Electronic Transactions on Numerical Analysis.
Volume 31, pp. 331-357, 2008.

ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

THE RCWA METHOD - A CASE STUDY WITH OPEN QUESTIONS AND PERSPECTIVES OF ALGEBRAIC COMPUTATIONS*

JOHN J. HENCH† AND ZDENĚK STRAKOŠ‡

Abstract. Diffraction of light on periodic media represents an important problem with numerous physical and engineering applications. The Rigorous Coupled Wave Analysis (RCWA) method assumes a specific form of gratings which enables a straightforward separation of space variables. Using Fourier expansions, the solutions of the resulting systems of ordinary differential equations for the Fourier amplitudes can be written, after truncation, in form of matrix functions, with an elegant formulation of the linear algebraic problem for integrating constants. In this paper, we present a derivation of the RCWA method, formulate open questions which still need to be addressed, and discuss perspectives of efficient solution of the related highly structured linear algebraic problems. A detailed understanding of the RCWA method for the two-dimensional grating is, in our opinion, necessary for the development of a successful generalization of the method to practical problems.

Key words. Diffraction of electromagnetic waves, Maxwell's equations, periodic gratings, RCWA, truncated Fourier expansions, matrix functions, structured matrices, scattering amplitude.

AMS subject classifications. 78A45, 42A20, 42A85, 35Q60, 65L10, 65F10, 65F30.

1. Introduction. There are many methods for the numerical modeling of the diffraction of electromagnetic waves on periodic gratings. Among those, a specific role is played by the so called Rigorous Coupled Wave Analysis (RCWA) method, which in its most basic two-dimensional form assumes very simple rectangular gratings. The history of the RCWA and related methods is given in the standard monograph [9], together with the description of fundamentals of the differential theory of gratings and several generalizations that can be applied to solving practical problems, see also the corresponding parts and references in [2, 6, 7, 8].

The simple rectangular form of a grating allows in RCWA an easy separation of space variables, and, using Fourier expansions for the space periodic part of the solution, a transformation of the problem described by the partial differential equations into the system of ordinary differential equations (ODE) for the Fourier amplitudes. In order to solve the problem numerically, the infinite dimensional continuous problem must be discretized. In RCWA this entails the truncation of the Fourier expansions, followed by a derivation of the finite dimensional representation of the problem. The solution of the resulting ODEs can be written in the form of elementary matrix functions with an elegant matrix formulation of the linear algebraic problems for the integrating constants.

Obviously, one should ask whether the solution of the discretized problem approximates to sufficient accuracy (in an appropriate sense) the solution of the original problem, which requires mathematical justification by rigorous analysis. The derivation of RCWA assumes smooth functions. In order to get a good agreement of the computed results with the physical reality in modeling the idealized surfaces of discontinuity (see [13, Chapter 9]), it is necessary to use truncation rules which *in the limit* remain valid in the presence of discontinuities. A step in this direction was taken by Li [6, 7], who proved convergence results for a particular truncation of the multiplied Fourier expansions, which led to the so called fast Fourier meth-

*Received January 31, 2008. Accepted March 2, 2009. Published online on September 18, 2009. Recommended by Oliver Ernst.

†KLA-Tencor Corporation, 160 Rio Robles, San Jose, CA 95134, U.S.A. (John.Hench@kla-tencor.com).

‡Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 18207 Prague, Czech Republic (strakos@cs.cas.cz). The work of this author was supported by the Institutional Research Plan AV0Z10300504, by the project IAA100300802 of the GAAS and by the donation of the KLA-Tencor in support of the basic research in the Institute of Computer Science AS CR.

ods[1], for their good performance in practical computations. What is even more important, Li proved that the discretizations that led to slow numerical computations are incorrect and the related discretization errors are responsible for the poor performance of the whole method observed in practice. This gives an illustrative example of a mathematical theory which not only justifies the intuitively derived results, but which also shows that intuition can, in an unfortunate case, mislead in the derivation of methods and algorithms in scientific computing. Without proper mathematical proofs to justify the choice of discretizations, wrong intuitive arguments can lead to algorithmic variations which are inefficient and inaccurate, wasting time and effort. Although the RCWA method has been used in practical computations for more than a decade, its mathematical justification has not been fully completed yet.

In our text, we present a derivation of the RCWA method for a simple two-dimensional rectangular grating. The simplicity of the grating model allows us to see more clearly the interconnections between the physical model with its assumptions, separation of variables, discretization, formulation of the algebraic problem and, finally, possible approaches to its efficient numerical solution. This is, in our opinion, necessary to identify the issues which have to be resolved in order to develop further efficient generalizations of the RCWA method, with some directions given, e.g., in [2, 9]. The RCWA approach is rich in mathematical problems from many disciplines, including numerical linear algebra, and building an efficient RCWA-based solver for practical problems will require a well-balanced solution of all of them.

The paper has a simple structure. After application of the basic theory of planar electromagnetic waves to our model problem in Section 2, we give in the subsequent structured Section 3 a step by step derivation of the RCWA method. Section 4 reviews some remaining open problems. The paper ends by discussing possible approaches for the efficient solution of linear algebraic problems resulting from the RCWA discretization.

2. Plane electromagnetic waves. We will start with Maxwell's equations of electrodynamics for a material with no free charges (see, e.g., [13, Section 21-2, (21-19)-(21-22)])

$$\begin{aligned} \operatorname{div} \widehat{\mathbf{D}} &= 0, & \operatorname{div} \widehat{\mathbf{B}} &= 0, \\ \operatorname{curl} \widehat{\mathbf{E}} &= -\frac{\partial \widehat{\mathbf{B}}}{\partial t}, & \operatorname{curl} \widehat{\mathbf{H}} &= \frac{\partial \widehat{\mathbf{D}}}{\partial t} + \widehat{\mathbf{J}}, \end{aligned} \tag{2.1}$$

where $\widehat{\mathbf{D}}$, $\widehat{\mathbf{E}}$, $\widehat{\mathbf{B}}$, $\widehat{\mathbf{H}}$ are the vectors of the displacement field, electric field, induction field and magnetic field, respectively, and $\widehat{\mathbf{J}}$ represents the free current. Throughout the paper we will consider linear isotropic materials for which the constitutive equations

$$\widehat{\mathbf{D}} = \varepsilon\, \widehat{\mathbf{E}}, \quad \widehat{\mathbf{B}} = \mu\, \widehat{\mathbf{H}}$$

hold. Moreover, the material will be considered magnetically homogeneous with $\mu = \mu_0$, where μ_0 is the magnetic permeability in vacuum. The electric permittivity ε will in general be considered space dependent, $\varepsilon = \varepsilon_0 \varepsilon_r$, where ε_0 is the electric permittivity in vacuum, $(\varepsilon_0 \mu_0)^{-1} = c^2$, and c is the speed of light in vacuum. Under these assumptions, (2.1) takes the form (see [13, Exercise 21-7, p. 362]),

$$\begin{aligned} \operatorname{div} \widehat{\mathbf{E}} &= -\widehat{\mathbf{E}} \cdot \frac{\operatorname{grad} \varepsilon}{\varepsilon}, & \operatorname{div} \widehat{\mathbf{H}} &= 0, \\ \operatorname{curl} \widehat{\mathbf{E}} &= -\mu \frac{\partial \widehat{\mathbf{H}}}{\partial t}, & \operatorname{curl} \widehat{\mathbf{H}} &= \varepsilon \frac{\partial \widehat{\mathbf{E}}}{\partial t} + \sigma \widehat{\mathbf{E}}, \end{aligned} \tag{2.2}$$

[1] This term is used in the optical engineering and physics literature. Since there is no relationship between the Fast Fourier Transform and the fast Fourier methods, the latter term being for mathematically oriented community rather confusing, we will avoid the appellation "fast Fourier methods" altogether.

where $\sigma\widehat{\mathbf{E}} \equiv \widehat{\mathbf{J}}$ accounts for the electric current caused by the electric field in the conductive material with conductivity σ in accordance with Ohm's law. Taking the curl of the last two equations in (2.2) and using simple vector calculus yields, under standard smoothness assumptions,

$$\Delta\widehat{\mathbf{E}} = \varepsilon\mu\frac{\partial^2\widehat{\mathbf{E}}}{\partial t^2} + \sigma\mu\frac{\partial\widehat{\mathbf{E}}}{\partial t} - \operatorname{grad}\left(\widehat{\mathbf{E}}\cdot\frac{\operatorname{grad}\varepsilon}{\varepsilon}\right), \tag{2.3}$$

$$\Delta\widehat{\mathbf{H}} = \varepsilon\mu\frac{\partial^2\widehat{\mathbf{H}}}{\partial t^2} + \sigma\mu\frac{\partial\widehat{\mathbf{H}}}{\partial t} - \operatorname{grad}\varepsilon \times \frac{\partial\widehat{\mathbf{E}}}{\partial t} - \operatorname{grad}\sigma \times \widehat{\mathbf{E}}. \tag{2.4}$$

REMARK 2.1. Except for the relationship between the space dependent vectors of electric and magnetic fields in Subsection 2.3, we will consider in the rest of Section 2 nonconductive materials, i.e., $\sigma = 0$. Then, the index of refraction of the materials is real (and positive), which simplifies the exposition. For conductive materials the derivation is analogous. The resulting individual equations for the electric and magnetic fields for lossless nonconductive materials, as they will be used in the description of the RCWA method to follow, are *formally* identical to the materials with losses due to their nonzero conductivity. The only difference is that in the latter case the index of refraction is complex, with positive real and nonnegative imaginary parts.

In a homogeneous material with losses, the real part of the index of refraction is used for the parametric description of propagating waves similarly as in lossless materials. A nonzero imaginary part describes the damping of the propagating field due to losses. Other differences are unimportant in the context of this text. For an instructive description of the theory of electromagnetic waves, including plane waves in conductive media and the use of a complex index of refraction, we refer to the basic textbook [13], in particular to Section 24.3.

2.1. Time-harmonic fields. We will consider only time-harmonic fields, where any field vector $\widehat{\mathbf{V}}(x, y, z, t)$ will be represented by its associated space dependent complex vector $\mathbf{V}(x, y, z)$ such that

$$\widehat{\mathbf{V}}(x, y, z, t) = \operatorname{Re}[\mathbf{V}(x, y, z)\exp(-\mathbf{i}\omega t)]; \tag{2.5}$$

see [13] and [9, Section I.2.1]. Here $\omega = 2\pi f$, $f\lambda = v$, therefore $\omega = 2\pi v\lambda^{-1}$, where λ is the wavelength, f the frequency of light, and v is the speed of light corresponding to the electric permittivity and the magnetic permeability. If the electric permittivity and the magnetic permeability are constant and $\sigma = 0$, (2.3)-(2.4) reduce to the wave equations for the electric and magnetic field in linear lossless isotropic homogeneous media, which gives

$$v = \frac{1}{\sqrt{\varepsilon\mu}} = \frac{1}{\sqrt{\varepsilon_r\mu_r}}\,\frac{1}{\sqrt{\varepsilon_0\mu_0}} = \frac{c}{n}, \quad n = \sqrt{\varepsilon_r\mu_r}, \quad c = \frac{1}{\sqrt{\varepsilon_0\mu_0}},$$

where n is the index of refraction of the given material.

Here we only consider what is called linear optics, where the time-harmonic setting is relevant and there are no time-frequency conversions, so that the different wavelengths may be treated independently of each other. In such a setting, (2.3)-(2.4) for the space-dependent vector fields take the form (recall $\sigma = 0$)

$$\Delta\mathbf{E} = -\varepsilon\mu\omega^2\mathbf{E} - \operatorname{grad}\left(\mathbf{E}\cdot\frac{\operatorname{grad}\varepsilon}{\varepsilon}\right), \tag{2.6}$$

$$\Delta\mathbf{H} = -\varepsilon\mu\omega^2\mathbf{H} - \frac{1}{\varepsilon}\operatorname{grad}\varepsilon \times \operatorname{curl}\mathbf{H}. \tag{2.7}$$

ETNA
Kent State University
http://etna.math.kent.edu

2.2. Plane waves, TE and TM polarization. We will consider a plane wave solution to Maxwell's equations. For a plane wave whose wave-front is moving in direction $\mathcal{D}$ the vectors $\mathbf{E}$, $\mathbf{H}$ and $\mathcal{D}$ form a right-handed orthogonal system, where $\mathbf{E}$ and $\mathbf{H}$ form a plane (wavefront) perpendicular to the direction $\mathcal{D}$. This paper considers planar diffraction on rectangular gratings in the x-z plane depicted in Figure 2.1, where the incident plane wave is moving in the direction $\mathcal{D}$ perpendicular to the third Cartesian coordinate y, with the angle θ between $\mathcal{D}$ and the vertical direction z.

