_{2m}), \tag{8.117}$$

and

$$\begin{cases} \alpha_m = \alpha + \dfrac{2m\pi}{d}, \quad \gamma_{\ell m} = \sqrt{(n_\ell k)^2 - (\alpha_m^2 + \beta^2)} \\ \qquad (\mathrm{Re}\,\gamma_{\ell m} \geq 0,\ \mathrm{Im}\,\gamma_{\ell m} \leq 0). \end{cases} \tag{8.118}$$

[35] $\mathbf{e}^{\mathrm{TE}}$ is perpendicular to the plane of incidence; the fact that the magnetic field accompanying $\mathbf{e}^{\mathrm{TM}}$ is orthogonal to the plane can be seen by manipulation.

Note that the functions in (8.115) are for constructing electric fields. For the magnetic fields we get

$$\boldsymbol{\psi}^{\mathrm{q}}_{\ell m}(\mathbf{r}) = \frac{1}{\omega\mu_0}\,\mathbf{k}_{\ell m} \times \boldsymbol{\varphi}^{\mathrm{q}}_{\ell m}(\mathbf{r}) \quad (\ell = 1, 2;\ \mathrm{q} = \mathrm{TE}, \mathrm{TM}) \tag{8.119}$$

through Maxwell's equations. Finite linear combinations of the modal functions define approximate solutions:

$$\begin{bmatrix} \mathbf{E}_{\ell N} \\ \mathbf{H}_{\ell N} \end{bmatrix}(\mathbf{r}) = \sum_{m=-N}^{N} A^{\mathrm{TE}}_{\ell m} \begin{bmatrix} \boldsymbol{\varphi}^{\mathrm{TE}}_{\ell m} \\ \boldsymbol{\psi}^{\mathrm{TE}}_{\ell m} \end{bmatrix}(\mathbf{r}) + \sum_{m=-N}^{N} A^{\mathrm{TM}}_{\ell m} \begin{bmatrix} \boldsymbol{\varphi}^{\mathrm{TM}}_{\ell m} \\ \boldsymbol{\psi}^{\mathrm{TM}}_{\ell m} \end{bmatrix}(\mathbf{r}) \quad (\ell = 1, 2) \tag{8.120}$$

Here, the number of modal functions M is neglected.

The unknown coefficients in (8.120) should be determined in order that the solutions satisfy the boundary condition (**CD4**) approximately in the mean-squares sense. For this purpose we first consider the cross section C between the grating surface $\mathrm{S_G}$ and a plane $Y = 0$. This is the same thing as the periodic curve C in Sect. 8.2.5. In a similar way to one in Sect. 8.2.5, we define the primary period S_1, whose boundary C_1 ($\subset$ C), the function space $\mathbf{H}$ consisting of all the square integrable functions on C_1, and the norm $\|f\|$ of a function $f(s)$. Then, we can state the least-squares problem that determines the unknown coefficients:

LSP 4: conical, dielectric grating. Find the coefficients $A^{\mathrm{TM}}_{\ell m}$ and $A^{\mathrm{TM}}_{\ell m}$ ($\ell = 1, 2;\ m = 0, \pm 1, \ldots, \pm N$) that minimize the mean-square error

$$E_N = \left\| \boldsymbol{\nu} \times \left(\tilde{\mathbf{E}}_{1N} + \tilde{\mathbf{E}}^{\mathrm{i}} - \tilde{\mathbf{E}}_{2N} \right) \right\|^2 + Z_0^2 \left\| \boldsymbol{\nu} \times \left(\tilde{\mathbf{H}}_{1N} + \tilde{\mathbf{H}}^{\mathrm{i}} - \tilde{\mathbf{H}}_{2N} \right) \right\|^2. \tag{8.121}$$

Here, Z_0 denotes the intrinsic impedance of vacuum and $\tilde{\mathbf{E}}^{\mathrm{i}}$ etc. mean periodic functions with respect to x defined in the same way as one in (8.69)–(8.71). The method of discretization and the solution method are found in Sect. 8.2.4.

