Diffraction électromagnétique par des réseaux et des surfaces rugueuses aléatoires : mise en œuvre deméthodes hautement efficaces pour la résolution de systèmes aux valeurs propres et de problèmesaux conditions initiales

Cihui Pan

NNT : 2015SACLV020

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A L'UNIVERSITE VERSAILLES SAINT-QUENTIN EN YVELINES

ÉCOLE DOCTORALE N°580
Sciences et technologies de l'information et de la communication
GRADUATE SCHOOL of STIC

Spécialité de doctorat : Informatique

Par

M. Cihui Pan

Electromagnetic scattering by gratings and random rough surfaces:
Implementation of high performance algorithms for solving eigenvalue problem and problems with initial conditions

Thèse présentée et soutenue à Saclay, le 2 Décembre 2015 :

Composition du Jury :

Mme V. Ciarletti, Professeur, LATMOS, UVSQ, Présidente du Jury
M. J. Demmel, Professeur, University of California, Berkeley, Rapporteur
M. D. Prémel, Recherche/HDR, CEA, Rapporteur
M. C. Bourlier, Directeur de recherche, CNRS, IEET, Nantes Examinateur
Mme V. Ciarletti, Professeur, LATMOS, UVSQ, Examinatrice
M. P. Ricoux, Directeur scientifique à TOTAL, Examinateur
Mme N. Emad, Professeur, PRiSM, MDLS, UVSQ, Directrice de thèse
M. R. Dusséaux, Professeur, LATMOS, UVSQ, Directeur de thèse

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Diffraction électromagnétique par des réseaux et des surfaces rugueuses aléatoires: Mise en œuvre de méthodes hautement efficaces pour la résolution de systèmes aux valeurs propres et de problèmes aux conditions initiales

Mots clés : Physique des ondes, Méthodes numériques, Calcul matriciel, Calcul parallèl et distribué à grande échelle

Résumé : Nous étudions la diffraction électromagnétique par des réseaux et des surfaces rugueuses aléatoire. Le méthode C est une méthode exacte développée pour ce but. Elle est basé sur des équations de Maxwell sous forme covariante écrites dans un système de coordonnées non orthogonales. Le méthode C conduisent à une matrice de diffusion dont il faut déterminer les valeurs propres.
Nous nous concentrons sur l'aspect numérique de la méthode C, en développant une mise en œuvre efficace de cette méthode exacte. En définissant un nouveau système de coordonnées non orthogonales, nous établissons une formation qui évite la résolutions d'un système aux valeur propres. Pour les réseaux de diffraction à une interface, nous montrons que cette nouvelle version de la méthode C conduit à un système différentiel avec les conditions initiales. Nous montrons que cette nouvelle version de la méthode C peut être utilisée pour l'étude des réseaux comme un empilement d'interfaces délimitant des couches homogènes.
Nous proposons un algorithme QR parallèle conçu spécifiquement pour la méthode C pour résoudre le problème de valeur propre. En perspective, nous proposons une méthode de projection spectrale pour résoudre le problème de valeurs propres efficacement. Cetee méthode proposeé afin de palier au problème de scalability de la méthode QR.
.

Title : Electromagnetic scattering by gratings and random rough surfaces: Implementation of high performance algorithms for solving eigenvalue problem and problems with initial conditions.

Keywords : Physics of waves, Numerical methods, Matrix computation, Large-scale parallel and distributed computing

Abstract : We study the electromagnetic diffraction by gratings and random rough surfaces. The C-method is an exact method for this aim. It is based on Maxwell's equations under covariant form written in a non-orthogonal coordinate system. The C-method leads to an eigenvalue problem, the solution of which gives the diffracted field.
We focus on the numerical aspect of the C-method, trying to develop an efficient application of the exact method. For gratings, we have developed a new version of the C-method which leads to a differential system with inital conditions. This new version of the C-method can be used to study multilayer gratings with a homogeneous médium.
We implmented high performance algorithms for the original version of the C-method. Especially, we have developed a specifically designed parallel QR algorithm for the C-method and spectral projection method to solve the eigenvalue problem more efficiently.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisors Prof. Richard Dusséaux and Prof. Nahid Emad for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. Their guidances helped me in all the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my Ph.D study.

Besides my advisors, I would like to thank the reviewers of my thesis: Porf. James Demmel and Dr. Denis Prémel, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives. I would like also to thank the rest of my thesis committee: Dr. Christophe Bourlier, Dr. Philippe Ricoux and Prof. Valérie Ciarletti.

I thank my fellow labmates in for the stimulating discussions, for all the fun we have had in the last four years.

Last but not the least, I would like to thank my family: my parents for supporting me spiritually throughout writing this thesis and my life in general.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Electromagnetic wave scattering is an active and interdisciplinary area of research with myriad practical applications in fields ranging from atomic physics to optics, medical imaging, geoscience and remote sensing [1–6]. In particular, the subject of wave scattering by gratings [5–12] and rough surfaces [13–25] presents great theoretical challenges due to the large number of degrees of freedom in these systems and a need to include multiple scattering effects accurately. In the past several decades, considerable theoretical progress has been made in elucidating and understanding the scattering processes involved in such problems. Diagrammatic techniques and effective medium theories remain essential for analytical studies; however, rapid advances in computer technology have opened new doors for researchers with the full power of Monte Carlo simulations in the numerical analysis of random media scattering [18–25]. Numerical simulations allow us to solve Maxwell's equations without the limitations of analytical approximations, whose regimes of validity are often difficult to assess [13–17, 26].

In this thesis, we study the electromagnetic diffraction by gratings and random rough surfaces. The C-method is an exact method developed for this aim. It is based on Maxwells equations under covariant form written in a nonorthogonal coordinate system [27–29]. Discretizing the Maxwells equations under the non-orthogonal coordinate system and separating variables lead to solving the eigenvalue problem of the high dimension, dense and non-symmetric scattering matrix. All the eigenvalues and eigenvectors of the scattering matrix are needed. The scattered field is expanded as a linear combination of eigensolutions satisfying the outgoing wave condition. The boundary conditions allow

the diffraction amplitudes to be determined. This method has been used for analyzing gratings used in optics [30–50] ,waveguides [51–54] and rough surfaces [55–65]

We focus on the numerical aspect of the C-method, trying to develop an efficient implementation of this exact method. Iterative eigensolvers, such as Krylov subspace methods or Jacobi-Davidson methods [66] have been developed to deal with large-scale eigenvalue problems. However, they have the possibility of missing some eigenvalues. So the standard iterative methods are ineffective for the C-method because all the eigenvalues and eigenvectors are needed. In contrast, the QR algorithm, which is based on similarity transformations, calculates all the eigenvalues and eigenvectors with very little danger and only with a warning of missing some eigensolutions. We propose a specifically designed parallel QR algorithm for the C-method to solve the eigenvalue problem.

This parallel QR algorithm is a variant of QR algorithm based on three techniques: early shift, parallel bulge chasing [67, 68] and parallel aggressive early deflation (AED) [68, 69]. We propose the "early shift" for the scattering matrix according to the property we have observed. That is the C-method and the physical interpretation behind the C-method provides very good approximations of some eigenvalues before any calculations. The "early shift" provides the possibility of quick deflation. We mixed the "early shift", Wilkinson's shift and exceptional shift together to accelerate the convergence. Especially, we use the "early shift" to have quick deflation of the approximated eigenvalues of the scattering matrix. They provide the possibility of quick deflation. For the bulge chasing, instead of only a single bulge, containing two shifts, a chain of several tightly coupled bulges, each containing two shifts, is chased in the course of one multishift QR iteration. This idea and the delay-and-accumulate technique [67, 68] allow performing most of the computational work in terms of matrix-matrix multiplications to benefit from level 3 Basic Linear Algebra Subprograms (BLAS, the level 3 contains matrix-matrix operations) [70]. Aggressive early deflation is a QR algorithm deflation strategy that takes advantage of matrix perturbations outside of the subdiagonal entries of the Hessenberg QR iterate. It identifies and deflates converged eigenvalues long before the classic small-subdiagonal strategy would. Aggressive early deflation can significantly enhance the convergence of the QR algorithm.

We also propose a spectral projection method to solve the eigenvalue problem efficiently. We propose a global eigensolver by a combination of the SS method (Sakuria and Sugiura

method, proposed by Sakuria and Sugiura in [71]) and MIRAMns (Multiple Implicitly Restarted Arnolid Method with nested subspaces, proposed by S. A. Shahzadeh Fazeli et al. in [72]). This proposed global eigensolver allows us to calculate a large number (or all) of the eigenvalues of a general matrix. According to our experiments presented in chapter 7, this method has the advantage of having very good scalability compared to the QR algorithm. This promising method can be continued in future work.

The original C-method is not very efficient when we are dealing with multilayer gratings. We want to find other solutions. Especially, we want to explore the potential parallelization of multilayer gratings. So, we propose a new version of C-method which leads to a differential system with initial conditions for gratings. We show that this new version of C-method can be used to study multilayer gratings with homogeneous medium. We show that this formulation is an interesting tool for analyzing perfectly conducting or dielectric gratings with deep grooves. The proposed method allows analyzing the complex phenomenon of incident energy absorption. We apply this method to multilayer gratings with an arbitrary number of interfaces. We show how to combine the local scattering matrix to obtain the global one. We validate our method by comparing experiments results with that from published paper. We show that this new version of C-method has very good accuracy as well as a nature of two level parallel property [73, 74]. This new version of C-method is an attractive alternative to analyze multilayered grating having parallel or non-parallel interfaces.