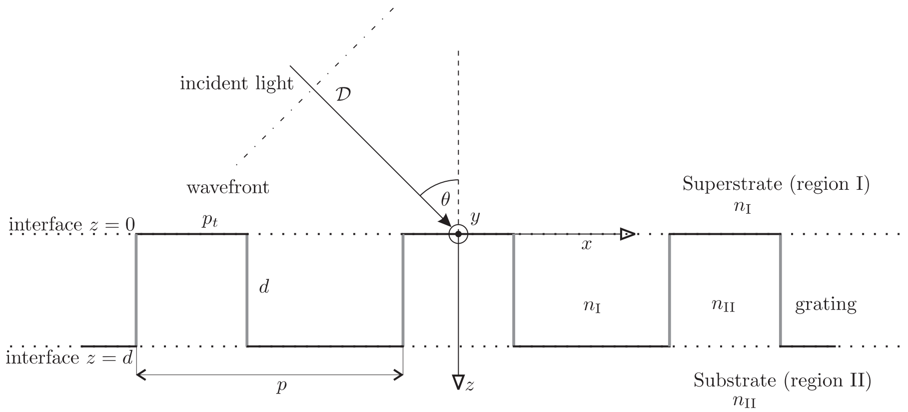

FIGURE 2.1. *Rectangular grating.*

The grating is uniformly extended from $-\infty$ to $+\infty$ in the y coordinate; see [8]. We will consider three subdomains: the superstrate $z < 0$, the grating region $0 \le z \le d$, and the substrate $z > d$, with two horizontal interfaces $z = 0$ (superstrate/grating interface) and $z = d$ (grating/substrate interface). Equations (2.6) and (2.7) will be solved on each domain *separately*, with subsequent matching of the solutions at $z = 0$ and $z = d$ in order to determine the integrating constants. Both materials which form the superstrate, the grooves and the ridges in the grating region, and the substrate are considered linear, isotropic, and *homogeneous*. Consequently, due to the geometry of the grating it is clear that the electric permittivity, which is constant in the superstrate and in the substrate, is in the grating region a function of x but not of z, $\varepsilon \equiv \varepsilon(x)$. This is essential for the RCWA method. It is furthermore *assumed* that $\varepsilon(x)$ is a differentiable function of x, i.e., the vertical lines in the grating region in Figure 2.1 do not represent discontinuities but a very thin layer with a smooth transition from n_{I} to n_{II}. The relevance of this assumption for physical models with the surfaces of discontinuity, as the vertical boundaries in the grating region in Figure 2.1, will be discussed later. Since the geometric structure of the grating is independent of the y coordinate, the electric and magnetic fields depend only on the variables x and z, i.e., $\mathbf{E} \equiv \mathbf{E}(x, z)$ and $\mathbf{H} \equiv \mathbf{H}(x, z)$. As before, the magnetic permeability $\mu = \mu_0$ is constant.

In order to describe the general case, it is sufficient to analyze two special polarizations when the vectors $\mathbf{E}$ and $\mathbf{H}$, respectively, are perpendicular to the plane of incidence x-z, i.e., when the vectors $\mathbf{E}$ and $\mathbf{H}$, respectively, are parallel to the direction of the third Cartesian coordinate y.

For the *Transverse Electric (TE) polarization*, $\mathbf{E} = (0, E_y, 0)$ is parallel to the y axis and $\mathbf{H}$ stays in the x-z plane. For such $\mathbf{E}$ and $\varepsilon \equiv \varepsilon(x)$ the inner product $(\mathbf{E} \cdot \operatorname{grad} \varepsilon)$ vanishes. We stress the point that *here* the geometry of the grating plays a crucial role. The equation (2.6) for $\mathbf{E}$ then reduces (in the superstrate, the grating region and the substrate) to

ETNA
Kent State University
http://etna.math.kent.edu

the wave equation for the single nonzero component E_y,

$$\Delta E_y = -\varepsilon\mu\omega^2 E_y.$$

For the *Transverse Magnetic (TM) polarization*, $\mathbf{H} = (0, H_y, 0)$ is parallel to the y axis and $\mathbf{E}$ stays in the x-z plane. Then,

$$\begin{aligned}
\operatorname{curl}\mathbf{H} &= \left(-\frac{\partial H_y}{\partial z}, 0, \frac{\partial H_y}{\partial x}\right),\\
\operatorname{grad}\varepsilon \times \operatorname{curl}\mathbf{H} &= \left(\frac{\partial \varepsilon}{\partial y}\frac{\partial H_y}{\partial x}, -\frac{\partial \varepsilon}{\partial z}\frac{\partial H_y}{\partial z} - \frac{\partial \varepsilon}{\partial x}\frac{\partial H_y}{\partial x}, \frac{\partial \varepsilon}{\partial y}\frac{\partial H_y}{\partial z}\right)\\
&= -\left(0, \frac{\partial \varepsilon}{\partial x}\frac{\partial H_y}{\partial x}, 0\right) = -(0, \operatorname{grad}\varepsilon\cdot\operatorname{grad}H_y, 0),
\end{aligned}$$

and (2.7) takes the form

$$\operatorname{div}\left(\frac{1}{\varepsilon}\operatorname{grad}H_y\right) = -\mu\omega^2 H_y;$$

see [9, equation (I.22)]. In our notation (recall $\varepsilon \equiv \varepsilon(x)$)

$$\Delta H_y - \frac{1}{\varepsilon(x)}\frac{d\varepsilon(x)}{dx}\frac{\partial H_y}{\partial x} = -\varepsilon(x)\mu\omega^2 H_y.$$

2.3. Summary. Considering $\mu_r = 1$, $\mu = \mu_0$ (this assumption is used throughout the text), $\varepsilon = \varepsilon_0\varepsilon_r$, $c = (\varepsilon_0\mu_0)^{-1/2}$, $\omega = 2\pi f$ and $f\lambda = c$, define

$$k_0^2 \equiv \varepsilon_0\mu_0\omega^2 = \frac{\omega^2}{c^2} = \left(\frac{2\pi f}{c}\right)^2 = \left(\frac{2\pi}{\lambda}\right)^2.$$

The electric field in the TE polarization is then described by the equation

$$\Delta E_y = -k_0^2\varepsilon_r(x)E_y, \qquad E_x = E_z = 0, \tag{2.8}$$

with the magnetic field

$$\mathbf{H} = -\frac{\mathbf{i}}{\mu_0\omega}\operatorname{curl}\mathbf{E},$$

giving

$$(H_x, 0, H_z) = \frac{\mathbf{i}}{\mu_0\omega}\left(\frac{\partial E_y}{\partial z}, 0, -\frac{\partial E_y}{\partial x}\right). \tag{2.9}$$

The magnetic field in the TM polarization is described by the equation

$$\Delta H_y - \frac{1}{\varepsilon_r(x)}\frac{d\varepsilon_r(x)}{dx}\frac{\partial H_y}{\partial x} = -k_0^2\varepsilon_r(x)H_y, \qquad H_x = H_z = 0. \tag{2.10}$$

As explained in Remark 2.1, the previous derivation assumed, for the purpose of simplifying the exposition, that $\sigma = 0$. The resulting equations (2.8)-(2.10) remain valid, with the difference that the corresponding index of refraction is complex, even with $\sigma \neq 0$. In determining the electric field from (2.10) (cf. (2.2)) σ cannot be omitted. Thus we have

$$\mathbf{E} = \frac{1}{-\mathbf{i}\varepsilon_0\varepsilon_r(x)\omega + \sigma(x)}\operatorname{curl}\mathbf{H},$$

ETNA
Kent State University
http://etna.math.kent.edu

which yields

$$(E_x, 0, E_z) = \frac{1}{-\mathbf{i}\varepsilon_0\varepsilon_r(x)\omega + \sigma(x)} \left(-\frac{\partial H_y}{\partial z}, 0, \frac{\partial H_y}{\partial x} \right). \tag{2.11}$$

The given description is valid in the superstrate, in the grating region, and in the substrate. In the following we will use equations (2.8)-(2.9) for the description of the electric and magnetic fields in the TE polarization, and equations (2.10)-(2.11) for the description of the magnetic and electric fields in the TM polarization. In the rest of the text the index of refraction of the substrate is generally complex, i.e., σ is presumed to have, in general, a nonzero value.

3. The RCWA method for a rectangular grating. We will consider the rectangular grating in the x-z plane described above (see Figure 2.1), with its extension from $-\infty$ to $+\infty$ in the y coordinate, where n_{I} and n_{II} denote the indices of refraction of the superstrate and substrate materials, respectively. Throughout the text we assume, consistently with the applications that motivate our work, that there are no losses in the superstrate, i.e., n_{I} is real. The substrate can be conductive, and n_{II} is generally complex with positive real and nonnegative imaginary parts.

The incident electric field is in the TE polarization normal to the plane of incidence, i.e., it is given by its y-component

$$E_y^{\text{inc}} = e^{\mathbf{i}k_0 n_{\mathrm{I}}(x\sin\theta + z\cos\theta)}, \tag{3.1}$$

where $x\sin\theta + z\cos\theta$ determines the phase along the direction $\mathcal{D}$ of the incident wave vector $\mathbf{k}_{\mathrm{I}}$,

$$\mathbf{k}_{\mathrm{I}} = n_{\mathrm{I}}\,\frac{\omega}{c}\,(\sin\theta, 0, \cos\theta),$$

with the wavenumber

$$k_{\mathrm{I}} = \|\mathbf{k}_{\mathrm{I}}\| = n_{\mathrm{I}}\frac{\omega}{c} = n_{\mathrm{I}}\frac{2\pi}{\lambda};$$

see [9, relation (I.16)]. Please note that with (2.5) this gives the time-harmonic field

$$\widehat{E}_y^{\text{inc}} = \mathrm{Re}[e^{\mathbf{i}\{k_0 n_{\mathrm{I}}(x\sin\theta + z\cos\theta) - \omega t\}}],$$

which corresponds to the wave propagating in the direction of increasing x and z, i.e., down and to the right. Similarly, in the TM polarization the incident magnetic field is normal to the plane of incidence,

$$H_y^{\text{inc}} = e^{\mathbf{i}k_0 n_{\mathrm{I}}(x\sin\theta + z\cos\theta)}. \tag{3.2}$$

The RCWA method will first be described assuming TE polarization, and then applied to TM polarization.

3.1. Planar diffraction: TE polarization. For completeness we will briefly derive some basics of the differential theory of gratings. A correspondence to the standard literature will be given by referencing the formulas and page numbers in [9]. An extensive survey of the literature can be found in [6, 7, 9].

Since the grating surface is periodic with period p and infinite, translation in the x coordinate from x to $x + p$ multiplies the incident wave (3.1) by the phase factor $e^{\mathbf{i}k_{\mathrm{I}}p\sin\theta}$, as depicted in Figure 3.1.

ETNA
Kent State University
http://etna.math.kent.edu

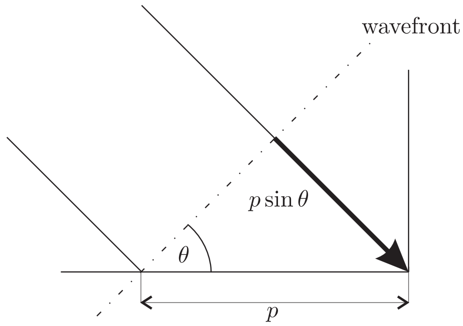

FIGURE 3.1. *Phase determining the periodicity.*

In linear optics, the transformation of the incident field into the total field is linear, therefore the total field $E_y(x,z)$ satisfies

$$E_y(x+p,z) = e^{\mathbf{i}k_{\mathrm{I}}p\sin\theta}\, E_y(x,z),$$

which must hold in the superstrate, in the grating and in the substrate; see [9, relation (I.27)]. Consequently the function $F(x,z) \equiv e^{-\mathbf{i}\,k_{\mathrm{I}}x\sin\theta}E_y(x,z)$ is periodic in x with period p,

$$F(x+p,z) = e^{-\mathbf{i}k_{\mathrm{I}}(x+p)\sin\theta}E_y(x+p,z) = e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta}E_y(x,z) = F(x,z).$$

This periodicity is used for *separation of the space variables* using the following Fourier expansion

$$F(x,z) = \sum_{s=-\infty}^{+\infty} f_s(z)\, e^{\mathbf{i}s\frac{2\pi}{p}x}, \quad \text{i.e.,} \quad E_y(x,z) = \sum_{s=-\infty}^{+\infty} f_s(z)e^{\mathbf{i}k_{xs}x}, \tag{3.3}$$

where $f_s(z)$ are the Fourier coefficients independent of x,

$$k_{xs} \equiv k_{\mathrm{I}}\sin\theta + s\frac{2\pi}{p} = k_0\left(n_{\mathrm{I}}\sin\theta + s\frac{\lambda}{p}\right), \quad s = 0,1,-1,\ldots; \tag{3.4}$$

see [9, relations (I.3), p. 3, and (I.29”), p. 22]. With n_{I} real, k_{I} and k_{xs} are also real. The relations in (3.4) are called the Floquet conditions. Since the wavenumber of the reflected field is preserved,

$$k_{\mathrm{I}}^2 = k_{xs}^2 + k_{\mathrm{I},zs}^2, \tag{3.5}$$

(3.4) determines the discrete diffraction angles for which

$$\sin\theta_s = \frac{k_{xs}}{k_{\mathrm{I}}} = \sin\theta + s\frac{\lambda}{n_{\mathrm{I}}p}; \tag{3.6}$$

see Figure 3.2 and [9, relation (I.5)].