Appendix 4: Comparison of Modal Functions and Algorithm of the SP

Here we show some results of effectiveness comparison between three kinds of modal functions in solving a sample problem[36]: E-wave scattering from a PC cylinder whose

[36] We can use monopole fields also in the grating problems discussed in Sect. 8.2.5. A countably infinite set of monopoles located periodically in x—i.e., the location is given by $(x_1 + \ell d, y_1)$ $(0 < x_1 \le d;\ y_1 < \eta(x_1);\ \ell = 0, \pm 1, \pm 2, \ldots)$—radiates a plane wave [4, 36] satisfying (**GD1**) and (**GD2**). If we let the monopoles be accompanied by phase factors $\exp(\mathrm{i}\ell kd \sin\theta)$, the plane wave meets the periodicity (**GD3**). Increasing the number of monopoles in the first strip region to M, i.e., $(x, y) = (x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_M, y_M)$, and repeating the same procedure, we have a set of M plane waves, which is the desired set of modal functions [28, 37].

cross section is given by[37]

$$\mathrm{C} : r_s = a(1 + 0.2\cos 3\theta_s). \tag{8.122}$$

Let us normalize every quantity having dimension of length by the total length of C. And, we assume the incident wave comes along the x-axis from the negative x-direction (i.e., $\iota = 0$).

The modal functions considered here are: (**a**) the separated solutions, which we defined by (8.7) in Sect. 8.2.2; (**b**) monopole fields defined by (8.9); (**c**) monopole fields whose poles are located densely near the convex part of C. Because the separated solutions are known widely, we explain the monopole fields below:

(**b**) Equally spaced poles. Let L be a similar curve to C with the ratio of similitude d $(0 < d < 1)$.[38] We arrange M poles on L at regular intervals. Then, the distance between two poles is L/M where L is the length of L.

(**c**) Concentration of poles near the convex parts of L. (**i**) First, we draw the similar curve L. (**ii**) Next,we calculate the curvature $\kappa(t)$ of L as a function of t $(\in \mathrm{L})$, and add some positive bias c in order that the biased curvature (BC) be no less than 0: $\tilde{\kappa}(t) = \kappa(t) + c$ (≥ 0). (**iii**) Thirdly, we define a probability density function by normalizing the BC.[39]

$$f(t) = \frac{\displaystyle\int_0^t \tilde{\kappa}(t')\,dt'}{\displaystyle\int_0^L \tilde{\kappa}(t')\,dt'}. \tag{8.123}$$

Thus we get the number of poles between t_1 and t_2 by

$$n(t_1, t_2) = M \int_{t_1}^{t_2} f(t)\,dt. \tag{8.124}$$

We have solved the problem using the method explained in Sects. 8.2.2 and 8.2.3. We used three kinds of modal functions (**a**), (**b**), and (**c**); and tried at two frequencies: $ka = 10$ and 30. The parameter d was set to be 0.87. To see the accuracy of a solution we calculated two kinds of errors: the normalized mean-square error $E_M(\mathrm{m})$ and

[37] Although the employment of polyphase wave functions is effective because of the periodicity, we do not use them for simplicity.

[38] According to the result of numerical computation, an optimum d was in the rage $[0.85, 0.90]$ when the total number of poles was between 40 and 120. If we increased (or decreased) the number of poles, the optimum d approached 0.90 (or 0.85). Note, however, that the trends were observed in solving a particular problem with specific computational parameters and are no more than reference data.

[39] The result of sample calculation has shown that the use of $|\mathrm{BC}|^\alpha$ $(\alpha > 1)$ instead of BC (i.e., further emphasis of the convex part in locating poles) gives better solutions.

the error on energy balance (or on the optical theorem) $e_M(\mathrm{m})$. The former is the same thing as one defined in (8.22) and (8.30)[40] except that the subscript shows the total number of modal functions. The latter shows the deviation from a proportional relation between the forward scattering amplitude and total cross section.[41] The argument m shows the type of modal functions: m = sep, esp, and pcc, which mean (**a**) separated solutions, (**b**) equally-spaced poles, and (**c**) poles concentrated near the convex parts.

Results at $ka = 10$. Because the obstacle size is handy, the E_M errors fall off rapidly: $E_{45}(\mathrm{sep})$, $E_{35}(\mathrm{esp})$, and $E_{31}(\mathrm{pcc})$ are below 1%. As for the e_M errors of the solutions, the situation is different. The solutions with esp or pcc modal functions converge rapidly as $e_M(\mathrm{esp})$ and $e_M(\mathrm{pcc})$ are below 1% at $M \simeq 30$. On the other hand, $e_{31}(\mathrm{sep})$ is about 10%. Increasing M to 70, we have: $e_{70}(\mathrm{esp}) = 9 \times 10^{-5}\%$; $e_{70}(\mathrm{pcc}) = 1 \times 10^{-5}\%$; and $e_{71}(\mathrm{sep}) = 4\%$.