During my Ph.D study, I worked in three laboratories: Laboratoire Atmosphères, Milieux, Observations Spatiales (LATMOS), Maison de la Simulation (MDLS) and Laboratoire Parallélisme, Réseaux, Systèmes, Modélisation (PRiSM). LATMOS is a laboratory specializing on the fundamental physical and chemical processes of atmospheres, environments, and spatial observations. MDLS is a laboratory specializing on scientific computing and simulations using HPC. PRiSM is a laboratory specializing on computer science.

This thesis is structured as follows. In chapter 2, we present the electromagnetic field theory fundamentals. In chapter 3, we present the scattering problem by gratings and by random rough surfaces. In chapter 4, we present the curvilinear coordinate method. In chapter 5, we propose the parallel QR algorithm for the C-method. In chapter 6, we present an implementation of our parallel QR algorithm and present the results of the

numerical experiments. In chapter 7, we propose an alternative to QR algorithm for solving the eigenvalue problem. In chapter 8, we propose a new version of C-method, which is the C-method as an initial value problem.

Chapter 2

Electromagnetic field theory fundamentals

2.1 Maxwell's equations

The general theory of electromagnetic phenomena is based on Maxwell's equations, which constitute a set of four coupled first order vector partial differential equations relating the space and time changes of electric and magnetic fields to their scalar source densities (divergence) and vector source densities (curl) [1]. Maxwell's equations are usually formulated in differential form (i.e., as relationships between quantities at the same point in space and at the same instant in time) or in integral form where, at a given instant, the relations of the fields with their source are considered over an extensive region of space [1]. The two formulations are related by the divergence and Stokes' theorems.

For stationary media, Maxwell's equations in differential and integral forms are:

Differential form of Maxwell's equation

$$\nabla \cdot \vec{D}(\vec{r},t) = \rho(\vec{r},t) \quad (Gauss' \ law) \tag{2.1}$$

$$\nabla \cdot \vec{B}(\vec{r},t) = 0 \quad (Gauss' \ law \ for \ magnetic \ fields) \tag{2.2}$$

$$\nabla \times \vec{E}(\vec{r},t) = -\frac{\partial \vec{B}(\vec{r},t)}{\partial t} \quad (Faraday's \ law) \tag{2.3}$$

$$\nabla \times \vec{H}(\vec{r},t) = \vec{J}(\vec{r},t) + \frac{\partial \vec{D}(\vec{r},t)}{\partial t} \quad (Generalized \ Ampere's \ law) \tag{2.4}$$

where ∇ is the Nabla operator, $\nabla \cdot \vec{D}$ designates the divergence of $\vec{D}$ and $\nabla \times \vec{E}$, the curl of $\vec{E}$.

Integral form of Maxwell's equations

$$\oint_S \vec{D}(\vec{r},t) \cdot d\vec{s} = Q_T(t) \quad (Gauss'\ law) \tag{2.5}$$

$$\oint_S \vec{B}(\vec{r},t) \cdot d\vec{s} = 0 \quad (Gauss'\ law\ for\ magnetic\ fields) \tag{2.6}$$

$$\oint_\Gamma \vec{E}(\vec{r},t) \cdot d\vec{l} = -\int_S \frac{\partial \vec{B}(\vec{r},t)}{\partial t} \cdot d\vec{s} \quad (Faraday's\ law) \tag{2.7}$$

$$\oint_\Gamma \vec{H}(\vec{r},t) \cdot d\vec{l} = \int_S (\vec{J}(\vec{r},t) + \frac{\partial \vec{D}(\vec{r},t)}{\partial t}) \cdot d\vec{s} \quad (Generalized\ Ampere's\ law) \tag{2.8}$$

where S is any fixed open surface and Γ is the associated boundary curve.

Maxwell's equations, involve only macroscopic electromagnetic fields and, explicitly, only macroscopic densities of free-charge, $\rho(\vec{r},t)$, which are free to move within the medium, giving rise to the free-current densities, $\vec{J}(\vec{r},t)$. The effect of the macroscopic charges and current densities bound to the medium's molecules is implicitly included in the auxiliary magnitudes $\vec{D}$ and $\vec{H}$ which are related to the electric and magnetic fields, $\vec{E}$ and $\vec{B}$ by the so-called constitutive equations that describe the behavior of the medium. In general, the quantities in these equations are arbitrary functions of the position $\vec{r}$ and the time t. The definition and units of these quantities are:

$$\vec{E} = electric\ field\ intensity \quad (volt/meter, V \cdot m^{-1})$$

$$\vec{B} = magnetic\ flux\ density \quad (weber/square\ meter, Wb \cdot m^{-2})$$

$$\vec{D} = electric\ flux\ density \quad (coulomb/square\ meter, C \cdot m^{-2})$$

$$\vec{H} = magnetic\ field\ density \quad (ampere/meter, A \cdot m^{-1})$$

$$\rho = free\ electric\ charge\ density \quad (coulomb/cubic\ meter, C \cdot m^{-3})$$

$$Q_T = net\ free\ charge\ inside\ any\ closed\ surface\ S \quad (coulomb, C)$$

$$\vec{J} = free\ electric\ current\ density \quad (ampere/square\ meter, A \cdot m^{-2})$$

The equations (2.1)-(2.4) or (2.5)-(2.8) as a whole are associated with the name of Maxwell's equations because he was responsible for completing them, adding to Ampere's original equation, $\nabla \times \vec{H}(\vec{r},t) = \vec{J}(\vec{r},t)$, the displacement current density term or, in short, the displacement current, $\frac{\partial \vec{D}}{\partial t}$, as an additional vector source for the field

$\vec{H}$. This term has the same dimensions as the free current density but its nature is different because no free charge movement is involved. Its inclusion in Maxwell's equation is fundamental to predict the existence of electromagnetic wave which can propagate through empty space at the constant velocity of light c. The concept of displacement is also fundamental to deduce from equation (2.4) the principle of charge conservation by means of the continuity equation:

$$\nabla \cdot \vec{J} = -\frac{\partial \rho}{\partial t} \tag{2.9}$$

or, in integral form:

$$\oint \vec{J} \cdot d\vec{s} = -\frac{dQ_t}{dt} \tag{2.10}$$

With these equations, Maxwell showed not only that the electric and magnetic fields are interrelated but also that they are in fact two aspects of a single concept, the electromagnetic field.

Maxwell's equations together with the Lorenz's force constitute the basic mathematical formulation of the physical laws that at a macroscopic level explain and predict all the electromagnetic phenomena which basically comprise the remote interaction of charges and currents taking place via the electric and/or magnetic fields that they produce.

In applications, Maxwell's equations have to be complemented by appropriate initial and boundary conditions. The initial conditions involve values or derivatives of the fields at $t = 0$, while the boundary conditions involve the values or derivatives of the fields on the boundary of the spatial region of interest. Usually, we consider the initial conditions as a form of boundary conditions and refer to the solution of Maxwell's equation, with all these conditions, as a boundary-value problem.

Next, we briefly describe the physical meaning of Maxwell's equations.

Gauss' law is a direct mathematical consequence of Coulomb's law, which states that the interaction force between electric charges depends on the distance r, between them, as r^{-2}. According to Gauss's law, the divergence of the vector field $\vec{D}$ is the volume density of free electric charges which are the sources or sinks of the field $\vec{D}$, i.e. the lines of $\vec{D}$ begin on positive charges ($\rho > 0$) and end on negative ($\rho < 0$). In its integral form,

Gauss' law relates the flux of the vector $\vec{D}$ through a closed surface S to the total free charge within that surface.

Gauss' law for magnetic fields states that the $\vec{B}$ field does not have scalar sources, i.e., it is divergenceless or solenoidal. This is because no free magnetic charges or monopoles have been found in nature which would be the magnetic analogues of electric charges for $\vec{E}$. Hence, there are no sources or sinks where the field lines of $\vec{B}$ start or finish, i.e., the field lines of $\vec{B}$ are closed. In its integral form, this fact indicates that the flux of the $\vec{B}$ field through any closed surface S is null.

Faraday's law establishes that a time-varying $\vec{B}$ field produces a non conservative electric field whose field lines are closed. In its integral form, Faraday's law states that the time variation of the magnetic flux ($\int \vec{B} \cdot d\vec{s}$) through any surface S bounded by an arbitrary closed loop Γ, induces an electromotive force given by the integral of the tangential component of the induced electric field around Γ. The line integration over the contour Γ must be consistent with the direction of the surface vector $d\vec{s}$ according to the right-hand rule. The minus sign in the equations of the law represents the feature by which the induced electric field, when it acts on charges, would produce an induced current that opposes the change in the magnetic flux (Lenz's law).

Ampere's generalized law, constitutes another connection, different from Faraday's law, between $\vec{E}$ and $\vec{B}$. It states that the vector sources of the magnetic field may be free currents, $\vec{J}$, and/or displacement currents, $\frac{\partial \vec{D}}{\partial t}$. Thus, the displacement current performs, as a vector source of $\vec{H}$, a similar role to that played by $\frac{\partial \vec{B}}{\partial t}$ as a source of $\vec{E}$. In its integral form, the left-hand side of the generalized Ampere's law equation represents the integral of the magnetic field tangential components along an arbitrary closed loop Γ and the right-hand side is the sum of the flux, through any surface S bounded by a closed loop Γ, of both currents: the free current $\vec{J}$ and the displacement current $\frac{\partial \vec{D}}{\partial t}$.