Relations (3.4), (3.6) represent the diffraction law for the grating. It replaces the common Snell’s law for specular surfaces which simply states that the tangential component k_x is preserved. Here k_{xs} can take different values (3.4) for different integers s.

It remains to determine the Fourier coefficients $f_s(z)$ in (3.3).

ETNA
Kent State University
http://etna.math.kent.edu

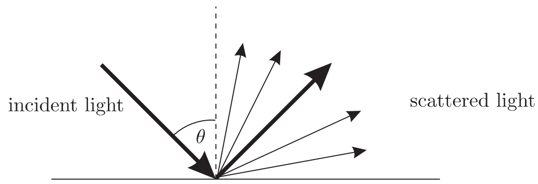

FIGURE 3.2. *Angles between the diffraction orders.*

3.2. Solution in the superstrate and in the substrate – TE polarization. In the homogeneous superstrate and substrate ε_r is constant, and (2.8) takes the form of the Helmholtz equation

$$\Delta E_y = -k_\ell^2\, E_y, \quad E_x = E_z = 0, \quad \ell = \mathrm{I}, \mathrm{II}. \tag{3.7}$$

Therefore, for $z < 0$ (superstrate) and $z > d$ (substrate) introducing the Fourier expansion (3.3) into (3.7) gives the infinite set of *uncoupled* ordinary differential equations for the unknown coefficients $f_s(z)$,

$$\left[\frac{d^2}{dz^2} + k_{\ell,zs}^2\right] f_s^{(\ell)}(z) = 0, \quad \ell = \mathrm{I}, \mathrm{II}, \quad s = 0, 1, -1, \dots, \tag{3.8}$$

where $k_{\ell,zs}^2 = k_\ell^2 - k_{xs}^2$; see (3.5). A general solution can be written as

$$f_s^{(\ell)}(z) = A_s^{(\ell)} e^{-\mathbf{i}k_{\ell,zs}z} + B_s^{(\ell)} e^{\mathbf{i}k_{\ell,zs}z}, \tag{3.9}$$

where $A_s^{(\ell)}$, $B_s^{(\ell)}$ are integrating constants. The physically meaningful solution is bounded when the waves propagate away from the grating, which means that the unbounded part of (3.9) is nonphysical and must be excluded.

Since the superstrate is lossless, the refraction index n_{I} is real, and therefore

$$k_{\mathrm{I},zs} = \sqrt{k_{\mathrm{I}}^2 - k_{xs}^2} \tag{3.10}$$

is real and positive if $k_{\mathrm{I}} > k_{xs}$, and zero or purely imaginary with positive imaginary part if $k_{\mathrm{I}} \le k_{xs}$, $k_{\mathrm{I}} = k_0 n_{\mathrm{I}}$. When $k_{\mathrm{I},zs}$ is real and positive the term $A_s^{(\ell)} e^{-\mathbf{i}k_{\ell,zs}z}$ corresponds to the wave propagating in the direction of decreasing z, i.e., going up, while the term $B_s^{(\ell)} e^{\mathbf{i}k_{\ell,zs}z}$ corresponds to the wave propagating in the direction of increasing z, i.e., going down. If $k_{\mathrm{I},zs}$ is zero or purely imaginary then there is no wave propagating in the z direction, the corresponding modes are evanescent and they will not be further considered.

Keeping a single incident wave (with $s = 0$) and considering no incidence from the substrate, the solution of (3.7) in the superstrate ($z < 0$) can finally be written in the form ($R_s \equiv A_s^{(I)}$)

$$E_y^{\mathrm{I}} = e^{\mathbf{i}k_{\mathrm{I}}(x\sin\theta + z\cos\theta)} + \sum_{s=-\infty}^{+\infty} R_s e^{\mathbf{i}k_{xs}x - \mathbf{i}k_{\mathrm{I},zs}z}. \tag{3.11}$$

Since k_{xs} is real and k_{II} corresponding to the substrate has a positive real and a nonnegative imaginary part,

$$k_{\mathrm{II},zs}^2 = k_{\mathrm{II}}^2 - k_{xs}^2$$

ETNA
Kent State University
http://etna.math.kent.edu

must also have a nonnegative imaginary part, with the real part positive or negative (the evanescent modes are not considered). Its square root is taken in the first quadrant, with positive real and nonnegative imaginary parts. Then, the resulting solution of (3.7) represented by the wave propagating in the substrate ($z > d$) in the direction of increasing z, i.e., travelling down, is given by ($\widehat{T}_s \equiv B_s^{(II)}$)

$$E_y^{\mathrm{II}} = \sum_{s=-\infty}^{+\infty} \widehat{T}_s e^{\mathbf{i}k_{xs}x + \mathbf{i}k_{\mathrm{II},zs}z}; \tag{3.12}$$

cf. [9, (I.35) respectively (I.38), p. 24]. Please note that E_y^{II} is bounded when $z \to +\infty$, which complies with the physical requirement. The fact that E_y^{I} is bounded when $z \to -\infty$ follows trivially since $k_{\mathrm{I},zs}$ is real or purely imaginary (with no wave propagating in the latter case).

Since the imaginary part of $k_{\mathrm{II},zs}$ is nonnegative, the real part of $e^{\mathbf{i}k_{\mathrm{II},zs}z}$ can be rather small for $z = d$, which can cause difficulties in *numerical computations*. Therefore, it might be convenient to consider the following scaling

$$E_y^{\mathrm{II}} = \sum_{s=-\infty}^{+\infty} T_s e^{\mathbf{i}k_{xs}x + \mathbf{i}k_{\mathrm{II},zs}(z-d)}, \tag{3.13}$$

where

$$\widehat{T}_s = T_s e^{-\mathbf{i}k_{\mathrm{II},zs}d}. \tag{3.14}$$

As a consequence, T_s can be expected to be much smaller in magnitude than $\widehat{T}_s$. It should also be noticed that the scaling is equivalent to moving the origin in the z direction by a distance d. In the following derivation we will continue with the scaled expansion (3.13), and we will comment on the effect of non-scaling on the derived algebraic system later.

The integrating constants R_s and T_s have to be determined from the boundary conditions on the top $(z = 0)$ and the bottom $(z = d)$ of the grating region.

REMARK 3.1. It should be noted that we use a different orientation of the z coordinate than the y coordinate in [9].

3.3. Infinite set of differential equations for the grating region – TE polarization. In the grating region, $\varepsilon_r(x)$ represents a periodic (sufficiently smooth) function with respect to x with period p. It can therefore be expressed by its Fourier series

$$\varepsilon_r(x) = \sum_{h=-\infty}^{+\infty} \epsilon_h e^{\mathbf{i}h\frac{2\pi}{p}x}. \tag{3.15}$$

Later (for the TM polarization) it will be convenient to consider also the subsequent Fourier expansions

$$\frac{1}{\varepsilon_r(x)} = \sum_{h=-\infty}^{+\infty} a_h e^{\mathbf{i}h\frac{2\pi}{p}x}. \tag{3.16}$$

If the geometry of the grating is symmetric with respect to x, as in Figure 2.1, the equality $\varepsilon_r(x) = \varepsilon_r(-x)$ gives

$$\epsilon_h = \epsilon_{-h} \quad \text{and} \quad a_h = a_{-h}, \quad h = 1, 2, \ldots \tag{3.17}$$

The symmetry of the grating is not substantial here; it affects only the properties of the matrix blocks in the resulting linear algebraic systems. With the Fourier expansion (3.3), (3.15) leads again to the separation of the x and z variables and to a reduction of the problem to a set of ordinary differential equations for the Fourier amplitudes $f_s(z)$, $s = 0, 1, -1, \ldots$, which, unlike (3.8), are coupled. The separation of variables is the key issue in the RCWA method. When the analytic solution of the *truncated* system of ODEs is expressed in a form of matrix functions, the boundary conditions formulated for $z = 0$ and $z = d$ give the linear algebraic systems for the integrating constants.

Inserting the Fourier expansions (3.3) and (3.15) into (2.8) then gives (cf. [9, (II.2), p. 38])

$$\left[\frac{d^2}{dx^2} + \frac{d^2}{dz^2}\right] \sum_{s=-\infty}^{+\infty} f_s(z) e^{\mathbf{i}k_{xs}x} \;=\; -k_0^2 \sum_{h=-\infty}^{+\infty} \epsilon_h e^{\mathbf{i}h\frac{2\pi}{p}x} \sum_{s=-\infty}^{+\infty} f_s(z) e^{\mathbf{i}k_{xs}x}.$$

Substituting for k_{xs} in the exponentials, straightforward manipulations lead to (we leave a discussion of some important details to Section 3.6)

$$\sum_{j=-\infty}^{+\infty} \left\{ \left[\frac{d^2}{dz^2} - k_{xj}^2\right] f_j(z) \right\} e^{\mathbf{i}j\frac{2\pi}{p}x} \;=\; -k_0^2 \sum_{j=-\infty}^{+\infty} \left\{ \sum_{s=-\infty}^{+\infty} \epsilon_{j-s} f_s(z) \right\} e^{\mathbf{i}j\frac{2\pi}{p}x}. \tag{3.18}$$

Equating for the index j gives the result

$$\frac{d^2 f_j(z)}{dz^2} \;=\; k_{xj}^2 f_j(z) - k_0^2 \sum_{s=-\infty}^{+\infty} \epsilon_{j-s} f_s(z). \tag{3.19}$$

Note that for any homogeneous medium in which only ϵ_0 is nonzero (and $\varepsilon_r(x)$ is constant), (3.19) decouples into the set of independent equations (3.8).

It is common to use the scaling $w = zk_0$. Using the new scaled variable w, (3.19) takes the form

$$\frac{d^2 f_j(w)}{dw^2} \;=\; \frac{k_{xj}^2}{k_0^2} f_j(w) - \sum_{s=-\infty}^{+\infty} \epsilon_{j-s} f_s(w), \quad j = 0, 1, -1, 2, -2, \ldots. \tag{3.20}$$

We emphasize the fact that under standard assumptions on the convergence of the Fourier expansions above, (3.20) represents one particular form, out of many mathematically equivalent forms, of writing the infinite set of differential equations for the Fourier amplitudes $f_j(w)$, $j = 0, 1, -1, 2, -2, \ldots$. After truncation, such mathematically equivalent forms can produce truncated finite dimensional problems which have *different approximation errors and convergence properties*. The next two subsections discuss the method of truncation used in the standard RCWA method. Open questions related to the truncation of the Fourier expansions and the infinite system of differential equations given above will be discussed later in Section 4.

3.4. Truncation – TE polarization. For numerical computations, it is necessary to truncate the infinite Fourier expansions. From this point forward, we will consider that the computed fields are described with sufficient accuracy by their $2N + 1$ Fourier components. The choice of N depends on the problem; the corresponding truncation error should be in balance with the accuracy of subsequent numerical computations, in particular with the accuracy of solving the system of ODEs (approximation of matrix functions) and the accuracy

ETNA
Kent State University
http://etna.math.kent.edu

of solving the final system of linear algebraic equations for integrating constants described below.