Results at $ka = 30$. The advantage of the monopole fields is clear in this range of frequency. Setting $M \simeq 100$, we have $E_{101}(\mathrm{sep}) = 4\%$, $E_{100}(\mathrm{esp}) = 2 \times 10^{-1}\%$, $E_{100}(\mathrm{pcc}) = 2 \times 10^{-3}\%$, $e_{101}(\mathrm{sep}) = 7\%$, $e_{100}(\mathrm{esp}) = 2 \times 10^{-1}\%$, and $e_{100}(\mathrm{pcc}) = 5 \times 10^{-3}\%$. The pcc modal functions seems to be the best choice in solving the problem. In fact, we can find an accurate solution with a $10^{-5}\%$ e_M error by setting m = pcc and $M = 120$.

These results mean that the potential of a combination of separated solutions is not so strong in describing scattered fields from obstacles deformed strongly from a circle. We have two ways to cope with this issue: (**i**) employment of a set of modal functions other than the separated solutions[42]; and (**ii**) employment of the SP.

The Algorithm of the SP

Here we include a guidance how to apply the SP in the boundary-matching process based on Yasuura's method of modal expansion for convenience. We start from **DLSP 1**, i.e., minimization of the numerator of (8.30), $\|\Phi\mathbf{A} - \mathbf{f}\|^2$. Instead of minimizing it directly, we force a constraint

$$(\mathbf{1}, \Phi\mathbf{A} - \mathbf{f}) = 0 \tag{8.125}$$

on the M-dimensional solution vector $\mathbf{A}$, where the parentheses mean an inner product and $\mathbf{1} = [1\ 1\ \cdots\ 1]^{\mathrm{T}}$ is a J-dimensional constant vector.

An operator of the smoothing procedure (in a discretized form) is a $J \times J$ matrix given by

$$\mathrm{K}^{(p)} = \left[K_{j\ell}^{(p)}\right], \tag{8.126}$$

[40] We have applied the rule $J = 2M$ and have omitted J.

[41] The relation is referred to as the optical theorem, which implies energy conservation.

[42] We have employed the monopole fields and have seen their effectiveness [30]. It is worth noting that inclusion of a few dipoles located near the convex part of L in addition to the monopoles improves the efficiency greatly. This might be related to Cadilhac-Petit's opinion [7] in locating the poles near an internal focus.

where p means the order of the SP. The explicit forms of the matrix elements for $p = 1, 2$, and 3 are[43]:

$$\begin{cases} K^{(1)}_{j\ell} = u(j-\ell) - \dfrac{j-\ell}{J} - \dfrac{1}{2}, \\ K^{(2)}_{j\ell} = -\dfrac{1}{2}\left[\dfrac{(j-\ell)^2}{J^2} - \dfrac{|j-\ell|}{J} + \dfrac{1}{6}\right], \\ K^{(3)}_{j\ell} = \dfrac{1}{6}\left[\dfrac{|j-\ell|^3}{J^3} - \dfrac{3(j-\ell)^2}{2J^2} + \dfrac{|j-\ell|}{2J} + \dfrac{1}{6}\right]. \end{cases} \tag{8.127}$$

Thus we can state a method of solution with the SP as follows:

DLSP 3: E-wave, PC, SP. Find the solution vector **A** that minimizes the discretized mean-square error

$$E_{MJ} = \frac{\|\mathrm{K}^{(p)}(\Phi\mathbf{A} - \mathbf{f})\|^2}{\|\mathrm{K}^{(p)}\mathbf{f}\|^2} \tag{8.128}$$

under the constraint (8.125).

Two ways are possible to solve this conditioned least-squares problem: (**i**) employment of Lagrange's multiplier; and (**ii**) elimination of a modal coefficient by using the constraint. Although (**i**) is a standard way in handling a constraint, we take (**ii**) because our constraint is simple and can eliminate one of the M unknowns to deduce a least-squares problem with $M - 1$ unknowns.

References

1. M. Bass (ed.), *Handbook of Optics; Volume II — Devices, Measurements, and Properties*, 2nd edn. (McGraw-Hill, 1995)
2. R.H.T. Bates, Analytic constraints on electromagnetic field computations. IEEE Trans. Microw. Theory Tech. **MTT-23**(8), 605–623 (1975)
3. R.H.T. Bates, J.R. James, I.N.L. Gallett, R.F. Millar, An overview of point matching. Radio Electron. Eng. **43**(3), 193–200 (1973)
4. A. Boag, Y. Leviatan, A. Boag, Analysis of two-dimensional electromagnetic scattering from a periodic grating of cylinders using a hybrid current method. Radio Sci. **23**(4), 612–624 (1988)
5. G.P. Bryan-Brown, J.R. Sambles, M.C. Hutley, Polarization conversion through the excitation of surface plasmons on a metallic grating. J. Modern Opt. **37**(7), 1227–1232 (1990)
6. A.P. Calderón, The multipole expansion of radiation fields. J. Ration. Mech. Anal. (J. Math. Mech.) **3**, 523–537 (1954)
7. M. Cadilhac, R. Petit, On the diffraction problem in electromagnetic theory: a discussion based on concepts of functional analysis including an example of practical application, in *Huygens'*