2.2 Constitutive equations

In the vacuum, Maxwell's equations can be written without using the artificial fields $\vec{D}$ and $\vec{H}$, as

$$\nabla \cdot \vec{E}(\vec{r},t) = \frac{\rho_{all}}{\varepsilon_0}(\vec{r},t) \tag{2.11}$$

$$\nabla \cdot \vec{B}(\vec{r},t) = 0 \tag{2.12}$$

$$\nabla \times \vec{E}(\vec{r},t) = -\frac{\partial \vec{B}(\vec{r},t)}{\partial t} \tag{2.13}$$

$$\nabla \times \vec{B}(\vec{r},t) = \mu_0 \vec{J}_{all}(\vec{r},t) + \mu_0 \varepsilon_0 \frac{\partial \vec{E}(\vec{r},t)}{\partial t} \tag{2.14}$$

where $\varepsilon_0 = \frac{10^{-9}}{36\pi}$ (farad/meter, $F \cdot m^{-1}$) and $\mu_0 = 4\pi 10^{-7}$ (henry/meter, $H \cdot m^{-1}$) are two constants called electric permittivity and magnetic permeability of free space, respectively. The subscript *all* indicates that all kinds of charges (free and bound) must be individually included in ρ and $\vec{J}$. These equations are, within the limits of classical electromagnetic theory, absolutely general. Nevertheless, in order to make it possible to study the interaction between an electromagnetic field and a medium and to take into account the discrete nature of matter, it is necessary to develop macroscopic models to obtain Maxwell's macroscopic equations, in which only macroscopic quantities are used and in which only the densities of free charges and currents explicitly appear as sources of the fields. To this end, the atomic and molecular physical properties, which fluctuate greatly over atomic distances, are averaged over microscopically large volume elements, Δv, so that these contain a large number of molecules but at the same time are macroscopically small enough to represent accurate spatial dependence at a macroscopic scale. As a result of this average, the properties of matter related to atomic and molecular charges and currents are described by the macroscopic parameters, electric permittivity ε, magnetic permeability μ, and electrical conductivity σ. These parameters, called constitutive parameters, are in general smoothed point functions. The derivation of the constitutive parameters of a medium from its microscopic properties is, in general, an involved process that may require complex models of molecules as well as quantum and statistical theory to describe their collective behavior. Fortunately, in most of the practical situations, it is possible to achieve good results using simplified microscopic models.

To define the electric permittivity and describe the behavior of the electric field in the presence of matter, we must introduce a new macroscopic field quantity, $\vec{P}$ $(C \cdot m^{-2})$, called electric polarization vector, such that

$$\vec{D} = \varepsilon_0 \vec{E} + \vec{P} \tag{2.15}$$

and defined as the average dipole moment per unit volume

$$\vec{P} = \lim_{\Delta v \to 0} \frac{\sum_{n=1}^{N\Delta v} \vec{p}_n}{\Delta v} \tag{2.16}$$

where N is the number of molecules per unit volume and the numerator is the vector sum of the individual dipolar moments, $\vec{p}_n$, if atoms and molecules contained in a macroscopically infinitesimal volume Δv. For many materials, called linear isotropic media, $\vec{P}$ can be considered collinear and proportional to the electric field applied. Thus we have

$$\vec{P} = \varepsilon_0 \chi_e \vec{E} \tag{2.17}$$

where the dimensionless parameter χ_e, called the electric susceptibility of the medium, describes the capability of a dielectric to be polarized. Equation (2.15) can be written in a more compact form as

$$\vec{D} = (1 + \chi_e)\varepsilon_0 \vec{E} \tag{2.18}$$

so that

$$\vec{D} = \varepsilon_0 \varepsilon_r \vec{E} = \varepsilon \vec{E} \tag{2.19}$$

where

$$\varepsilon_r = 1 + \chi_e \tag{2.20}$$

and

$$\varepsilon = \varepsilon_0 \varepsilon_r \tag{2.21}$$

are the relative permittivity and the permittivity of the medium, respectively.

To define the magnetic permeability and describe the behavior of the magnetic field in the presence of magnetic materials, we must introduce another macroscopic field

quantity, called magnetization vector $\vec{M}$ $(A \cdot m^{-1})$, such that

$$\vec{H} = \frac{\vec{B}}{\mu_0} - \vec{M} \tag{2.22}$$

where $\vec{M}$ is defined, in a similar way to that of the electric polarization vector, as the average magnetic dipole moment per unit volume

$$\vec{M} = \lim_{\Delta v \to 0} \frac{\sum_{n=1}^{N\Delta v} \vec{m}_n}{\Delta v} \tag{2.23}$$

where N is the number of atomic current elements per unit volume and the numerator is the vector sum of the individual magnetic moments, $\vec{m}_n$ contained in a macroscopically infinitesimal volume Δv.

In general, $\vec{M}$ is a function of the history of $\vec{B}$ or $\vec{H}$, which is expressed by the hysteresis curve. Nevertheless, many magnetic media can be considered isotropic and linear, such that

$$\vec{M} = \chi_m \vec{H} \tag{2.24}$$

where χ_m is the dimensionless magnetic susceptibility magnitude, being negative and small for diamagnetic, positive and small for paramagnet, and positive and large for ferromagnet. Thus

$$\vec{H} = \frac{1}{(1+\chi_m)\mu_0}\vec{B} = \frac{1}{\mu}\vec{B} \tag{2.25}$$

where

$$\mu_r = (1 + \chi_m) \tag{2.26}$$

and

$$\mu = \mu_r \mu_0 \tag{2.27}$$

are the relative magnetic permeability and the permeability of the medium, respectively.

In a vacuum, or free space, $\varepsilon_r = 1$, $\mu_r = 1$, and therefore the fields vectors $\vec{D}$ and $\vec{E}$, as well as $\vec{B}$ and $\vec{H}$, are related by

$$\vec{D} = \varepsilon_0 \vec{E} \tag{2.28}$$

$$\vec{B} = \mu_0 \vec{H} \tag{2.29}$$

Very often the relation between an electric field and the conduction current density $\vec{J}_c$

that it generates is given, at any point of the conducting material, by the phenomenological relation, called Ohm's law

$$\vec{J_c} = \sigma \vec{E} \tag{2.30}$$

so that $\vec{J}$ is linearly related to $\vec{E}$ through the proportionality factor σ called the conductivity of the medium. Conductivity is measured in Siemens per meter ($S \cdot m^{-1} = \Omega^{-1} \cdot m^{-1}$) or mhos per meter ($mho \cdot m^{-1}$). Media in which equation (2.30) is valid are called ohmic media. A typical example of ohmic media are metals where (2.30) holds in a wide range of circumstances. However, in other materials, such as semiconductors, the Ohm's law may not be applicable. For most metals σ is a scalar with a magnitude that depends on the temperature and that, at room temperature, has a very high value of the order of $10^7 mho \cdot m^{-1}$. Then very often metals are considered as perfect conductors with an infinite conductivity.

The relations between macroscopic quantities, (2.15), (2.25) and (2.30), are called constitutive relations. Depending on the characteristics of the constitutive macroscopic parameters ε, μ and σ, which are associated with the macroscopic response of atoms and molecules in medium, this medium can be classified as:

- Inhomogeneous or homogeneous: according to whether or not the constitutive parameter of interest is a function of the position, $\varepsilon = \varepsilon(\vec{r})$, $\mu = \mu(\vec{r})$, $\sigma = \sigma(\vec{r})$.
- Anisotropic or isotropic: according to whether or not the response of the medium depends on the orientation of the field. In isotropic media all magnitudes of interest are parallel, i.e., $\vec{E}$ and $\vec{D}$, $\vec{E}$ and $\vec{J_c}$, $\vec{B}$ and $\vec{H}$. In anisotropic materials the constitutive parameter of interest is a tensor.
- Non linear or linear: according to whether or not the constitutive parameters depend on the magnitude of the applied fields. For instance $\varepsilon(E)$, $\sigma(E)$ and $\mu(H)$. $\sigma \neq \sigma(t)$.
- Dispersive: according to whether or not, for time-harmonic fields, the constitutive parameters depend on the frequency, $FT(\varepsilon) = FT(\varepsilon)(\omega), FT(\mu) = FT(\mu)(\omega), FT(\sigma) = FT(\sigma)(\omega)$, here, $FT()$ represents the Fourier transformation. The materials in which these parameters are functions of the frequency are called dispersive.
- Magnetic medium: if $\mu \neq \mu_0$. Otherwise the medium is called nonmagnetic because its only significant reaction to the electromagnetic field is polarization.

Fortunately, in many cases the medium in which the electromagnetic field exists can be considered homogeneous, linear, isotropic, non dispersive and non magnetic. Indeed, this assumption is not very restrictive since many electromagnetic phenomena can be studied using this simplification.

Equations (2.15) and (2.22) are simplified. In practice, we must write [2]:

$$\vec{D} = \vec{D}(\vec{E}, \vec{B}) \tag{2.31}$$

$$\vec{H} = \vec{H}(\vec{E}, \vec{B}) \tag{2.32}$$

In this thesis, we only consider:

- Non magnetic medium for which,

$$\vec{H}(\vec{r}, t) = \frac{1}{\mu_0}\vec{B}(\vec{r}, t) \tag{2.33}$$

- Linear, isotropic, homogeneous and time invariant medium with respect to electrical properties,

$$\vec{D}(\vec{r}, t) = \varepsilon_0 \varepsilon_r(t) * \vec{E}(\vec{r}, t) = \varepsilon_0 \int_{t'<t} \varepsilon_r(t - t')\vec{E}(\vec{r}, t')dt' \tag{2.34}$$

  where $\varepsilon(t)$ is the impulse dielectric permittivity.

2.3 Boundary conditions

As is evident from the Maxwell's equation, in general the fields $\vec{E}, \vec{B}, \vec{D}$ and $\vec{H}$ are discontinuous at points where ε, μ and σ also are. Hence the field vectors will be discontinuous at a boundary between two media with different constitutive parameters.