In the superstrate and in the substrate (see (3.11) and (3.13))

$$
\begin{aligned}
E_y^{\mathrm{I}} &= e^{\mathbf{i}k_{\mathrm{I}}(x\sin\theta + z\cos\theta)} + \sum_{s=-N}^{N} R_s e^{\mathbf{i}k_{xs}x - \mathbf{i}k_{\mathrm{I},zs}z} \\
&= e^{\mathbf{i}k_{\mathrm{I}}z\cos\theta}\, e^{\mathbf{i}k_{\mathrm{I}}x\sin\theta} + \sum_{s=-N}^{N} \left(R_s e^{-\mathbf{i}k_{\mathrm{I},zs}z}\right) e^{\mathbf{i}k_{xs}x} \\
&\equiv \sum_{s=-N}^{N} u_{\mathrm{I},y}^{(s)}(z) e^{\mathbf{i}k_{xs}x}, \qquad (3.21)\\
E_y^{\mathrm{II}} &= \sum_{s=-N}^{N} T_s\, e^{\mathbf{i}k_{xs}x + \mathbf{i}k_{\mathrm{II},zs}(z-d)} = \sum_{s=-N}^{N} \left(T_s e^{\mathbf{i}k_{\mathrm{II},zs}(z-d)}\right) e^{\mathbf{i}k_{xs}x} \\
&\equiv \sum_{s=-N}^{N} u_{\mathrm{II},y}^{(s)}(z) e^{\mathbf{i}k_{xs}x}. \qquad (3.22)
\end{aligned}
$$

We use here for simplicity the same notation for E_y^{I} and E_y^{II} as in (3.12) and (3.13), i.e., we omit in (3.21) and (3.22) the index N denoting the truncation order of the Fourier modes. Denoting

$$
r_{\mathrm{TE}} = \begin{bmatrix} R_{-N} \\ \vdots \\ R_0 \\ \vdots \\ R_N \end{bmatrix} \in \mathbb{C}^{2N+1}, \quad t_{\mathrm{TE}} = \begin{bmatrix} T_{-N} \\ \vdots \\ T_0 \\ \vdots \\ T_N \end{bmatrix} \in \mathbb{C}^{2N+1}, \qquad (3.23)
$$

$$
\begin{aligned}
Y_{\mathrm{I}} &= \mathrm{diag}(k_{\mathrm{I},zs}/k_0) \in \mathbb{C}^{(2N+1)\times(2N+1)}, \qquad (3.24)\\
Y_{\mathrm{II}} &= \mathrm{diag}(k_{\mathrm{II},zs}/k_0) \in \mathbb{C}^{(2N+1)\times(2N+1)}, \qquad (3.25)
\end{aligned}
$$

the parts in the truncated Fourier expansions (3.21) and (3.22) dependent on the z variable can be written, using vector notation, as

$$
u_y^{\mathrm{I}} = \begin{bmatrix} u_{\mathrm{I},y}^{(-N)} \\ \vdots \\ u_{\mathrm{I},y}^{(0)} \\ \vdots \\ u_{\mathrm{I},y}^{(N)} \end{bmatrix} = \begin{bmatrix} R_{-N} e^{-\mathbf{i}k_{\mathrm{I},z(-N)}z} \\ \vdots \\ R_0 e^{-\mathbf{i}k_{\mathrm{I}z0}z} \\ \vdots \\ R_N e^{-\mathbf{i}k_{\mathrm{I},zN}z} \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ e^{\mathbf{i}k_{\mathrm{I}}z\cos\theta} \\ \vdots \\ 0 \end{bmatrix},
$$

$$
u_y^{\mathrm{II}} = \begin{bmatrix} u_{\mathrm{II},y}^{(-N)} \\ \vdots \\ u_{\mathrm{II},y}^{(N)} \end{bmatrix} = \begin{bmatrix} T_{-N} e^{\mathbf{i}k_{\mathrm{II},z(-N)}(z-d)} \\ \vdots \\ T_N e^{\mathbf{i}k_{\mathrm{II},zN}(z-d)} \end{bmatrix},
$$

where $k_{x0} = k_{\rm I} \sin\theta$, $k_{{\rm I},z0} = k_{\rm I}\cos\theta$; see (3.4) and (3.10). With the scaling $w = zk_0$, and using matrix exponentials,

$$u_y^{\rm I} = e^{-\mathbf{i}Y_{\rm I}w} r_{\rm TE} + e^{\mathbf{i}Y_{\rm I}w} e_0, \tag{3.26}$$

$$u_y^{\rm II} = e^{\mathbf{i}Y_{\rm II}(w-dk_0)} t_{\rm TE}, \tag{3.27}$$

where the last term in (3.26), $e_0 = [0,\dots,0,1,0,\dots,0]^T$, corresponds to the incident plane wave given above (with the single nonzero spectral mode).

Similarly to (3.21)-(3.22) we consider in the grating region the truncated expansion (see (3.3))

$$E_y^G = \sum_{s=-N}^{N} f_s(w) e^{\mathbf{i}k_{xs}x}, \quad u_y^G(w) \equiv \begin{bmatrix} f_{-N}(w) \\ \vdots \\ f_N(w) \end{bmatrix}. \tag{3.28}$$

The $2N+1$ differential equations for the parts of the Fourier expansion dependent on z in (3.20), $j = -N,\dots,N$, can be written in matrix form

$$\frac{d^2 u_y^G}{dw^2} = -Cu_y^G, \quad C = \Upsilon - Y_G^2 \in \mathbb{C}^{(2N+1)\times(2N+1)}, \tag{3.29}$$

where

$$\begin{aligned} Y_G &= \mathrm{diag}(k_{xs}/k_0) \\ &= \mathrm{diag}\Big(n_{\rm I}\sin\theta + N\frac{\lambda}{p},\dots,n_{\rm I}\sin\theta,\dots,n_{\rm I}\sin\theta - N\frac{\lambda}{p}\Big), \end{aligned} \tag{3.30}$$

$$(\Upsilon)_{js} = \epsilon_{j-s}, \quad j,s = -N,\dots,0,\dots,N. \tag{3.31}$$

Here Υ represents a Toeplitz matrix with the entries determined by the Fourier expansion of the relative permittivity in the grating region. Since for the simple geometry of the grating (3.17) holds, Υ and, consequently, C are complex symmetric. A general solution of (3.29) is then given in matrix form by

$$u_y^G = e^{\mathbf{i}\sqrt{C}w} g_{\rm TE}^+ + e^{-\mathbf{i}\sqrt{C}w} \widehat{g}_{\rm TE}^-, \tag{3.32}$$

where $g_{\rm TE}^+$ and $\widehat{g}_{\rm TE}^-$ represent the corresponding vectors of the integrating constants.

Assume, for a moment, that $\sqrt{C}$ is a single complex number with a *positive real* and a nonzero imaginary parts. Then, the first term in (3.32) corresponds to the downward and the second part to the upward wave in the grating region ($0 \le w \le dk_0$). The fact that the signal can only be damped, not amplified, which means that the energy of the signal cannot grow in the direction of its propagation, requires in both cases the *positive imaginary part of the square root*; cf. [13, Section 24.3, relations (24.37), (24.38), (24.51) and (24.55)]. It should be realized, however, that only if the real part of the square root is *positive*, then with our choice in (2.5) the first part in (3.32) corresponds to the downward and the second part to the upward wave in the grating region. With the positive imaginary part of $\sqrt{C}$ the wave corresponding to $e^{\mathbf{i}\sqrt{C}w} g_{\rm TE}^+$ is then damped with increasing w, while the wave corresponding to $e^{-\mathbf{i}\sqrt{C}w}\widehat{g}_{\rm TE}^-$ is damped with *decreasing* w, which is in agreement with the waves propagating downwards and upwards, respectively.

If, however, the real part of the square root $\sqrt{C}$ is negative, then with our choice in (2.5) the second part in (3.32) corresponds to the downward and the first part to the upward wave in the grating region. Then the requirement of non-amplification of the signal (which is frequently in the engineering literature identified with stability) implies that the imaginary part of $\sqrt{C}$ must be negative.

It should be noted that the non-amplification of the signal requires the real and imaginary parts of $\sqrt{C}$ *to have the same sign*. If C is in the upper half plane, then the principal square root of C lies in the first quadrant, and the solution of the discretized problem (3.32) has a straightforward physical interpretation. If, however, C is in the lower half part of the complex plane, i.e., it has a negative imaginary part, then the real and imaginary parts of the square root $\sqrt{C}$ cannot have the same sign, *no matter which branch of the complex square root is considered*. In such a case the physical meaning of the discretized solution is unclear, since the signal is inevitably amplified in one of the directions of its propagation.

In the RCWA method, C is a *matrix*, and the considerations above apply to every individual eigenvalue of C; cf. [4] and [5, Section 6.2]. Indeed, denoting by $C = UJU^{-1}$ the Jordan canonical form of C, (3.32) in fact means

$$u_y^G = Ue^{\mathbf{i}\sqrt{J}w}U^{-1}g_{\mathrm{TE}}^{+} + Ue^{-\mathbf{i}\sqrt{J}w}U^{-1}g_{\mathrm{TE}}^{-}. \tag{3.33}$$

If all eigenvalues of C lie in the upper half plane, then the principal value of the complex square root will be in the first quadrant for *all* eigenvalues, and it make sense to state that the square root in (3.32) corresponds to the branch with the positive imaginary part.

Here *we assume* that C indeed has all its eigenvalues in the upper half plane. Whether such an assumption restricts the applicability of the RCWA method is yet to be found (see the discussion below) and we pose it as an open problem.

REMARK 3.2. Discussions of the choice of a branch of the complex square root lacks completeness in the literature on RCWA known to us. In particular, the consequences of the fact that non-amplification of the signal in the direction of propagation links together the signs of *both* real and imaginary parts of the eigenvalues in (3.33), with its consequences for $\sqrt{C}$, are not clearly explained. Sometimes the signs of the *real parts* of the eigenvalues of $\sqrt{C}$ are ignored, and the positive imaginary parts of the eigenvalues of $\sqrt{C}$ are identified with damping, independently of the direction in which the signal propagates. Such an approach is not correct. For example, the negative real part of the eigenvalue of $\sqrt{C}$ corresponds in $e^{\mathbf{i}\sqrt{C}w}g_{\mathrm{TE}}^{+}$ to the wave propagating in the direction of decreasing w, and therefore the positive imaginary part of the eigenvalue of $\sqrt{C}$ means in such a case an amplification, not damping, of the signal in the direction of propagation. Similarly, the negative real part of of the eigenvalue of $\sqrt{C}$ corresponds in $e^{-\mathbf{i}\sqrt{C}w}\widehat{g}_{\mathrm{TE}}^{-}$ to the wave going in the direction of increasing w, and the positive imaginary part of the eigenvalue of $\sqrt{C}$ means in this case an amplification, not damping, of the signal in the direction of propagation.

Positive imaginary parts of the eigenvalues of $\sqrt{C}$ can cause numerical difficulties; cf. Section 3.2. We will therefore use, as above, the following scaling

$$u_y^G = e^{\mathbf{i}\sqrt{C}w}g_{\mathrm{TE}}^{+} + e^{-\mathbf{i}\sqrt{C}(w-dk_0)}g_{\mathrm{TE}}^{-}, \tag{3.34}$$

where, comparing with (3.32),

$$\widehat{g}_{\mathrm{TE}}^{-} = e^{\mathbf{i}\sqrt{C}dk_0}g_{\mathrm{TE}}^{-}. \tag{3.35}$$

In the following derivation we will continue with the scaled expansion (3.34), and we will comment on the effect of non-scaling to the derived algebraic system later.

Summarizing, (3.26), (3.27), and (3.34) describe the w (or z) dependent $2N+1$ Fourier coefficients of the truncated Fourier expansion (in the variable x) of the electric field E_y in the superstrate, substrate, and in the grating region, respectively.

3.5. Matching on the boundaries and formulation of the algebraic problem – TE polarization. In order to determine the integrating constants, which represent the vectors $r_{\rm TE}$, $t_{\rm TE}$, $g^+_{\rm TE}$ and $g^-_{\rm TE}$, each of length $2N+1$, we have two sets of $2N+1$ equations for matching the electric field at $z=0$ and $z=d$ (top and bottom of the grating region). Two missing sets of $2N+1$ equations can be obtained by matching the tangential components H_x of the magnetic field [9, pp. 39–40] given by (see (2.9))

$$H_x = \frac{\mathbf{i}}{\mu_0\omega}\,\frac{\partial E_y}{\partial z} = \mathbf{i}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2}\,\frac{\partial E_y}{\partial w}.$$

Unlike in some other methods for computing the diffraction of light on gratings, the RCWA method deals with the grating region *mathematically* as a single region with the electric permittivity dependent on x. Consequently, there are no other boundary conditions to consider.