[43] We got the elements under the assumption that the length of C is 1. This is convenient in mathematical analysis and does not affect applications to obstacles made of a lossless material including PC. In dealing with a lossy material, in particular a metal in light frequency, the normalization should be accompanied by a law of similitude in time-dependent EM field [19] and, hence, the use of actual length might be appropriate.

Principle 1690–1990: Theory and Applications, Studies in Mathematical Physics, ed. by H. Blok, et al. (Elsevier, Amsterdam, 1992)
8. G. Hass, L. Hardley, Optical properties of metal, in *American Institute of Physics Handbook*, ed. by D.E. Gray, 2nd ed. (McGraw-Hill, 1963), pp. 6–107
9. J.P. Hugonin, R. Petit, M. Cadilhac, Plane-wave expansions used to describe the field diffracted by a grating. J. Opt. Soc. Am. **71**(5), 593–598 (1981)
10. H. Ikuno, K. Yasuura, Numerical calculation of the scattered field from a periodic deformed cylinder using the smoothing process on the mode-matching method. Radio Sci. **13**(6), 937–946 (1978)
11. H. Ikuno, M. Gondoh, M. Nishimoto, Numerical analysis of electromagnetic wave scattering from an indented body of revolution. Trans. IEICE Electron. **E74-C**(9), 2855–2863 (1991)
12. T. Inagaki, J.P. Goudonnet, J.W. Little, E.T. Arakawa, Photoacoustic study of plasmon-resonance absorption in a bigrating. J. Opt. Soc. Am. B **2**(3), 433–439 (1985)
13. M. Kawano, H. Ikuno, M. Nishimoto, Numerical analysis of 3-D scattering problems using the Yasuura method. Trans. IEICE Electron. **E79-C**(10), 1358–1363 (1996)
14. A.N. Kolmogorov, S.V. Fomin, *Elements of the Theory of Functions and Functional Analysis* (Dover, New York, 1999)
15. C.L. Lawson, R.J. Hanson, *Solving Least Squares Problems* (Prentice-Hall, New Jersey, 1974)
16. T. Matsuda, D. Zhou, Y. Okuno, Numerical analysis of plasmon-resonance absorption in a bisinusoidal metal grating. J. Opt. Soc. Am. A **19**(4), 695–701 (2002)
17. A. Matsushima, Y. Momoka, M. Ohtsu, Y. Okuno, Efficient numerical approach to electromagnetic scattering from three-dimensional periodic array of dielectric spheres using sequential accumulation. Progr. Electromagn. Res. **69**, 305–322 (2007)
18. R.F. Millar, Rayleigh hypothesis and a related least-squares solution to scattering problems for periodic surfaces and other scatterers. Radio Sci. **8**(8–9), 785–796 (1973)
19. H. Nakano, Frequency-independent antennas: spirals and log-periodics, in *Modern Antenna Handbook*, ed. by C.A. Balanis (Wiley, New Jersey, 2008), pp. 263–323
20. Y. Nakata, M. Koshiba, M. Suzuki, Finite-element analysis of plane wave diffraction from dielectric gratings. Trans. IEICE Jpn. **J69-C**(12), 1503–1511 (1986)
21. M. Neviér, The homogeneous problems, in *Electromagnetic Theory of Gratings*, ed. by R. Petit (Springer, Berlin, 1980), pp. 123–157
22. M. Ohtsu, Y. Okuno, A. Matsushima, T. Suyama, A Combination of up- and down-going Floquet modal functions used to describe the field inside grooves of a deep grating. Progr. Electromagn. Res. **64**, 293–316 (2006)
23. Y. Okuno, A numerical method for solving edge-type scattering problems. Radio Sci. **22**(6), 941–946 (1987)
24. Y. Okuno, The mode-matching method, in *Analysis Methods in Electromagnetic Wave Problems*, ed. by E. Yamashita (Artech House, 1990), pp. 107–138
25. Y. Okuno, An introduction to the Yasuura method, in *Analytical and Numerical Methods in Electromagnetic Wave Theory*, ed. by M. Hashimoto, M. Idemen, O.A. Tretyakov (Science House, 1993), pp. 515–565
26. Y. Okuno, H. Ikuno, Completeness of the boundary values of equivalent sources. Mem. Fac. Eng. Kumamoto Univ. **38**(1), 1–8 (1993)
27. Y. Okuno, H. Ikuno, Yasuura's method, its relation to the fictitious source methods, and its advancements in the solution of 2D problems, in *Generalized Multipole Techniques for Electromagnetic and Light Scattering*, ed. T. Wriedt (Elsevier, Amsterdam, 1999)
28. Y. Okuno, T. Matsuda, T. Kuroki, Diffraction efficiency of a grating with deep grooves, in *Proceedings of the 1995 Sino-Japanese Joint Meeting on Optical Fiber Science and Electromagnetic Theory (OFSET'95)*, vol. 1 (Tianjin, China, 1995), pp. 106–111
29. Y. Okuno, T. Suyama, R. Hu, S. He, T. Matsuda, Excitation of surface plasmons on a metal grating and its application to an index sensor. Trans. IEICE Electron. **E90-C**(7), 1507–1514 (2007)
30. Y. Okuno, H. Yamaguchi, The idea of equivalent sources in the Yasuura method, in *Proceedings 1992 International Symposium on Antennas Propagat (ISAP'92)*, vol. 1E3-2 (Sapporo, Japan, 1992)