The integral form of Maxwell's equations can be used to determine the relations, called boundary conditions, of the normal and tangential components of the fields at the interface between two regions with different constitutive parameters ε, μ and σ where surface density of sources may exist along the boundary.

The boundary condition for $\vec{D}$ can be calculated using a very thin, small pill-box that crosses the interface of the two media. Applying the divergence theorem [1] to (2.1) and

we have:

$$\oint \vec{D} \cdot d\vec{s} = \int_{Base1} \vec{D}_1 \cdot d\vec{s} + \int_{Curved\ surface} \vec{D} \cdot d\vec{s} + \int_{Base2} \vec{D}_2 \cdot d\vec{s} = \int \rho dv \qquad (2.35)$$

where $\vec{D}_1$ denotes the value of $\vec{D}$ in medium 1, and $\vec{D}_2$ the value in medium 2. Since both bases of the pillbox can be made as small as we like, the total outward flux of $\vec{D}$ over them is $(D_{n1} - D_{n2})ds = (\vec{D}_1 - \vec{D}_2) \cdot \hat{n}ds$, where these D_n are the normal drawn from medium 2 to medium 1 and $\hat{n}$ is the unit normal vector. At the limit, by taking a shallow enough pillbox, we can disregard the flux over the curved surface, whereupon the sources of $\vec{D}$ reduce to the density of surface free charge ρ_s on the interface,

$$\hat{n} \cdot (\vec{D}_1 - \vec{D}_2) = \rho_s \qquad (2.36)$$

Hence the normal component of $\vec{D}$ changes discontinuously across the interface by an amount equal to the free charge surface density ρ_s on surface boundary.

Similarly the boundary condition for $\vec{B}$ can be established using the Gauss' law for magnetic fields. Since the magnetic field is solenoidal, it follows that the normal components of $\vec{B}$ are continuous across the interface between two media,

$$\hat{n} \cdot (\vec{B}_1 - \vec{B}_2) = 0 \qquad (2.37)$$

The behavior of the tangential components of $\vec{E}$ can be determined using an infinitesimal rectangular loop at the interface which has sides of length dh, normal to the interface, and sides of length dl parallel to it. From the integral form of the Faraday's law and defining $\hat{t}$ as the unit tangent vector parallel to the direction of integration on the upper side of the loop, we have:

$$(\vec{E}_1 \cdot \hat{t} - \vec{E}_2 \cdot \hat{t})dl + contributions\ of\ sides\ dh = -\frac{\partial \vec{B}}{\partial t} \cdot d\vec{s} \qquad (2.38)$$

In the limit, as $dh \to 0$, the area $ds = dhdl$ bounded by the loop approaches zero and, since $\vec{B}$ is finite, the flux of $\vec{B}$ vanishes. Hence $(\vec{E}_1 - \vec{E}_2) \cdot \hat{t} = 0$ and we conclude that the tangential components of $\vec{E}$ are continuous across the interface between two media.

In term of the normal $\hat{n}$ to the boundary, this can be written as:

$$\hat{n} \times (\vec{E}_1 - \vec{E}_2) = \vec{0} \tag{2.39}$$

where the symbol $\times$ designate the cross product.

Analogously, using the same infinitesimal rectangular loop, it can be deduced from the generalized Ampere's law that

$$(\vec{H}_1 \cdot \hat{t} - \vec{H}_2 \cdot \hat{t})dl + contributions\ of\ sides\ dh = -(\frac{\partial \vec{D}}{\partial t} + \vec{J}) \cdot d\vec{s} \tag{2.40}$$

where, since $\vec{D}$ is finite, its flux vanishes. Nevertheless, the flux of the surface current can have a non-zero value when the integration loop is reduced to zero, if the conductivity σ of the medium 2, and consequently $\vec{J}_s$, is finite. This requires the surface to be a perfect conductor. Thus,

$$\hat{n} \times (\vec{H}_1 - \vec{H}_2) = \vec{J}_s \tag{2.41}$$

the tangential component of $\vec{H}$ is discontinuous by the amount of surface current density $\vec{J}_s$. For finite conductivity, the tangential magnetial field is continuous across the boundary.

A summary of the boundary conditions are given for the general case and for the case when the medium 2 is a perfect conductor:

General boundary conditions

$$\hat{n} \times (\vec{E}_1 - \vec{E}_2) = \vec{0} \tag{2.42}$$

$$\hat{n} \times (\vec{H}_1 - \vec{H}_2) = \vec{J}_s \tag{2.43}$$

$$\hat{n} \cdot (\vec{D}_1 - \vec{D}_2) = \rho_s \tag{2.44}$$

$$\hat{n} \cdot (\vec{B}_1 - \vec{B}_2) = 0 \tag{2.45}$$

where $\vec{J}_s$ and ρ_s are potential surface density of charge or current.

Boundary conditions when the medium 2 is a perfect conductor ($\sigma_2 \to \infty$)

$$\hat{n} \times \vec{E}_1 = \vec{0} \tag{2.46}$$

$$\hat{n} \times \vec{H}_1 = \vec{J}_s \tag{2.47}$$

$$\hat{n} \cdot \vec{D}_1 = \rho_s \tag{2.48}$$

$$\hat{n} \cdot \vec{B}_1 = 0 \tag{2.49}$$

2.4 The conservation of energy

Poynting's theorem represents the electromagnetic energy-conservation law. To derive the theorem, let us calculate the divergence of the vector field $\vec{E} \times \vec{H}$ in a homogeneous, linear and isotropic finite region V bounded by a closed surface S. If we assume that V contains power sources generating currents $\vec{J}$, then, from Maxwell's equations, we get:

$$\nabla \cdot (\vec{E} \times \vec{H}) = \vec{H} \cdot \nabla \times \vec{E} - \vec{E} \cdot \nabla \times \vec{H} = -\vec{H} \cdot \frac{\partial \vec{E}}{\partial t} - \vec{E} \cdot \frac{\partial \vec{D}}{\partial t} - \vec{E} \cdot (\sigma \vec{E} + \vec{J}) \tag{2.50}$$

where $\vec{J}$ represents the source current density distribution which is the primary origin of the electromagenetic fields, while the induced conduction current density is written as $\vec{J}_c = \sigma \vec{E}$.

As the medium is assumed to be linear and no dispersive, the derivatives with respect to time can be written as

$$\vec{E} \cdot \frac{\partial \vec{D}}{\partial t} = \varepsilon \vec{E} \cdot \frac{\partial \vec{E}}{\partial t} = \frac{\partial}{\partial t}(\frac{\varepsilon E^2}{2}) = \frac{\partial}{\partial t}(\frac{\vec{E} \cdot \vec{D}}{2}) \tag{2.51}$$

$$\vec{H} \cdot \frac{\partial \vec{B}}{\partial t} = \mu \vec{H} \cdot \frac{\partial \vec{H}}{\partial t} = \frac{\partial}{\partial t}(\frac{\mu H^2}{2}) = \frac{\partial}{\partial t}(\frac{\vec{B} \cdot \vec{H}}{2}) \tag{2.52}$$

By introducing the equations (2.51) and (2.52) into (2.50), integrating over the volume V, applying the divergence theorem, and then rearranging terms, we have

$$\int_V \vec{J} \cdot \vec{E} dv = -\frac{\partial}{\partial t} \int_V \frac{1}{2}(\vec{E} \cdot \vec{D} + \vec{B} \cdot \vec{H}) dv - \int_V \sigma E^2 dv - \oint_S (\vec{E} \times \vec{H}) \cdot d\vec{s} \tag{2.53}$$

To interpret this result we accept that

$$U_{ev} = \frac{\vec{E} \cdot \vec{D}}{2} \tag{2.54}$$

and

$$U_{mv} = \frac{\vec{B} \cdot \vec{H}}{2} \tag{2.55}$$

represent, as a generalization of their expression for static fields, the instantaneous electric energy density, U_{ev}, and magnetic energy density, U_{mv}, stored in the respective fields. Let us recall the empirical Lorenz force equation, which gives the electromagnetic force density, $\vec{f}$ (in $N \cdot m^{-3}$), acting on a volume charge density ρ moving at a velocity u (in $m \cdot s^{-1}$) in a region where an electromagnetic filed exists,

$$\vec{f} = \rho(\vec{E} + \vec{u} \times \vec{B}) = \rho\vec{E} + \vec{J} \times \vec{B} \tag{2.56}$$

where $\vec{J} = \rho\vec{u}$ is the current density in terms of the mean drift velocity of the particles, which is independent of any random velocity due to collisions. The total force $\vec{F}$ exerted on a volume of charge is calculated by integrating $\vec{f}$ in this volume. For a single particle with charge q the Lorentz force is:

$$\vec{F} = q(\vec{E} + \vec{u} \times \vec{B}) \tag{2.57}$$

The work done by the electromagnetic field that acting on a volume density ρ inside a volume dv during a time interval dt is

$$dW = \vec{f} \cdot \vec{u}dtdv = \rho(\vec{E} + \vec{u} \times \vec{B}) \cdot \vec{u}dtdv = \rho\vec{E} \cdot \vec{u}dtdv = \vec{E} \cdot \vec{J}dtdv \tag{2.58}$$

This work is transformed into heat. The corresponding power density P_v (in $W \cdot m^{-3}$) that the electromagnetic field supplies to the charge distribution is:

$$P_v = \frac{dP}{dv} = \frac{dW}{dtdv} = \vec{E} \times \vec{J} \tag{2.59}$$

This equation is known as the point form of Joule's law. So the left side of equation (2.53) represents the total electromagnetic power supplied by all the sources within the volume V. Regarding the right side of equation (2.53), the first term represents the change rate of the stored electromagnetic energy within the volume, the second term

represents the dissipation rate of the electromagnetic energy within the volume, and the third term represents the flow of electromagnetic energy per second (power) through the surface S that bounds volume V. Defining Poynting's vector $\vec{\mathcal{P}}$ as

$$\vec{\mathcal{P}} = \vec{E} \times \vec{H} \ (W \cdot m^{-2}) \tag{2.60}$$

we can write

$$\oint_S (\vec{E} \times \vec{H}) \cdot d\vec{s} = \oint_S \vec{\mathcal{P}} \cdot d\vec{s} \tag{2.61}$$

This equation represents the total flow of power passing through the closed surface S and, consequently, we conclude that $\vec{\mathcal{P}} = \vec{E} \times \vec{H}$ represents the power passing through a unit area perpendicular to the direction of $\vec{\mathcal{P}}$.