Using the truncated Fourier expansions for E_y, (see (3.21), (3.22), and (3.28)), and differentiating the Fourier coefficients (3.26), (3.27), and (3.34) gives

$$\frac{\partial u_y^{\rm I}}{\partial w} = -\mathbf{i}Y_{\rm I}e^{-\mathbf{i}Y_{\rm I}w}r_{\rm TE} + \mathbf{i}Y_{\rm I}e^{\mathbf{i}Y_{\rm I}w}e_0, \tag{3.36}$$

$$\frac{\partial u_y^{\rm II}}{\partial w} = \mathbf{i}Y_{\rm II}e^{\mathbf{i}Y_{\rm II}(w-dk_0)}t_{\rm TE}, \tag{3.37}$$

$$\frac{\partial u_y^{\rm G}}{\partial w} = \mathbf{i}\sqrt{C}e^{\mathbf{i}\sqrt{C}w}g^+_{\rm TE} - \mathbf{i}\sqrt{C}e^{-\mathbf{i}\sqrt{C}(w-dk_0)}g^-_{\rm TE}. \tag{3.38}$$

Finally, writing the boundary matching conditions

$$-E_y^{\rm I}(x,0) + E_y^{\rm G}(x,0) = 0, \quad -H_x^{\rm I}(x,0) + H_x^{\rm G}(x,0) = 0$$

at $z=0$, and

$$+E_y^{\rm G}(x,d) - E_y^{\rm II}(x,d) = 0, \quad +H_x^{\rm G}(x,d) - H_x^{\rm II}(x,d) = 0$$

at $z=d$ into one matrix equations for the unknown integrating constants $r_{\rm TE}$, $g^+_{\rm TE}$, $g^-_{\rm TE}$, $t_{\rm TE}$, gives the large $4(2N+1)\times 4(2N+1)$ linear algebraic system (the second and the fourth block equations have been multiplied by $-\mathbf{i}$)

$$\begin{bmatrix} -I & I & e^{\mathbf{i}\sqrt{C}dk_0} & 0 \\ Y_I & \sqrt{C} & -\sqrt{C}e^{\mathbf{i}\sqrt{C}dk_0} & 0 \\ 0 & e^{\mathbf{i}\sqrt{C}dk_0} & I & -I \\ 0 & \sqrt{C}e^{\mathbf{i}\sqrt{C}dk_0} & -\sqrt{C} & -Y_{\rm II} \end{bmatrix} \begin{bmatrix} r_{\rm TE} \\ g^+_{\rm TE} \\ g^-_{\rm TE} \\ t_{\rm TE} \end{bmatrix} = \begin{bmatrix} e_0 \\ n_{\rm I}\cos\theta\, e_0 \\ 0 \\ 0 \end{bmatrix}, \tag{3.39}$$

denoted in the following as

$$A_{\rm TE}\xi_{\rm TE} = b_{\rm TE}.$$

It should be noted that in most practical measurements one does not actually need the full solution $\xi_{\rm TE}$. Typically, only the zeroth order mode of $r_{\rm TE}$, which can be expressed as

$$r_{0,\rm TE} = (e_0^T, 0)\xi_{\rm TE} = e_0^T r_{\rm TE},$$

is required.

If we use the unscaled blocks of unknowns $\widehat{g}^-_{\rm TE}$ and $\widehat{t}_{\rm TE}$ (see (3.14) and (3.35)), the matrix of the linear algebraic system (3.39) will have the last two columns multiplied by the corresponding factors, which will increase its condition number and make it less suitable for numerical computations.

3.6. Subtleties of the discretization. Now we describe in more detail the steps which lead to the system of the linear algebraic equations (3.39). Using Maxwell's equations and assuming a plane time-harmonic wave, we have derived the second order equation

$$\Delta E_y = -k_0^2 \varepsilon_r(x) E_y \tag{3.40}$$

for the electric field component E_y. Then we have considered Fourier expansions

$$E_y(x,z) = \sum_{s=-\infty}^{\infty} f_s(z) e^{\mathbf{i}\left(k_{\mathrm{I}} \sin\theta + s\frac{2\pi}{p}\right)x}, \quad \varepsilon_r(x) = \sum_{h=-\infty}^{\infty} \epsilon_h e^{\mathbf{i}h\frac{2\pi}{p}x}, \tag{3.41}$$

where the second reflects the dependence of ε_r on x in the grating region. Since ε_r is constant in the superstrate and in the substrate, substitution for E_y into (3.40) yields decoupled second order differential equations for the unknown coefficients f_s, $s = 0, -1, 1, \ldots$; see (3.8). Writing down the solution for a *finite* subset

$$f_{-N}, \ldots, f_0, \ldots, f_N,$$

which means *truncation* of the first expansion in (3.41), gives finally the truncated approximation E_y^{I} and E_y^{II} to the solution in the superstrate and in the substrate respectively; see (3.21)-(3.27).

In the grating region the situation is more complicated due to the fact that $\varepsilon_r(x)$ is not constant there, and $\varepsilon_r E_y$ represents the product of two Fourier series (3.41),

$$\begin{aligned} e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta}\varepsilon_r E_y &= \sum_{h=-\infty}^{+\infty} \epsilon_h e^{\mathbf{i}h\frac{2\pi}{p}x} \sum_{s=-\infty}^{+\infty} f_s(z) e^{\mathbf{i}s\frac{2\pi}{p}x} \\ &= \sum_{j=-\infty}^{+\infty} \left(\sum_{s=-\infty}^{+\infty} \epsilon_{j-s} f_s(z) \right) e^{\mathbf{i}j\frac{2\pi}{p}x} \\ &= \lim_{N\to\infty} \sum_{j=-N}^{N} \left(\lim_{M\to\infty} \sum_{s=-M}^{M} \epsilon_{j-s} f_s(z) \right) e^{\mathbf{i}j\frac{2\pi}{p}x}, \end{aligned} \tag{3.42}$$

where the last line represents the precise formulation. Considering the particular *simultaneous truncation* with a fixed $M = N$, we obtain the truncated approximation E_y^{G} to the solution in the grating region. Matching E_y^{I}, E_y^{G} and E_y^{II}, H_x^{I}, H_x^{G} and H_x^{II} on the boundaries gives the algebraic system (3.39) for the integrating constants.

The last line of the identity (3.42) represents one of the crucial points of the whole derivation. The two functions $\varepsilon_r(x)$ and $e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta}E_y(x,z)$ are periodic in the x direction with period p; these are expanded into Fourier series and then multiplied. Their product is expressed as a Fourier series and then approximated by *the simultaneous truncation*

$$e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta}\varepsilon_r(x)E_y(x,z) = \lim_{N\to\infty} \sum_{j=-N}^{N} \psi_{1,j}^{(N)}(z) e^{\mathbf{i}j\frac{2\pi}{p}x}, \tag{3.43}$$

where

$$\psi_{1,j}^{(N)}(x) = \sum_{s=-N}^{N} \epsilon_{j-s} f_s(z)$$

is *also a truncated approximation* of the true Fourier amplitude

$$\psi_{1,j}(z) = \sum_{s=-\infty}^{+\infty} \epsilon_{j-s} f_s(z),$$

known in the literature as Laurent's rule [3, p. 240], [6, 7], [9, Chapter IV], though the principle can be connected to Cauchy's summation rule; see [3, p. 227]. Here everything relies upon the convergence of the limit of the simultaneously truncated expansion in (3.43).

In general, when the multiplied functions are piecewise smooth bounded periodic functions which have no common discontinuities, which is satisfied in (3.43) using our assumption that $\varepsilon_r(x)$ and $E_y(x, z)$ are sufficiently smooth, the series converges [6, Theorem 1, p. 1872], [7, Theorem 4.3, p. 122]. Then the infinite set of differential equations (3.20) is truncated into the set of $2N + 1$ differential equations for $2N + 1$ unknown functions (see (3.29)) and the solution u_y^{G} is expressed in matrix form by (3.34). In other words, the ODE problem (3.20) for an infinite number of unknown functions $f_j(w)$ is approximated using the *truncated Laurent's rule* by the set of $2N + 1$ ordinary differential equations for $u_y^{\mathrm{G}}(w) = [f_{-N}(w), \ldots, f_N(w)]^T$. The whole solution process is justified by the convergence of the limit on the right hand side of (3.43) to the function on the left hand side of that identity. Without convergence and equality in (3.43), the truncation would lead to an incorrect result, since the solution of the truncated problem would in general not converge for $N \to \infty$ to the solution of the original problem. Here the convergence is meant point-wise, not in a norm which ignores sets of measure zero; see [7, Section 4.4.2].

The considerations above may seem obvious, but it is useful to include them here. Though the matter is explained in some mathematically oriented papers [6, 7], and also, using less rigorous arguments, in a more practically focused book [9, Chapter IV], the consequences do not seem fully realized by the community of practitioners. In particular, if we have two piecewise smooth bounded periodic functions which have common jump discontinuities, then the truncated Laurent's rule cannot be applied; see [6, Theorem 2, p. 1872], [7, Theorem 4.4, pp. 122–123]. If, however, the *product* of the two functions is continuous at the points of their common discontinuities, then, under some nonsingularity assumptions, it can be expressed as a Fourier expansion using the *truncated inverse multiplication rule* [6, Theorem 3, p. 1872, relation (22)], [7, Theorem 4.5, p. 123, relation (4.32) and the examples in Section 4.4.4]. In the derivation of (2.6)-(2.7) we have assumed *smooth functions* and therefore the discussion of discontinuities may seem irrelevant. In physics, however, one has to deal with modelling of the idealized surfaces of discontinuity; see [13, Chapter 9]. In order to get a good match of the computed results with physical reality, it is therefore necessary to use truncation rules which *in the limit* remain valid in the presence of discontinuities.

It should be emphasized that the truncated inverse multiplication rule, which will be applied in the following section, cannot be viewed as a mechanical rule derived simply by the truncation on both sides of the rearranged identities, using Laurent's rule followed by the inversion of the matrix of truncated coefficients, as inaccurately interpreted in [9, Section IV.2.1, p. 82, relation (IV.10) and its derivation given there]; see also [11]. Though such derivation may give the correct result, it is neither complete nor mathematically correct. It does not prove *the convergence* of the resulting approximation of the Fourier expansion; see [7, proof of Theorem 4.5, Appendix A, pp. 136–137].

The common subtle mistake, which has led to incorrectly discretized formulations used in practice, is caused by overlooking the following fact. Let $[\![\Gamma]\!]^{(N)}$ denote the Toeplitz matrix

$$[\![\Gamma]\!]^{(N)}_{js} = \gamma_{j-s}, \quad j, s = -N, \ldots, 0, \ldots, N, \tag{3.44}$$

ETNA
Kent State University
http://etna.math.kent.edu

generated by the Fourier coefficients of some given function

$$\Gamma(x) = \sum_{s=-\infty}^{+\infty} \gamma_s e^{\mathbf{i} s \frac{2\pi}{p} x}.$$

Assume that Γ^{-1} has no singularities and its Fourier expansion is given by

$$\Gamma^{-1}(x) = \sum_{s=-\infty}^{+\infty} \delta_s e^{\mathbf{i} s \frac{2\pi}{p} x},$$

with the corresponding Toeplitz matrix defined analogously to (3.44),

$$[\![\Gamma^{-1}]\!]^{(N)}_{js} = \delta_{j-s}, \quad j,s = -N,\dots,0,\dots,N.$$

Then, in general,

$$\left([\![\Gamma]\!]^{(N)}\right)^{-1} \neq [\![\Gamma^{-1}]\!]^{(N)}.$$

There are various mathematically well justified identities and formulas containing *infinite matrices* which can be useful here; see [7, Theorems 4.1 and 4.2, Section 3.3]. Classical treatment of the spectral theory of infinite matrices related to the mathematical foundations of the matrix formulation of quantum mechanics can be found in [14], together with extensive comments on historical developments and literature. For a comprehensive introduction to infinite Toeplitz matrices, with very valuable comments on the existing literature, see [1, Chapter 1].

Without a mathematically rigorous justification, identities valid for infinite matrices cannot (in general) be “truncated”, and then freely manipulated in further derivations, with the ambiguous argument that the obtained results hold “in the limit”. The papers by Li [6, 7] are invaluable in demonstration of possible consequences of not taking into account the fact that numerical approximations do not solve the original problem [6, p. 1876], [7, Summary, p. 133]. A rigorous clarification of the relationship between the solution of the original problem and its numerically computed approximation is an imperative, not an option which may be left aside.

We end this section rewriting (3.29) using the notation analogous to (3.44),

$$\frac{d^2 u_y^G}{dw^2} \;=\; -\left[[\![\varepsilon_r]\!]^{(N)} - Y_G^2\right] u_y^G, \quad [\![\varepsilon_r]\!]^{(N)} \equiv \Upsilon.$$

3.7. TM polarization. Here we will briefly summarize the derivation of the linear algebraic system analogous to (3.39) for the TM polarization, while pointing out subtle differences between both cases. Since TE and TM polarizations are treated separately, we can use, where appropriate, similar notations for the magnetic field in the TM polarization as for the electric field in the TE polarization without confusion.