31. Y. Okuno, K. Yasuura, Numerical algorithm based on the mode-matching method with a singular-smoothing procedure for analysing edge-type scattering problems. IEEE Trans. Antennas Propagat. **30**(4), 580–587 (1982)
32. R. Petit (ed.), *Electromagnetic Theory of Gratings* (Springer, Berlin, 1980)
33. R. Petit, M. Cadilhac, Electromagnetic theory of gratings: some advances and some comments on the use of the operator formalism. J. Opt. Soc. Am. A **7**(9), 1666–1674 (1990)
34. H. Raether, Surafce plasmon and roughness, in *Surface Polaritons — Electromagnetic Waves at Surfaces and Interfaces*, ed. by V.M. Agranovich, D.L. Mills (North Holland, 1982), pp. 331–403
35. M. Tomita, K. Yasuura, The Rayleigh expansion theorem for the boundary value problem in two media. Kyushu Univ. Tech. Rep. **52**(2), 142–154 (1979)
36. J.R. Wait, Reflection from a wire grid parallel to a conducting plane. Can. J. Phys. **32**, 571–579 (1954)
37. X. Xu, B.W. Chen, R. Gong, M. Zheng, Use of auxiliary source fields in Yasuura's method, in *Proceedings of the 2017 IEEE International Conference on Computational Electromagnetics (ICCEM2017)*, vol. 2C1.2 (Kumamoto, Japan, 2017)
38. K. Yasuura, T. Itakura, Approximation method for wave functions (I). Kyushu Univ. Tech. Rep. **38**(1), 72–77 (1965)
39. K. Yasuura, T. Itakura, Complete set of wave functions – approximation method for wave functions (II). Kyushu Univ. Tech. Rep. **38**(4), 378–385 (1966)
40. K. Yasuura, T. Itakura, Approximation algorithm by complete set of wave functions – approximation method for wave functions (III). Kyushu Univ. Tech. Rep. **39**(1), 51–56 (1966)
41. K. Yasuura, H. Ikuno, Smoothing process on the mode-matching method for solving two-dimensional scattering problems. Mem. Fac. Eng. Kyushu Univ. **37**(4), 175–192 (1977)
42. K. Yasuura, Y. Okuno, Singular-smoothing procedure on Fourier analysis. Mem. Fac. Eng. Kyushu Univ. **41**(2), 123–141 (1981)
43. K. Yasuura, M. Tomita, Convergency of approximate wave functions on the boundary – the case of inner domain. Kyushu Univ. Tech. Rep. **52**(1), 79–86 (1979)
44. K. Yasuura, M. Tomita, Convergency of approximate wave functions on the boundary – the case of outer domain. Kyushu Univ. Tech. Rep. **52**(1), 87–93 (1979)
45. K. Yasuura, M. Tomita, Numerical analysis of plane wave scattering from dielectric cylinders. Trans. IECE Jpn. **62-B**(2), 132–139 (1979)
46. K.A. Zaki, A.R. Neureuther, Scattering from a perfectly conducting surface with a sinusoidal height profile: TE polarization. IEEE Trans. Antennas Propagat. **AP-19**(2), 208–214 (1971)