Note that equation (2.53) was deduced by assuming a linear medium and that the loss occurs only through conduction currents. Otherwise the equation should be modified to include other kinds of losses such as those due to hysteresis or possible transformations of the electromagnetic energy into mechanical energy, etc. When there are no sources within V, equation (2.53) represents an energy balance of that flowing through S versus that stored and dissipated in V.

2.5 Time-harmonic electromagnetic fields

A particular case of great interest is on in which the sources vary sinusoidally in time. In linear media, the time-harmonic dependence of the source gives rise to fields which, once having reached the steady state, also vary sinusoidally in time. However, time-harmonic analysis is important not only because many electromagnetic systems operate with signals that are practically harmonic, but also because arbitrary periodic time functions can be expanded into Fourier series of harmonic sinusoidal components while transient nonperiodic functions can be expressed as Fourier integrals. Thus, since the Maxwell's equations are linear differential equations, the total fields can be synthesized from its Fourier components.

Analytically, the time-harmonic variation is expressed using the complex exponential notation based on Euler's formula, where it is understood that the physical fields are obtained by taking the real part, whereas their imaginary part is discarded. For example,

an electric field with time-harmonic dependence given by $cos(\omega t + \varphi)$, where ω is the angular frequency, is expressed as

$$\vec{E} = Re(\vec{\mathbf{E}}e^{j\omega t}) = \frac{1}{2}(\vec{\mathbf{E}}e^{j\omega t} + (\vec{\mathbf{E}}e^{j\omega t})^*) = \vec{E}_0 cos(\omega t + \varphi) \tag{2.62}$$

where $\vec{\mathbf{E}}$ is the complex phasor,

$$\vec{\mathbf{E}} = \vec{E}_0 e^{j\varphi} \tag{2.63}$$

of amplitude E_0 and phase φ, which will in general be a function of angular frequency and coordinates. The asterisk $*$ indicates the complex conjugate, and $Re()$ represents the real part of what is in the brackets.

Assuming $e^{j\omega t}$ time dependence, we can get the phasor form or time-harmonic form of Maxwell's equations simply by changing the operator $\frac{\partial}{\partial t}$ to the factor $j\omega$ and eliminating the factor $e^{j\omega t}$. Maxwell's equations in differential and integral forms for time-harmonic fields are given below.

Differential form of Maxwell's equation for time-harmonic fields

$$\nabla \cdot \vec{\mathbf{D}} = \rho \quad (Gauss' \ law) \tag{2.64}$$

$$\nabla \cdot \vec{\mathbf{B}} = 0 \quad (Gauss' \ law \ for \ magnetic \ fields) \tag{2.65}$$

$$\nabla \times \vec{\mathbf{E}} = -j\omega\vec{\mathbf{B}} \quad (Faraday's \ law) \tag{2.66}$$

$$\nabla \times \vec{\mathbf{H}} = \vec{\mathbf{J}} + j\omega\vec{\mathbf{D}} \quad (Generalized \ Ampere's \ law) \tag{2.67}$$

Integral form of Maxwell's equation for time-harmonic fields

$$\oint_S \vec{\mathbf{D}} \cdot d\vec{s} = Q_T \quad (Gauss' \ law) \tag{2.68}$$

$$\oint_S \vec{\mathbf{B}} \cdot d\vec{s} = 0 \quad (Gauss' \ law \ for \ magnetic \ fields) \tag{2.69}$$

$$\oint_\Gamma \vec{\mathbf{E}} \cdot d\vec{l} = -j\omega \int_S \vec{\mathbf{B}} \cdot d\vec{s} \quad (Faraday's \ law) \tag{2.70}$$

$$\oint_\Gamma \vec{\mathbf{H}} \cdot d\vec{l} = \int_S (\vec{\mathbf{J}} + j\omega\vec{\mathbf{D}}) \cdot d\vec{s} \quad (Generalized \ Ampere's \ law) \tag{2.71}$$

For the linear, homogeneous and invariant electrical medium, the constitutive relations become:

$$\vec{H}(\vec{r}, f) = \frac{1}{\mu_0}\vec{B}(\vec{r}, f) \tag{2.72}$$

$$\vec{D}(\vec{r}, f) = \varepsilon_c \vec{E}(\vec{r}, f) \tag{2.73}$$

with

$$\varepsilon_c = \varepsilon_0 \hat{\varepsilon}_r(f), \quad \hat{\varepsilon}_r(f) = FT(\varepsilon_r) \tag{2.74}$$

where $\hat{\varepsilon}(f)$ depends on the frequency for a dispersive medium. We can write $\hat{\varepsilon}_r = \varepsilon_r' + j\varepsilon_r''$ where $\varepsilon_r'' < 0$ for $f > 0$. $\hat{\varepsilon}_r$ is the relative complex permittivity.

Similar process occurs in magnetic and conducting media, and, within a given frequency range, there may be a phase shift between $\vec{\mathbf{E}}$ and $\vec{\mathbf{J}}_c$ or between $\vec{\mathbf{B}}$ and $\vec{\mathbf{H}}$ which, at the macroscopic level, is reflected in the corresponding complex constitutive parameters $\sigma_c = \sigma' + j\sigma''$ and $\mu_c = \mu' + j\mu''$.

For a medium with complex permittivity, the complex phasor form of the displacement current is:

$$j\omega\vec{\mathbf{D}} = j\omega\varepsilon_c\vec{\mathbf{E}} = \omega\varepsilon''\vec{\mathbf{E}} + j\omega\varepsilon'\vec{\mathbf{E}} \tag{2.75}$$

with $\varepsilon' = \varepsilon_0\varepsilon_c'$ and $\varepsilon'' = \varepsilon_0\varepsilon_c''$. While the sum, of the displacement and conduction current, called total induced current, $\vec{\mathbf{J}}_i$, is

$$\vec{\mathbf{J}}_i = \sigma\vec{\mathbf{E}} + j\omega\varepsilon_c\vec{\mathbf{E}} = (\sigma + \omega\varepsilon'')\vec{\mathbf{E}} + j\omega\varepsilon'\vec{\mathbf{E}} = \vec{\mathbf{J}}_d + \vec{\mathbf{J}}_r \tag{2.76}$$

where $\vec{\mathbf{J}}_d$, called the dissipative current,

$$\vec{\mathbf{J}}_d = (\sigma + \omega\varepsilon'')\vec{\mathbf{E}} \tag{2.77}$$

in phase with the electric field, is the real part of the induced current $\vec{\mathbf{J}}_i$ while $\vec{\mathbf{J}}_r$, called the reactive current,

$$\vec{\mathbf{J}}_r = j\omega\varepsilon'\vec{\mathbf{E}} \tag{2.78}$$

is the imaginary part of the induced current which is in phase quadrature with the electric field. The dissipative current can be expressed in a more compact form as

$$\vec{\mathbf{J}}_d = \sigma_e\vec{\mathbf{E}} \tag{2.79}$$

where σ_e is the effective or equivalent conductivity

$$\sigma_e = \sigma + \omega\varepsilon'' \tag{2.80}$$

which includes the ohmic losses due to σ and the damping losses due to $\omega\varepsilon''$. Thus the induced current, can be written as

$$\vec{\mathbf{J}}_i = \sigma_e\vec{\mathbf{E}} + j\omega\varepsilon'\vec{\mathbf{E}} = \sigma_{ec}\vec{\mathbf{E}} \tag{2.81}$$

where σ_{ec} is the complex effective conductivity, defined as

$$\sigma_{ec} = \sigma_e + j\omega\varepsilon' \tag{2.82}$$

Thus a medium with conductivity σ_{ec} and null permittivity is formally equivalent to one with conductivity and permittivity, σ and ε_c, respectively.