In the superstrate and in the substrate the electric permittivity is constant. Therefore the equation (2.10) for the magnetic field H_y in the superstrate and in the substrate is in the TM polarization fully analogous to the equation (2.8) for the electric field in the TE polarization. With the incident magnetic field H_y^{inc} given by (3.2) and the Fourier expansion for $H_y(x,z)$ analogous to (3.3), the solution H_y^{I} in the superstrate and H_y^{II} in the substrate is given by the right hand sides of the identities (3.11) and (3.13). After truncation (similarly to (3.21) and (3.22) we omit the index N)

$$H_y^{\mathrm{I}} = \sum_{s=-N}^{N} u_{\mathrm{I},y}^{(s)}(z) e^{\mathbf{i}k_{xs}x}, \tag{3.45}$$

$$H_y^{\mathrm{II}} = \sum_{s=-N}^{N} u_{\mathrm{II},y}^{(s)}(z) e^{\mathbf{i}k_{xs}x}, \tag{3.46}$$

where

$$u_y^{\mathrm{I}} = [u_{\mathrm{I},y}^{(-N)}, \ldots, u_{\mathrm{I},y}^{(N)}]^T = e^{-\mathbf{i}Y_{\mathrm{I}}w} r_{\mathrm{TM}} + e^{\mathbf{i}Y_{\mathrm{I}}w} e_0, \tag{3.47}$$

$$u_y^{\mathrm{II}} = [u_{\mathrm{II},y}^{(-N)}, \ldots, u_{\mathrm{II},y}^{(N)}]^T = e^{\mathbf{i}Y_{\mathrm{II}}(w-dk_0)} t_{\mathrm{TM}}, \tag{3.48}$$

Y_{I} and Y_{II} are given by (3.24) and (3.25), respectively, and

$$r_{\mathrm{TM}} = [R_{-N}, \ldots, R_0, \ldots, R_N]^T \in \mathbb{C}^{2N+1},$$
$$t_{\mathrm{TM}} = [T_{-N}, \ldots, T_0, \ldots, T_N]^T \in \mathbb{C}^{2N+1},$$

which are, in general, different from the vectors r_{TE} and t_{TE} given by (3.23). We use in (3.48) the same scaling as in (3.13).

In order to derive the truncated approximate solution in the grating, we rewrite the equation (2.10) for $H_y(x,z)$ in the form

$$\frac{\partial^2 H_y}{\partial z^2} = -\varepsilon_r(x) \left\{ \frac{\partial}{\partial x} \left[\frac{1}{\varepsilon_r(x)} \frac{\partial H_y}{\partial x} \right] + k_0^2 H_y \right\}. \tag{3.49}$$

Now we need to substitute for H_y and $\partial H_y/\partial x$ the Fourier expansions

$$H_y(x,z) = \sum_{s=-\infty}^{+\infty} f_s(z) e^{\mathbf{i}(k_{\mathrm{I}} \sin\theta + s\frac{2\pi}{p})x}, \tag{3.50}$$

$$\frac{\partial H_y}{\partial x}(x,z) = \mathbf{i} \sum_{s=-\infty}^{+\infty} k_{xs} f_s(z) e^{\mathbf{i}(k_{\mathrm{I}} \sin\theta + s\frac{2\pi}{p})x},$$

and for $\varepsilon_r(x)$ and $1/\varepsilon_r(x)$ the expansions (3.15) and (3.16) respectively. We observe that, *in the idealized sense* (see [13, Chapter 9] and the discussion in Section 3.6) $1/\varepsilon_r(x)$ and $\partial H_y/\partial x$ are piecewise continuous with common discontinuities. Their product is continuous. From [6, Theorem 3, p. 1872], [7, Theorem 4.5, p. 123] (the nonsingularity assumptions in the statements of the theorems from [6, 7] are satisfied from the physics of the problem), under our smoothness assumptions as well as in the idealized sense, the product can be written using the *truncated inverse multiplication rule*

$$e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta} \frac{1}{\varepsilon_r(x)} \frac{\partial H_y}{\partial x} = \lim_{N\to\infty} \sum_{h=-N}^{+N} \psi_{2,h}^{(N)}(z) e^{\mathbf{i}h\frac{2\pi}{p}x},$$

where

$$\psi_{2,h}^{(N)}(z) = \mathbf{i} \sum_{s=-N}^{N} \left([\![\varepsilon_r]\!]^{(N)} \right)_{hs}^{-1} k_{xs} f_s(z). \tag{3.51}$$

ETNA
Kent State University
http://etna.math.kent.edu

Consequently,

$$\frac{1}{\varepsilon_r(x)}\frac{\partial H_y}{\partial x} = \lim_{N\to\infty}\sum_{h=-N}^{+N}\psi_{2,h}^{(N)}(z)e^{\mathbf{i}k_{x,h}x},$$

which gives

$$e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta}\frac{\partial}{\partial x}\left[\frac{1}{\varepsilon_r(x)}\frac{\partial H_y}{\partial x}\right] = \mathbf{i}\lim_{N\to\infty}\sum_{h=-N}^{N}\psi_{2,h}^{(N)}(z)k_{xh}e^{\mathbf{i}h\frac{2\pi}{p}x}. \tag{3.52}$$

The product of (the idealized discontinuous) $\varepsilon_r(x)$ with the rest of the right hand side of (3.49) is again continuous, because the left hand side of (3.49) is continuous in x. It therefore can be handled by the truncated inverse multiplication rule. Here, however, the *true Fourier amplitudes* for the function $\frac{\partial}{\partial x}\left[\frac{1}{\varepsilon_r(x)}\frac{\partial H_y}{\partial x}\right]$ are not available, and we replace them by their *truncated inverse multiplication rule approximations* $\mathbf{i}\psi_{2,h}^{(N)}(z)k_{xh}$ from (3.52), which depend on the truncation limit N,

$$e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta}\frac{\partial^2 H_y}{\partial z^2} = -\lim_{N\to\infty}\sum_{\nu=-N}^{N}\psi_{3,\nu}^{(N)}(z)\, e^{\mathbf{i}\nu\frac{2\pi}{p}x},$$

where

$$\psi_{3,\nu}^{(N)}(z) = \sum_{h=-N}^{N}\left(\left[\!\left[\frac{1}{\varepsilon_r}\right]\!\right]^{(N)}\right)^{-1}_{\nu h}(\mathbf{i}\psi_{2,h}^{(N)}(z)k_{xh} + k_0^2 f_h(z)).$$

Substituting for H_y the expansion (3.50) and for $\psi_{2,h}^{(N)}$ the expansion (3.51), we obtain after truncation

$$\frac{\partial^2 f_j(z)}{\partial z^2} = \sum_{h=-N}^{N}\left(\left[\!\left[\frac{1}{\varepsilon_r}\right]\!\right]^{(N)}\right)^{-1}_{jh}\left\{\sum_{s=-N}^{N}\left[\left([\![\varepsilon_r]\!]^{(N)}\right)^{-1}_{hs}k_{xs}k_{xh}\right]f_s(z) - k_0^2 f_h(z)\right\},$$
$$j = -N,\ldots,0,\ldots,N. \tag{3.53}$$

With the scaling $w = zk_0$ and the matrix-vector notation for the truncated expansion

$$H_y^G = \sum_{s=-N}^{N} f_s(w)e^{\mathbf{i}k_{xs}x} \equiv \sum_{s=-N}^{N} u_{G,y}^{(s)}(w)e^{\mathbf{i}k_{xs}x},$$
$$u_y^G(w) = [u_{G,y}^{(-N)}(w),\ldots,u_{G,y}^{(N)}(w)]^T \equiv [f_{-N}(w),\ldots,f_N(w)]^T,$$

the $2N+1$ differential equations in (3.53), $j = -N,\ldots,0,\ldots,N$, can be written as

$$\frac{d^2 u_y^G}{dw^2} = -Qu_y^G,$$

where

$$Q \equiv \left(\left[\!\left[\frac{1}{\varepsilon_r}\right]\!\right]^{(N)}\right)^{-1}\left[I - Y_G\left([\![\varepsilon_r]\!]^{(N)}\right)^{-1}Y_G\right] \equiv Z^{-1}[I - Y_G\Upsilon^{-1}Y_G], \tag{3.54}$$

ETNA
Kent State University
http://etna.math.kent.edu

while

$$(\Upsilon)^{-1} = \left(\llbracket \varepsilon_r \rrbracket^{(N)}\right)^{-1} \quad \text{and} \quad Z^{-1} = \left(\left\llbracket \frac{1}{\varepsilon_r} \right\rrbracket^{(N)}\right)^{-1} \tag{3.55}$$

represent, respectively, the inverse of the Toeplitz matrix (3.31), and the inverse of the Toeplitz matrix whose entries are determined by the Fourier expansion of the inverse of the relative permittivity in the grating region; see (3.16). Analogously to Υ, the matrix Z is complex symmetric. It should be noted that the inverse of a Toeplitz matrix is generally not Toeplitz. A solution to (3.54) may be given in matrix form by

$$u_y^G = e^{\mathbf{i}\sqrt{Q}w} g_{\mathrm{TM}}^+ + e^{-\mathbf{i}\sqrt{Q}(w-dk_0)} g_{\mathrm{TM}}^-,$$

where we use in the second term the same scaling as in (3.34). The square root function corresponds to the branch with the positive imaginary part. If all eigenvalues of Q are in the upper half plane, then the signal is not amplified in the direction of propagation; see the discussion in Subsection 3.4. In some experiments, however, we have observed some eigenvalues of Q also in the third quadrant, which can be considered as an artificial loss of passivity due to the discretization. A full analysis of that observation is yet to be undertaken.

We also need to find the tangential component of the electric field E_x. Using (2.11),

$$E_x = -\frac{1}{-\mathbf{i}\varepsilon_0\varepsilon_r(x)\omega + \sigma(x)} \frac{\partial H_y}{\partial z} = -\mathbf{i}\left(\frac{\mu_0}{\varepsilon_0}\right)^{1/2} \frac{1}{\varepsilon_r(x) + \mathbf{i}\sigma(x)/(\varepsilon_0\omega)} \frac{\partial H_y}{\partial w}.$$

Since $\mu_r = 1$, in the superstrate and in the substrate, apart from the thin transition regions (see [13, Chapter 9]), $\varepsilon_r + \mathbf{i}\sigma/(\varepsilon_0\omega) = n_{\mathrm{I}}^2/\mu_r = n_{\mathrm{I}}^2$ and $\varepsilon_r + \mathbf{i}\sigma/(\varepsilon_0\omega) = n_{\mathrm{II}}^2/\mu_r = n_{\mathrm{II}}^2$ respectively. Then we can immediately write the truncated approximation for E_x, using (3.45) and (3.46),

$$E_x^{\mathrm{I}} = -\mathbf{i}\left(\frac{\mu_0}{\varepsilon_0}\right)^{1/2} \frac{1}{n_{\mathrm{I}}^2} \sum_{s=-N}^{N} \frac{\partial u_{\mathrm{I},y}^{(s)}}{\partial w} e^{\mathbf{i}k_{xs}x},$$

$$E_x^{\mathrm{II}} = -\mathbf{i}\left(\frac{1}{\varepsilon_0}\right)^{1/2} \frac{1}{n_{\mathrm{II}}^2} \sum_{s=-N}^{N} \frac{\partial u_{\mathrm{II},y}^{(s)}}{\partial w} e^{\mathbf{i}k_{xs}x},$$

where the derivatives $\partial u_y^{\mathrm{I}}/\partial w$ and $\partial u_y^{\mathrm{II}}/\partial w$ are given by (3.36) and (3.37), respectively.