For harmonic signals the boundary conditions of the normal and tangential components of the fields at the interface between two regions with different constitutive parameters ε, μ and σ, become:

General boundary conditions

$$\hat{n} \times (\vec{\mathbf{E}}_1 - \vec{\mathbf{E}}_2) = \vec{0} \tag{2.83}$$

$$\hat{n} \times (\vec{\mathbf{H}}_1 - \vec{\mathbf{H}}_2) = \vec{\mathbf{J}}_s \tag{2.84}$$

$$\hat{n} \cdot (\vec{\mathbf{D}}_1 - \vec{\mathbf{D}}_2) = \rho_{\mathbf{s}} \tag{2.85}$$

$$\hat{n} \cdot (\vec{\mathbf{B}}_1 - \vec{\mathbf{B}}_2) = 0 \tag{2.86}$$

Boundary conditions when the medium 2 is a perfect conductor ($\sigma_2 \to \infty$)

$$\hat{n} \times \vec{\mathbf{E}}_1 = \vec{0} \tag{2.87}$$

$$\hat{n} \times \vec{\mathbf{H}}_1 = \vec{\mathbf{J}}_s \tag{2.88}$$

$$\hat{n} \cdot \vec{\mathbf{D}}_1 = \rho_{\mathbf{s}} \tag{2.89}$$

$$\hat{n} \cdot \vec{\mathbf{B}}_1 = 0 \tag{2.90}$$

In formulating the conservation energy equation for time-harmonic fields, it is convenient

to find, first, the time-average Poynting vector over a period, i.e. the time-average power passing through a unit area perpendicular to the directon of $\vec{\mathcal{P}}$. We have:

$$\vec{E} = Re(\vec{\mathbf{E}}e^{j\omega t}) = \frac{1}{2}(\vec{\mathbf{E}}e^{j\omega t} + (\vec{\mathbf{E}}e^{j\omega t})^*) \tag{2.91}$$

$$\vec{H} = Re(\vec{\mathbf{H}}e^{j\omega t}) = \frac{1}{2}(\vec{\mathbf{H}}e^{j\omega t} + (\vec{\mathbf{H}}e^{j\omega t})^*) \tag{2.92}$$

Thus, the instantaneous Poynting vector can be written as:

$$\vec{\mathcal{P}} = \vec{E} \times \vec{H} = Re(\vec{\mathbf{E}}e^{j\omega t}) \times Re(\vec{\mathbf{H}}^{j\omega t}) = \frac{1}{2}Re(\vec{\mathbf{E}} \times \vec{\mathbf{H}}^* + \vec{\mathbf{E}} \times \vec{\mathbf{H}}e^{2j\omega t}) \tag{2.93}$$

The time-average value of the instantaneous Poynting vector can be calculated integrating the above equation over period, i.e.,

$$\vec{\mathcal{P}}_{av} = \frac{1}{T}\int_0^T \vec{\mathcal{P}}dt = \frac{1}{2T}\int_0^T Re(\vec{\mathbf{E}}\times\vec{\mathbf{H}}^* + \vec{\mathbf{E}}\times\vec{\mathbf{H}}e^{2j\omega t})dt = \frac{1}{2}Re(\vec{\mathbf{E}}\times\vec{\mathbf{H}}^*) = Re(\vec{\mathcal{P}}_c) \tag{2.94}$$

since the time average of $\vec{\mathbf{E}} \times \vec{\mathbf{H}}e^{2j\omega t}$ vanishes. The magnitude,

$$\vec{\mathcal{P}}_c = \frac{1}{2}\vec{\mathbf{E}} \times \vec{\mathbf{H}}^* \tag{2.95}$$

is termed as the complex Poynting vector. Thus the time-average of the Poynting vector is equal to half of the real part of the complex Poynting vector.

2.6 Plane wave and propagation equations

In this thesis, we work with the time-harmonic electromagnetic fields. The time dependence of the plane wave is $e^{j\omega t}$, it will be omitted in the calculus. To exhibit the propagation equations, we apply the following mathematical formula to Maxwell's equations [1]:

$$\nabla \times (\nabla \times \vec{V}) = \nabla(\nabla \cdot \vec{V}) - \nabla^2\vec{V} \tag{2.96}$$

Then the propagation equations of $\vec{E}$ and $\vec{H}$ can be written as:

$$\nabla^2\vec{E} + \varepsilon_c\mu_0\omega^2\vec{E} = \frac{1}{\varepsilon_c}\nabla\rho + j\omega\mu_0\vec{J} \tag{2.97}$$

$$\nabla^2\vec{H} + \varepsilon_c\mu_0\omega^2\vec{H} = -\nabla \times \vec{J} \tag{2.98}$$

If the medium contains neither free-charge nor free-current ($\rho = 0, \vec{J} = 0$), then the propagation equations become:

$$\nabla^2 \vec{E} + \varepsilon_c \mu_0 \omega^2 \vec{E} = \vec{0} \tag{2.99}$$

$$\nabla^2 \vec{H} + \varepsilon_c \mu_0 \omega^2 \vec{H} = \vec{0} \tag{2.100}$$

These are the Helmholtz equations[2–4]. They have a particular solution in the following form:

$$\vec{E}(\vec{r}, t) = \vec{E}_0 e^{-j\vec{k}\cdot\vec{r}} \tag{2.101}$$

$$\vec{H}(\vec{r}, t) = \vec{H}_0 e^{-j\vec{k}\cdot\vec{r}} \tag{2.102}$$

where $\vec{E}_0$ and $\vec{H}_0$ are independent of $\vec{r}$.

The vector $\vec{k}$ is the wave vector of propagation medium. We have the following equation:

$$\vec{k}^2 = \varepsilon_0 \hat{\varepsilon}_r \mu_0 \omega^2 \tag{2.103}$$

If the medium is transparent, then $\hat{\varepsilon_r}$ and $\vec{k}$ are real. We have that:

$$k = \frac{2\pi}{\lambda} \tag{2.104}$$

where λ is the wavelength. From the two equations above, we have:

$$\varepsilon_0 \hat{\varepsilon}_r \mu_0 v^2 = 1 \tag{2.105}$$

where v is a constant velocity characterizing the propagation medium. In particular, if the medium is the vacuum, then we have:

$$\varepsilon_0 \hat{\varepsilon}_r c^2 = 1 \tag{2.106}$$

where c is the velocity of light. For a lossless medium with optical index $\nu = \sqrt{\hat{\varepsilon}_r}$, the constant v is:

$$v = \frac{c}{\nu} \tag{2.107}$$

Equation (2.101) and (2.102) express a monochromatic plane wave with the propagation direction given by the wave vector $\vec{k}$. For the plane wave described by $\vec{k} \cdot \vec{r} = Constant$,

the amplitudes of the components of the wave are independent of the position and remain constant.

From Maxwell's equations, we know that:

$$\vec{E} = \frac{\vec{H} \times \vec{k}}{\varepsilon_0 \hat{\varepsilon}_r \omega} \tag{2.108}$$

With the definition of impedance of medium: $Z = \sqrt{\frac{\mu_0}{\varepsilon_0 \hat{\varepsilon}_r}}$, we have:

$$\vec{E} = Z\vec{H} \times \frac{\vec{k}}{k} \tag{2.109}$$

For the wave propagation, if we consider the Cartesian coordinate $(Oxyz)$, with orthogonal basis $(\vec{u}_x, \vec{u}_y, \vec{u}_z)$. An incident monochromatic plane wave propagates in the space constituted of two media that are separated by an interface. The incident wave vector $\vec{k}_0$ is located in the plane (xOz). The direction of propagation of the incident wave is represented by the angle θ_0 with respect to the Oz axis. The polarization of a plane wave is then determined based on the curve that is going to describe the electric field $\vec{E}$ in a wave plane. This polarization is in general elliptical and can be decomposed into a combination of two linear polarizations: horizontal and vertical.

Horizontal polarization corresponds to the case where the electric field $\vec{E}$ is perpendicular to the plane of incidence formed by the couple of vectors $(\vec{k}_0, \vec{u}_z)$ where $\vec{u}_z$ is the unit vector of Oz axis. It is also called the transverse electric polarization (denoted TE, h or S). We shall call polarization E parallel and will be denoted by $E_{//}$ because the $\vec{E}$ field is parallel to the plane (xOy). The situation is similar with vertical polarization where $\vec{H}$ replaces $\vec{E}$.

The following table gives convention of notation, the notations in the same column represent the same polarization.

Horizontal Polarization	Vertical Polarization
h	v
$E_{//}$	$H_{//}$
TE	TM
S	P

The wave vector $\vec{k}_0$ can be represented by the incidence zenith angle θ_0 and the azimuth angle φ_0:

$$\vec{k}_0 = \alpha_0 \vec{u}_x + \beta_0 \vec{u}_y - \gamma_0 \vec{u}_z \tag{2.110}$$

with

$$\begin{cases} \alpha_0 = k_0 \sin\theta_0 \cos\varphi_0 \\ \beta_0 = k_0 \cos\theta_0 \\ \gamma_0 = k_0 \sin\theta_0 \sin\varphi_0 \\ k_0 = \frac{2\pi}{\lambda} \end{cases} \tag{2.111}$$

In fact, all the solutions of the wave propagation problem can be expressed as a combination of elementary plane waves with different amplitudes and wave vectors. We will discuss this in details in the next chapter.

2.7 Conclusion

In this chapter, we present the electromagnetic field theory fundamentals. Especially, we present the Maxwell's equations and the interaction of an electromagnetic field with an object.

Despite their apparent simplicity, Maxwell's equations are in general not easy to solve. In fact, even in the most favorable situation of homogeneous, linear and isotropic media, there are not many problems of interest that can be analytically solved except for those presenting a high degree of geometrical symmetry. Moreover, the frequency range of scientific and technological interest can vary by many orders of magnitude, expanding from frequency value of zero (or very low) to roughly $10^{14} Hertz$. The behavior and values of the constitutive parameters can change very significantly in this frequency range. Conductivity, for example, can vary from 0 to $10^7 S \cdot m^{-1}$. It is even possible to build artificial materials, called metamaterials, which present electromagnetic properties that are not found in nature. Examples of such metamaterials are those characterized with both negative permittivity ($\varepsilon < 0$)) and negative permeability ($\mu < 0$). These media are called double-negative metamaterials and, owing to their unusual electromagnetic properties, they present many potential technological applications.

Another important factor to study the interaction of an electromagnetic field with an object is the electric size of the body, i.e., the relationship between the wavelength and the body size, which can also vary by several orders of magnitude. All these circumstances make it in general necessary to use analytical, semi-analytical or numerical methods appropriate to each situation. In particular, numerical methods are fundamental for simulating and solving complex problems that do not admit analytical solutions.

In the next chapter, we will present the scattering by gratings and by random rough surfaces.

Chapter 3

Scattering by gratings and by random rough surfaces

In this thesis, the main aim is to study the diffraction by gratings and the scattering by random rough surfaces illuminated by a electromagnetic plane wave. In this chapter, I will introduce the fundamental theory about these two aspects.

3.1 The theory of diffraction gratings

3.1.1 Introduction

Diffraction gratings are optical components used to separate light into its component wavelengths. Diffraction gratings are used in spectroscopy, or for integration into spectrophotometers or monochromators. Diffraction gratings consist of a series of closely packed grooves that have been engraved or etched into the grating surface. Diffraction gratings can be either transmissive or reflective. As light transmits through or reflects off a grating, the grooves cause the light to diffract, dispersing the light into its component wavelengths [5].

For practical applications, gratings generally have ridges or rulings on their surface rather than dark lines. Such gratings can be either transmissive or reflective. Gratings which modulate the phase rather than the amplitude of the incident light are also produced, frequently using holography [5].

3.1.2 The grating equation

When monochromatic light is incident on a grating surface (i.e. a periodic surface), it is diffracted into discrete directions. We can picture each grating groove as being a very small, slit-shaped source of diffracted light. The light diffracted by each groove combines to form set of diffracted wavefronts. The usefulness of grating depends on the fact that there exists a unique set of discrete angles along which, for a given spacing D between grooves, the diffracted light from each facet is in phase with the light diffracted from any other facet, leading to constructive interface [1].

Diffraction by a grating can be visualized from the geometry in the figure 3.1, which shows a light ray of wavelength λ incident at an angle θ_0 and diffracted by a grating (of groove spacing D, also called the pitch) along a set of angles θ_n. These angles are measured from the grating normal. The sign convention for these angles depends on whether the light is diffracted on the same side or the opposite side of the grating as the incident light.

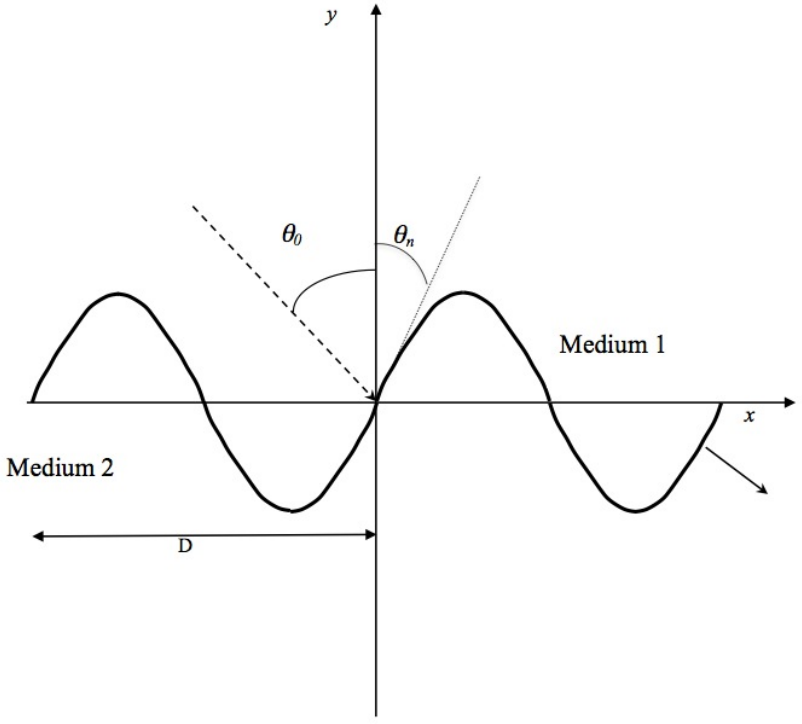

FIGURE 3.1: Diffraction by a grating where incident angle θ_0 and period D

The geometry path difference between light from adjacent grooves is seen to be $-D\sin\theta_0 + D\sin\theta_n$. The principle of constructive interface dictates that only when this difference equals the wavelength λ of the light, or some integral multiple thereof, will the light from adjacent grooves be in the phase (leading to constructive interface). All other angles the wavelets originating from the groove facets will interface destructively.

These relationships are expressed by the equation:

$$-1 < \sin\theta_n = \sin\theta_0 + n\frac{\lambda}{D} < 1 \tag{3.1}$$

which governs the angular locations of the principal intensity maxima when light of wavelength λ is diffracted from a gratings of groove spacing D. Here n is the diffraction order (or spectral order), which is an integer. For a wavelength λ, all values of n for which $|\sin\theta_0 + n\lambda/D| < 1$ correspond to propagating (rather than evanescent) diffraction orders. The special case $n = 0$ leads to the law of reflection $\theta_n = \theta_0$.

It is sometimes convenient to write the grating equation as

$$Gn\lambda = -\sin\theta_0 + \sin\theta_n \tag{3.2}$$

where $G = 1/D$ is the groove frequency or groove density, more commonly called "groove per millimeter".

Equation 3.1 and 3.2 are the common forms of the grating equation, but their validity is restricted to cases in which the incident and diffracted rays lie in a plane which is perpendicular to the grooves (at the center of grating). The majority of grating systems fall within this category, which is called classical diffraction. If the incident light beam is not perpendicular to the grooves, the grating equation must be modified:

$$Gn\lambda = \cos\varepsilon(-\sin\theta_0 + sin\theta_n) \tag{3.3}$$

Here, ε is the angle between the incident light path and the plane perpendicular to the groove at the grating center. In geometries, for which $\varepsilon \neq 0$, the diffracted spectra lie on a cone rather than in a plane, so such cases are termed conical diffraction.

For a grating of groove spacing D, there is a purely mathematical relationship between the wavelength and the angles of incidence and diffraction. In a given spectral order n,

the different wavelength of polychromatic wavefronts incident at angle θ_0 are separated in angle

$$\theta_n(\lambda) = \sin^{-1}(\frac{n\lambda}{D} + \sin\theta_0) \tag{3.4}$$

When $n = 0$, the grating acts as a mirror, and the wavelength are not separated ($\theta_n = \theta_0$ for all λ), this is called specular reflection or simply the zeroth order.

A special but common case is that in which the light is diffracted back towards the direction from which it came (i.e. $-\theta_0 = \theta_n$), this is called the Littrow configuration, for which the grating equation becomes:

$$n\lambda = -2d\sin\theta_0 \tag{3.5}$$

3.1.3 Diffraction orders

Generally several integers n will satisfy the grating equation. We call each of these values a diffraction order [5].

For a particular groove spacing D, wavelength λ and incidence angle θ_0, the grating equation 3.1 is generally satisfied by more than one diffraction angle θ_n. In fact, subject to restrictions discussed below, there will be several discrete angles at which the condition for constructive interference is satisfied. The physical significance of this is that the constructive reinforcement of wavelets diffracted by successive grooves merely requires that each ray be retarded (or advanced) in phase with every other, this phase difference must therefore correspond to a real distance (path difference) which equals an integral multiple of the wavelength. This happens, for example, when the path difference is one wavelength, in which case, we speak of the positive first diffraction order ($n = 1$) or the negative first diffraction order ($n = -1$), depending on whether the rays are advanced or retarded as we move from groove to groove.

The grating equation reveals that only those spectral orders for which $|\sin\theta_0 + n\lambda/D| < 1$ can exist. This restriction prevents light of wavelength from being diffracted in more than finite number of orders. Specular reflection ($n = 0$) is always possible. In most cases, the grating equation allows light of wavelength λ to be diffracted into both negative

and positive orders as well. Explicitly, spectra of all orders n exist for which,

$$(-1 - \sin\theta_0)D < n\lambda < (1 - \sin\theta_0)D, \quad n \text{ is an integer} \tag{3.6}$$

For $\lambda/D \ll 1$, a large number of diffracted orders will exist.

The most troublesome aspect of multiple order behavior is that successive spectral overlap. It is evident from the grating equation that light of wavelength λ diffracted by a grating along direction θ_n will be accompanied by integral fraction $\lambda/2$, $\lambda/3$, etc. That is for any grating instrument configuration, the light of wavelength λ diffracted in the $n = 1$ order will coincide with the light of wavelength $\lambda/2$ in the $n = 2$ order, etc. This superposition of wavelengths, which would lead to ambiguous spectroscopic data, is inherent in the grating equation itself and must be prevented by suitable filtering (called order sorting), since the detector cannot generally distinguish between light of different wavelengths incident on it.

3.1.4 Rayleigh expansion

Suppose the interface of the two media is described by the function $z = a(x, y)$, outside the deformation, the diffracted field $(\vec{E}, \vec{H})$ could be represented by the so-called Rayleigh expansion. For example, in the medium 1, when $z > \max a(x, y)$, the electric field $\vec{E}$ and the magnetic field $\vec{H}$ in $E_{//}$ could be represented with the help of the particular solutions as in equation (2.101) and (2.102). These are called the Rayleigh expansions [6]. In the discrete case, they can be represented as a linear combination of elementary plane waves. In the continuous case, they can be represented as a integral of the elementary plane waves. The Rayleigh expansion is only valid outside the modulated zone (i.e. $z > \max a(x, y)$ or $z < \min a(x, y)$) [7–12].