In the grating region the derivation requires more care. Since (the idealized) $\partial H_y/\partial w$ is continuous, substituting the Fourier expansions (3.16) and (3.50) gives

$$\begin{aligned} e^{-\mathbf{i}k_{\mathrm{I}}x\sin\theta} \frac{1}{\varepsilon_r(x)} \frac{\partial H_y}{\partial w} &= \sum_{h=-\infty}^{+\infty} a_{rh} e^{\mathbf{i}h\frac{2\pi}{p}x} \sum_{s=-\infty}^{+\infty} \frac{\partial f_s(w)}{\partial w} e^{\mathbf{i}s\frac{2\pi}{p}x} \\ &= \lim_{N\to\infty} \sum_{\nu=-N}^{N} \psi_{4,\nu}^{(N)}(w) e^{\mathbf{i}\nu\frac{2\pi}{p}x}, \end{aligned}$$

where

$$\psi_{4,\nu}^{(N)} = \sum_{s=-N}^{N} a_{r(\nu-s)} \frac{\partial f_s(w)}{\partial w}.$$

ETNA
Kent State University
http://etna.math.kent.edu

Consequently, after truncation,

$$u_y^{\mathrm{G}} = \left[\!\left[\frac{1}{\varepsilon_r}\right]\!\right]^{(N)} \left\{ \mathbf{i}\sqrt{Q} e^{\mathbf{i}\sqrt{Q}w} g_{\mathrm{TM}}^{+} - \mathbf{i}\sqrt{Q} e^{-\mathbf{i}\sqrt{Q}(w-dk_0)} g_{\mathrm{TM}}^{-} \right\}, \tag{3.56}$$

$$E_x^{\mathrm{G}} = -\mathbf{i}\left(\frac{\mu_0}{\varepsilon_0}\right)^{1/2} \sum_{s=-N}^{N} u_{G,y}^{(s)} e^{\mathbf{i}k_{xs}x}.$$

Finally, writing (similarly as in the TE polarization above) the boundary matching conditions

$$-H_y^{\mathrm{I}}(x,0) + H_y^{\mathrm{G}}(x,0) = 0, \quad -E_x^{\mathrm{I}}(x,0) + E_x^{\mathrm{G}}(x,0) = 0$$

at $z = 0$, and

$$+H_y^{\mathrm{G}}(x,d) - H_y^{\mathrm{II}}(x,d) = 0, \quad +E_x^{\mathrm{G}}(x,d) - E_x^{\mathrm{II}}(x,d) = 0$$

at $z = d$, into one matrix equation for the unknown integrating constants r_{TM}, g_{TM}^{+}, g_{TM}^{-}, and t_{TM}, gives the $4(2N+1) \times 4(2N+1)$ linear algebraic system similar to (3.39), which, recalling $\left[\!\left[\frac{1}{\varepsilon_r}\right]\!\right]^{(N)} \equiv Z$ and multiplying the second and the fourth equation by $-\mathbf{i}$, can be written as

$$\begin{bmatrix} -I & I & e^{\mathbf{i}\sqrt{Q}dk_0} & 0 \\ \frac{1}{n_{\mathrm{I}}^2}Y_{\mathrm{I}} & Z\sqrt{Q} & -Z\sqrt{Q}e^{\mathbf{i}\sqrt{Q}dk_0} & 0 \\ 0 & e^{\mathbf{i}\sqrt{Q}dk_0} & I & -I \\ 0 & Z\sqrt{Q}e^{\mathbf{i}\sqrt{Q}dk_0} & -Z\sqrt{Q} & -\frac{1}{n_{\mathrm{II}}^2}Y_{\mathrm{II}} \end{bmatrix} \begin{bmatrix} r_{\mathrm{TM}} \\ g_{\mathrm{TM}}^{+} \\ g_{\mathrm{TM}}^{-} \\ t_{\mathrm{TM}} \end{bmatrix} = \begin{bmatrix} e_0 \\ \frac{\cos\theta}{n_{\mathrm{I}}}e_0 \\ 0 \\ 0 \end{bmatrix}, \tag{3.57}$$

and will be denoted in the following as

$$A_{\mathrm{TM}}\xi_{\mathrm{TM}} = b_{\mathrm{TM}}.$$

As in the TE polarization, in practical measurements one typically needs only the zeroth order mode of r_{TM},

$$r_{0,\mathrm{TM}} = (e_0^T, 0)\xi_{\mathrm{TM}} = e_0^T r_{\mathrm{TM}}.$$

If we use the unscaled blocks of unknowns

$$\widehat{t}_{\mathrm{TM}} = e^{-\mathbf{i}Y_{\mathrm{II}}dk_0} t_{\mathrm{TM}}, \quad \widehat{g}_{\mathrm{TM}}^{-} = e^{\mathbf{i}\sqrt{Q}dk_0} g_{\mathrm{TM}}^{-},$$

the last two columns of the matrix of the linear algebraic system (3.57) must be scaled accordingly.

3.8. Numerical experiments. In this section we present some of the results obtained with the RCWA method. Our aim is to illustrate the numerical behavior of the method using a representative example, and not necessarily to present an overview of the efficiency of the method. Nevertheless, the importance of the issue of efficiency of the numerical computations will be apparent, and will motivate the following sections which will close the paper.

In our experiment, we apply RCWA to the problem of computing the zeroth order reflection coefficient from a rectangular grating, such as the one depicted in Figure 2.1. This experiment has its basis in the semiconductor industry, where optical instruments measure

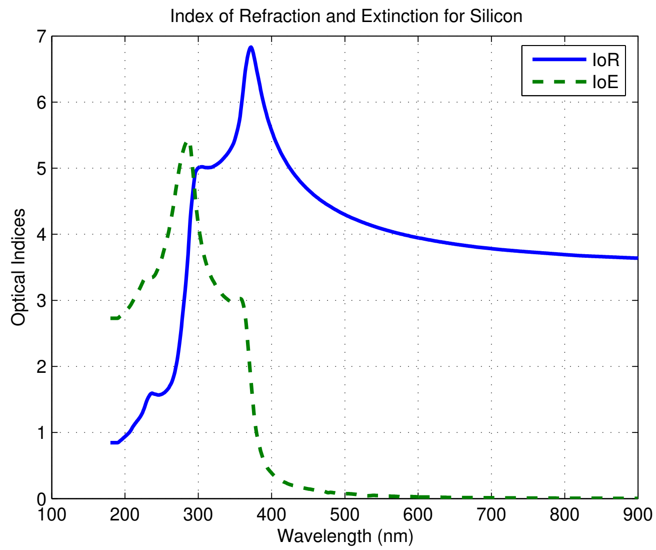

FIGURE 3.3. *Indices of refraction and extinction for silicon.*

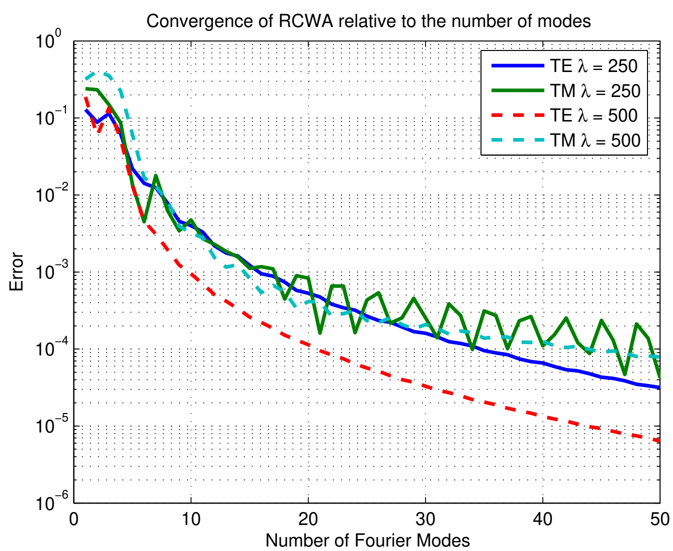

FIGURE 3.4. *Approximation error as a function of the number of Fourier modes.*

the reflection coefficients from periodic structures on silicon wafers, and through an inverse problem, determine the geometry of the measured feature.

In this simple example, the substrate is silicon and the superstrate is air. The material for the substrate is chosen not merely because of its importance in the semiconductor industry, but also because the material exhibits a number of interesting properties. First, it has a very high index of refraction relative to most materials; for example, at a wavelength of 500 nm the index of refraction is over 4. Compare this to the index of refraction of glass, which is ap-

ETNA
Kent State University
http://etna.math.kent.edu

proximately 1.5. Second, at wavelengths in the ultraviolet region (below 280 nm) the index of extinction (the imaginary part of the complex index of refraction) dominates, with the material behaving more like a conductor than a dielectric. To illustrate the behavior of the electromagnetic fields for these two different regimes, we compute a solution to Maxwell's equation via RCWA at two wavelengths, 250 and 500 nm. For these wavelengths, the complex indices of refraction for silicon have been determined experimentally, with $n_{\mathrm{II}} = 1.580 + 3.632\mathbf{i}$ and $n_{\mathrm{II}} = 4.2975 + 0.07297\mathbf{i}$ at wavelengths of 250 and 500 nm, respectively. For reference, a plot of the indices of refraction and extinction for silicon as a function of wavelength is depicted in Figure 3.3.

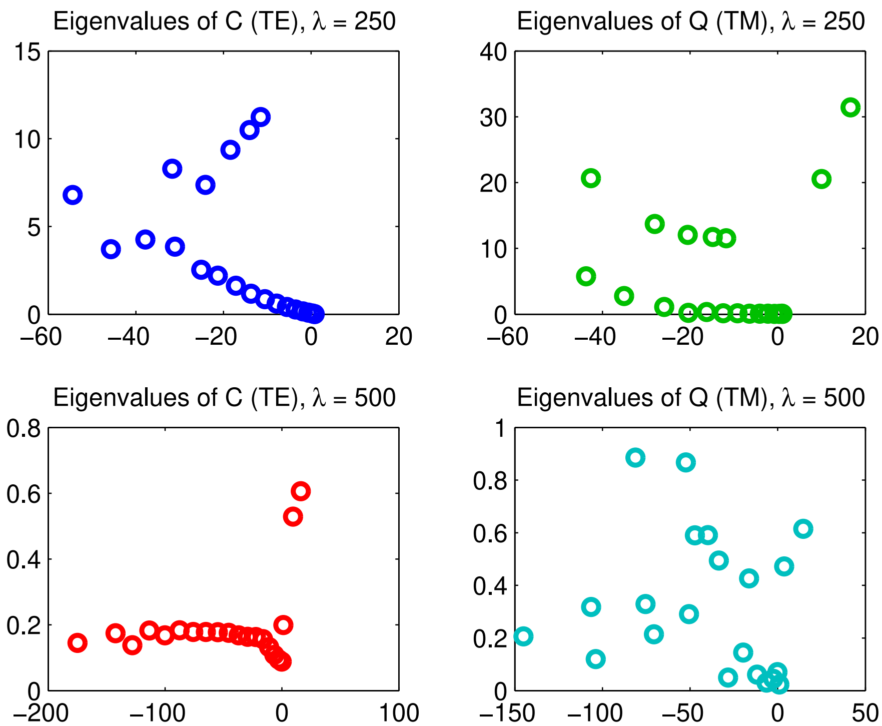

FIGURE 3.5. *Eigenvalues of the matrices C and Q.*

The values of the parameters used to describe the rectangular grating are as follows: the period $p = 400$, space fraction $q = p_t/p = 125/400$, and the height $d = 300$. All geometrical distances are in nm. As it happens, these values are also representative of the measurement targets one might find in the semiconductor industry. The incidence angle for all experiments in this paper is $\theta = 70$ degrees.

Figure 3.4 compares the convergence of the zeroth order reflection coefficient R_0 for the TE and TM modes of both wavelengths as a function of the number of Fourier modes used to compute the fields. The approximation error is computed as the modulus of the difference between the zeroth order reflection coefficient and that which is computed for 100 Fourier modes. Note that the convergence is faster for the TE than the TM modes for both wavelengths, that the convergence of the TE method is faster for silicon in its dielectric regime ($\lambda = 500\ nm$) than in its more metallic regime ($\lambda = 250\ nm$), and that the convergence

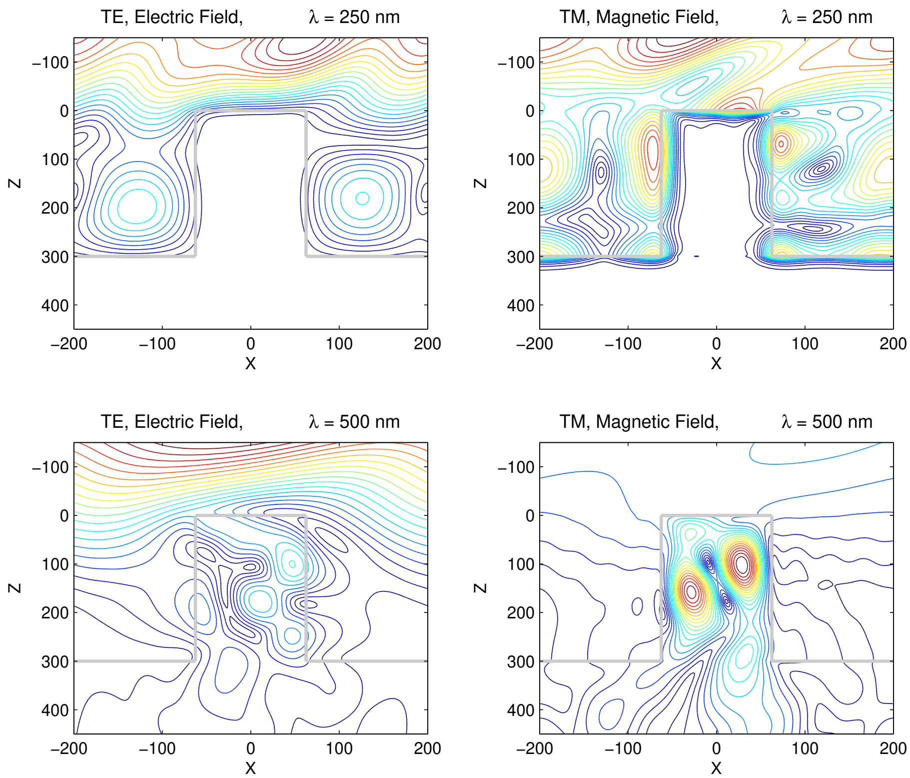

FIGURE 3.6. *Contour plots of the electric and magnetic fields.*

of the TM method is more monotonic for silicon in its dielectric regime than in its more metallic regime. Finally, note that the solution converges quickly to the relative accuracy of about 10^{-3}–10^{-4} for relatively few Fourier modes. This property is of particular interest in the semiconductor industry, due to the importance of the speed of the solution. Any greater accuracy of the solution is superfluous due to the measurement precision of its instruments.

Figure 3.5 plots the complex eigenvalues of the system matrices and for both wavelengths. We note that, as expected, the eigenvalues are in the upper half plane. It has been observed, however, that for cases with materials with large index of extinction, some of the eigenvalues of Q can drift into the third quadrant of the complex plane, which causes difficulties in physical interpretation of the computed solution described above. Interestingly, we have not yet encountered a situation in which the eigenvalues of C fall outside the upper half-plane, or when any eigenvalue of Q falls within the fourth quadrant.

Finally, we plot in Figure 3.6 a contour map of the transverse electric and magnetic fields for the wavelengths $\lambda = 250\ nm$ and $\lambda = 500\ nm$. The fields are computed with the field expansion truncated to 10 Fourier modes. Let us point out a few features in these plots. First, note that the fields hardly penetrate the silicon structure at the wavelength $\lambda = 250\ nm$, which demonstrates the property of finite skin-depth for conductive materials [13]. Second, note that at wavelength $\lambda = 500\ nm$, the magnitude of the magnetic field in the dielectric region is much higher than that of the surrounding air. This is a consequence of the high index

ETNA
Kent State University
http://etna.math.kent.edu

of refraction of silicon at this wavelength. Third, note that the contour lines in the TE mode are smooth across the material boundaries (shown in light grey), while for the TM mode the contour lines are almost discontinuous. This is due to the continuity properties of the TE and TM fields across material boundaries, with the TE field being smooth and the TM field being almost discontinuous. Finally, note the wavy nature of the contour line for the TM field at coordinate position $x \approx 150$ and $z \approx 50$ for the wavelength $\lambda = 500\, nm$. This is an artifact of the Fourier decomposition. This feature gradually disappears as the number of Fourier modes kept in the field expansion grows larger.

4. Open problems in the analysis of the RCWA method. As is stated in [6], the process of discretization in RCWA presented here can lead to some nontrivial ambiguities. Some of the approaches found in the literature are not well-justified mathematically, and have the potential of yielding incorrect results without proper analysis. While we have dealt with many of these issues here, a number of issues remain open, which we list below.

One issue is that of the smoothness of the permittivity function within the grating region. As mentioned, we have presumed for the sake of the derivation that it is smooth; however, in the literature it is given an idealized mathematical description as discontinuous at the interface between two distinct materials. The discontinuity idealization has been treated in the physics literature by referring to the integral form of Maxwell's equation and taking appropriate limits; cf. [13, Chapter 9]. Other means might be through the periodic convolution of the permittivity function with a distribution which in an appropriate limit becomes the Dirac delta function. A more detailed treatment of the discontinuity of the fields and permittivities is left, however, as an open problem.

The reduction of the problem from a countably infinite set of ordinary differential equation to a finite set yields the problem of how to formally multiply the two series and take their truncations. Hardy [3, Chapter X] provides a set of formal rules, and discusses their convergence properties. The applicability of these rules to RCWA remains an open problem.

Another issue that is not addressed is the issue of the possible additional truncations. In standard RCWA, if the fields are truncated to order N (consisting of $2N + 1$ components), Fourier modes up to order $2N$ (consisting of $4N + 1$ components) are used in the matrices Υ and Z. As is pointed out in [11, Appendix A], this inconsistency between the number of components for the fields and for the permittivity function is a consequence of the representation of the truncated problem. It is clear, for example, that fewer modes for the permittivity could be used, reducing the matrices Υ and Z to banded matrices. Taking a cue from signal processing literature, it might be advantageous for reasons of convergence or conditioning to multiply the Fourier components of the permittivity function by a suitable windowing function. Again, such approaches and their analysis remain open.

We have not addressed here the systematic treatment of loss of passivity that one can observe in the solution of the truncated problem, e.g., the loss of passivity associated with the eigenvalues of the matrices C and Q that lie in the third quadrant. The standard approach in RCWA enforces a type of passivity with the eigenvalues of $\sqrt{C}$ and $\sqrt{Q}$ lying in the upper half-plane. While this ensures that the matrix exponentials in (3.38) and (3.56) remain bounded as w becomes large, it has the consequence of mixing waves with different propagation directions, as those eigenmodes associated with third quadrant eigenvalues of C and Q have a different propagation direction than those associated with first and second quadrant eigenvalues of C and Q. While this seems to produce an acceptable solution of the numerical problem, a complete analysis of this issue and its physical interpretation is yet to be undertaken.

5. Perspectives of algebraic computations within RCWA. The paper presents, within our abilities, a mathematically justified derivation of the RCWA discretized approximation to

ETNA
Kent State University
http://etna.math.kent.edu

the problem of light diffraction on a simple rectangular grating. Without a mathematically correct derivation of the truncated approximation there is a possibility that the discretized problem may not be formed correctly. From a practical point of view, the message is that although the discretization can be motivated intuitively or empirically, its justification requires mathematical rigor. Indeed, it has been observed in practice that an intuitive derivation can fail. Therefore a step-by-step detailed mathematical examination of methods used in practical computations is useful.

The solution of Maxwell's equation by the RCWA method on a simple rectangular grating given here forms the basis for computing the solution for *more complicated shapes*. To extend RCWA to these shapes, it is necessary to approximate the original shape by a set of vertical regions, or slabs, within which the permittivity is constant as a function of height. For example, a trapezoid is approximated by a shape in the form of a ziggurat. This approximation requires the solution of boundary conditions at the interface of each slab, which results in a system matrix akin to (3.39) or (3.57), i.e., block tridiagonal with the number of diagonal blocks proportional to the number of slabs; see, e.g., [2]. Typically, a trade-off must be made in the approximation of the shape by slabs. A good geometrical approximation of the shape of the grating may be obtained with many slabs of small height. On the other hand, using such a large number of slabs makes the computation of integrating constants more demanding.

The dominance of RCWA in the field of scatterometry has been attributed to two factors. First, RCWA has been shown to be remarkably robust: it is able to reliably compute the reflection coefficients for the wide range of wavelengths, for arbitrary shapes and incidence angles. Second, it is able to compute the reflection coefficients to a relative accuracy of about 10^{-4} in a reasonably short time. This is crucial as the industrial application is that of an inverse problem, whereby the reflection coefficients as functions of the wavelengths (called the reflection spectra) computed for a parameterized structure are matched to the measured information. The time allotted for the solution of the inverse problem is determined by the hardware, typically between 2 and 10 seconds. In one instantiation of the problem, the matching is performed by an on-line optimization algorithm, which takes many steps to converge and requires in input not only the measured and computed reflection spectra, but also its Jacobian, i.e., the derivative of the computed reflection spectra as a function of the parameter vector. The Jacobian is typically approximated using finite differences. Since the number of optimization parameters is usually between 5 and 10, the number of wavelengths used in the reflection spectra computation is approximately 100, and the number of steps for convergence is between 10 and 20, the solution of the inverse problems requires the computation of a number of individual reflection coefficient of the order of 10,000. Even when one accounts for the parallelizability of the problem, the need for computational efficiency is clear.

From this setting a number of interesting problems in numerical linear algebra arise:

- One problem is computing the solution of the linear system which yields the Fourier components of the fields. Two general approaches can be taken. One is the solution a linear system consisting of the block tridiagonal matrix as is written in this paper, another uses a method of scattering matrix propagation [9, Section III.6]. Typically, the latter method is used in industry and is equivalent to a sequential block elimination algorithm.
- Another problem is the issue of efficiently computing the function of matrices, such as the matrix exponential or square root, which is necessary to fill the blocks in the system matrix. Since the matrix functions are computed over the domain of highly structured matrices, there is the possibility of computing these matrix functions more efficiently than for an arbitrary dense matrix.
- A third issue is that of providing an a priori estimate for the number of slabs and/or

number of Fourier modes required to achieve a given accuracy. This remains an open problem. Also, the issue of the convergence of the eigenvalues of the matrices C and Q, (3.29) and (3.54), respectively, as a function of the number of Fourier modes is also not well understood.

- Another issue that arises is related to the computation of solutions for gratings with highly conductive materials. In such as case the matrices Υ and Z, (3.31) and (3.55), respectively, can be ill-conditioned with respect to inversion. Recently this phenomenon was studied and attributed to spurious eigenvalues with small magnitudes; see [10]. A suitable regularization method has yet to be devised.
- The most pressing problem facing the industrial use of RCWA is the *speed of the solution* for three-dimensional (doubly periodic) structures. In these structures an arbitrary two dimensional shape is tiled on the x-y plane, and requires a two-dimensional Fourier decomposition of the permittivity and the fields. In such a case, the system matrices become much larger, as the dimensionality of the field vectors scale as the *product of the number of retained Fourier components* in x and y. Thus, techniques for improving the speed of the solution for the two-dimensional (singly periodic) problems become essential for three-dimensional problems.

We close our discussion with the note that approximation of the solution of the linear systems $A_{\rm TE}\xi_{\rm TE} = b_{\rm TE}$ and $A_{\rm TM}\xi_{\rm TM} = b_{\rm TM}$ has in RCWA a very particular meaning. We need to approximate only one element of the solution vector, namely, the one which is associated with the zeroth diffraction order. This indicates that there may be a number of suitable fast iterative methods to find that element with sufficient accuracy. Recent results [12] indicate that this quantity can be computed to a given level of accuracy considerably faster by a moment-matching method, than by explicitly computing the solution of the linear system.

Acknowledgment. The authors are indebted to Petr Tichý for several useful comments and to Oliver Ernst for numerous suggestions which have led to a significant improvement of the exposition.

REFERENCES

[1] A. Böttcher and S. M. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM, Philadelphia, 2005.
[2] A. David, H. Benisty, and C. Weisbuch, *Fast factorization rule and plane-wave expansion method for two-dimensional photonic crystals with arbitrary hole-shape*, Phys. Rev. B, 73 (2006), pp. 1–7.
[3] G. H. Hardy, *Divergent Series*, Clarendon Press, Oxford, 1949.
[4] N. J. Higham, *Functions of Matrices*, SIAM, Philadelphia, 2008.
[5] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.
[6] L. Li, *Use of Fourier series in the analysis of discontinuous periodic structures*, J. Opt. Soc. Amer. A, 13 (1996), pp. 1870–1876.
[7] L. Li, *Mathematical reflections on the Fourier modal method in grating theory*, in Mathematical Modeling in Optical Science, vol. 22 of Frontiers Appl. Math., SIAM, Philadelphia, 2001, pp. 111–139.
[8] M. G. Moharam, E. B. Grann, and D. A. Pommet, *Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings*, J. Opt. Soc. Amer. A, 12 (1995), pp. 1068–1076.
[9] M. Neviere and E. Popov, *Light Propagation in Periodic Media*, Marcel Dekker Ltd., New York, 2002.
[10] E. Popov, B. Chernov, and M. Neviere, *Differential theory: application to highly conducting gratings*, J. Opt. Soc. Amer. A, 21 (2004), pp. 199–206.
[11] E. Popov and M. Neviere, *Grating theory: new equations in Fourier space leading to fast converging results for TM polarization*, J. Opt. Soc. Amer. A, 17 (2000), pp. 1773–1784.
[12] Z. Strakoš and P. Tichý, *On efficient numerical approximation of the scattering amplitude $c^*A^{-1}b$ via matching moments*, (2009). Submitted.
[13] R. K. Wangsness, *Electromagnetic Fields*, J. Willey and Sons, New York, 1986.
[14] A. Wintner, *Spektraltheorie der Unendlichen Matrizen*, Verlag von S. Hirzel, Leipzig, 1929.