For example, if we consider only the one-dimensional interface for simplicity. In the Cartesian referential $Oxyz$, the grating is represented by a periodic cylindrical surface $y = a(x)$. This surface separates the air (medium 1) from the medium with a real or complex refractive index (medium 2). The grating of period D is illuminated by a monochromatic plane wave under the incidence θ_0. The incident wave vector lies in the xOy plane. The letter m denotes indifferently the upper medium ($m = 1$) or the lower medium ($m = 2$). Henceforth, $n^{(m)}, Z^{(m)}$ and $k^{(m)}$ indicate the optical index, the

impedence and the wave number of medium m. In this case, when $y > max(a(x))$ and $y < min(a(x))$, the diffracted field can be represented by a combination of elementary plane waves, the Rayleigh expansion:

$$\begin{cases} F_c^{(m)}(x,y) = \sum_n (c_n^{(m+)} exp(-j\alpha_n x) exp(-j\beta_n^{(m)} y) + c_n^{(m-)} exp(-j\alpha_n x) exp(j\beta_n^{(m)} y)) \\ G_c^{(m)}(x,y) = \sum_n \frac{\beta_n^{(m)}}{k^{(m)}} (c_n^{(m+)} exp(-j\alpha_n x) exp(-j\beta_n^{(m)} y) + c_n^{(m-)} exp(-j\alpha_n x) exp(j\beta_n^{(m)} y)) \end{cases} \tag{3.7}$$

The subscript (c) denotes the Cartesian components of electromagnetic field. In $E_{//}$ polarization, $F_c^{(m)}(x,y) = E_z^{(m)}(x,y)$, $G_c^{(m)} = Z^{(m)} H_z^{(m)}(x,y)$ and in $H_{//}$ polarization, $F_c^{(m)}(x,y) = Z^{(m)} H_z^{(m)}(x,y)$, $G_c^{(m)}(x,y) = -E_x^{(m)}(x,y)$. Superscripts $(+)$ and $(-)$ denotes a plane wave moving in direction along the y-axis and inverse the y-axis, respectively. The propagation coefficients of the n-th order diffraction are presented by α_n and $\beta_n^{(m)}$ with the relation

$$\alpha_n^2 + (\beta_n^{(m)})^2 = k^{(m)2} \tag{3.8}$$

where $Im(\beta_n^{(m)}) < 0$ and $\alpha_n = k^{(1)} sin\theta_0 + n\frac{2\pi}{D}$. $(\alpha_n, \beta_n^{(m)})$ are the propagation coefficients of the wave vector $\vec{k}_n^{(m)}$ of the elementary plane wave associated with the n^{th} diffraction order. $c_n^{(m\pm)}$ are the diffraction amplitudes of elementary plane waves. The propagation coefficient $\beta_n^{(m)}$ defines the nature of the plane wave: a propagation wave if $\beta_n^{(m)}$ is real, and an evanescent wave if $\beta_n^{(m)}$ is imaginary or complex.

3.1.5 The scattering matrix $\boldsymbol{S}$

With the discrete version of Rayleigh expansion, we can define the scattering matrix($\boldsymbol{S}$-matrix) to relate the amplitudes of outgoing plane waves to those of incoming waves. Take the one-dimensional case for simplicity, we have:

$$\begin{pmatrix} \boldsymbol{c}^{(1+)} \\ \boldsymbol{c}^{(2-)} \end{pmatrix} = \boldsymbol{S} \begin{pmatrix} \boldsymbol{c}^{(1-)} \\ \boldsymbol{c}^{(2+)} \end{pmatrix} \tag{3.9}$$

here we use $\boldsymbol{c}^{(m\pm)}$ to represent a vector containing the scattering amplitudes $c_n^{(m\pm)}$. For a perfectly conduction surface, the scattering matrix is given by:

$$\boldsymbol{c}^{(1+)} = \boldsymbol{S}\boldsymbol{c}^{(1-)} \tag{3.10}$$

Figure (3.2) illustrates the link between incoming and outcoming plane waves.

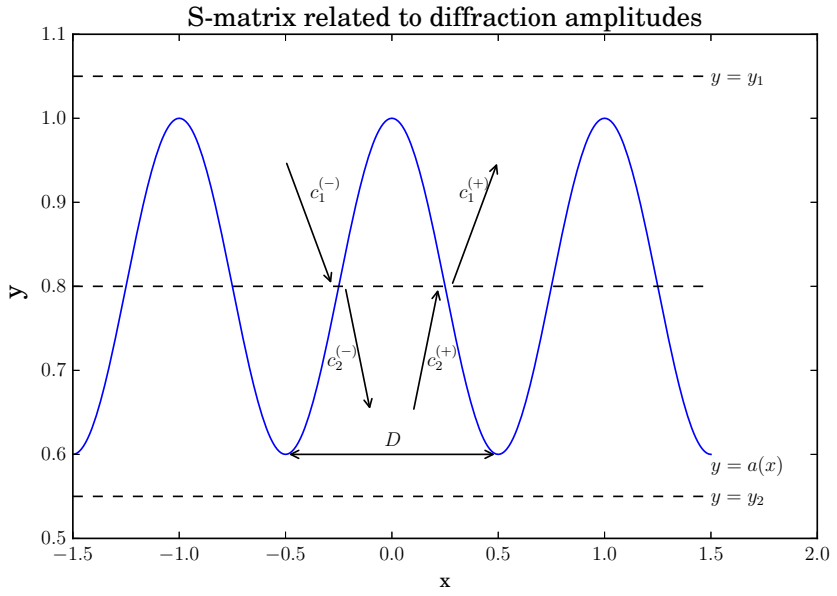

FIGURE 3.2: Scattering matrix S. Matrix associates the amplitudes of outgoing plane waves and those of incoming waves

3.1.6 The efficiency

For a lossless medium (m) with a real optical index $\nu^{(m)}$, the diffracted far field is expressed as a finite sum of plane waves propagating without attenuation, the directions of which form angles $\theta_n^{(m)}$ with the Oy axis. The aim is to determine the efficiencies that represent the incident power distribution in the different diffraction orders characterized by angles $\theta_n^{(m)}$ with $\cos\theta_n^{(m)} = \beta_n^{(m)}/k^{(m)}$. According to the definition of the complex Poynting vector (2.95), for an incidence angle θ_0, the efficiency $\epsilon_n^{(m)}$ is given by [6]:

$$\epsilon_n^{(m)} = \frac{\nu^{(m)}\cos\theta_n^{(m)}}{\cos\theta_0}|c_n^{(m)}|^2 \tag{3.11}$$

For a lossless grating, the sum of efficiencies is equal to 1 according with the conservation of power [6]. For the two media case, considering the propagation plane waves, we have:

$$\sum_n \epsilon_n^{(1)} + \sum_n \epsilon_n^{(2)} = 1 \tag{3.12}$$

3.2 Scattering from random rough surfaces

3.2.1 Introduction

The problem of electromagnetic scattering from random rough surfaces has aroused the interest of physicists and engineers for many years because of its wide range of applications in optics, material science, communications, oceanography and remote sensing. The three classical analytical methods commonly used in random rough surfaces scattering are the small-perturbation method [13], the Kirchhoff method [14–16] and the small slope approximation [17].

The electromagnetic analysis of rough surfaces with parameters close to the incident wavelength requires a rigorous formalism. Numerical methods based on Monte Carlo simulations are available for the study of electromagnetic wave scattering from one-dimensional and two-dimensional random rough surfaces. In the frequency domain, the boundary integral method can be used to analyze the scattering problem by rough surfaces. In this case, the electric or magnetic field integral equation is converted into matrix equations using the method of moments (MoM) [18, 19]. The average mesh length determines the accuracy of the MoM solution. The number of unknowns N is proportional to surface area in square wavelength. Several fast methods have been proposed to reduce the CPU time [20–25] and lead to a computational efficiency of $\mathcal{O}(N \log N)$. These methods are fully capable of describing the field scattered from surfaces of large size. Other numerical approaches are also suggested and for topical reviews, see [18] and [19]. Exact methods require solutions for many realizations of two-dimensional rough surfaces. The Monte-Carlo is used and the average scattered power is estimated over results of several surface realizations.

3.2.2 Random rough surface generation

In order to study the diffraction phenomenon of random rough surfaces, we use a numerical solution method based on exact electromagnetic model. This method requires the inputs describing surfaces to be analyzed. These inputs are numerical representation of the actual surfaces.

If we take the method of small perturbations (SPM) for example [14], which is an approximate analytical method, we only need the geometric parameters (statistics) which characterize the surface to be analyzed (standard deviation of heights, length correlation ...) to study its average response to electromagnetic excitation. In our case, these parameters will not intervene directly in the electromagnetic treatment. They will be used during a preliminary step in the numeric generation of surfaces that we want to analyze. These numeric profiles will be the inputs of the electromagnetic model that we have developed.

In a space based on the Cartesian orthonormal $(Oxyz)$, a surface whose generating line is based on a function $y = a(x)$ is called as cylindrical surface or one-dimensional surface. This surface is invariant along the z direction. For the two-dimensional surface, we use the equation $z = a(x, y)$ to represent the surface function.

In this thesis, we consider the one-dimensional or two-dimensional bounded supported random surfaces. They are expected from random process, verifying some assumptions that we explain later in this chapter. The randomness of these surfaces requires a statistical study to characterize it. To better understand the interaction between electromagnetic waves and the rough surface will require a good description of them.

In signal theory, a random process represents the evolution of a random variable with time or space. It is symbolized by a random function depending on time and/or on space and on a parameter $\mathcal{W}$ reflecting the randomness. A random process depending only on the space (time-independent) is noted as $\xi(r, \mathcal{W})$. For a given value $\mathcal{W}_0$ of $\mathcal{W}$, we obtain a realization of the random process $\xi(r, \mathcal{W}_0)$. This realisation is deterministic. If we vary the random parameter $\mathcal{W}$, we get a set of realisations from the same random process.

The statistical description of a random spatial process, whether in the one-dimensional or two-dimensional case, can be done by studying its spatial fluctuations depending on

the position. For a given position random functions $\xi(x_0, \mathcal{W})$ for the one-dimensional case, and $\xi(x_0, y_0, \mathcal{W})$ for the two-dimensional case, are random variables.

The distribution function of the random variable $\xi(\vec{r}_0, \mathcal{W})$ is defined as:

$$F_\xi(h) = Prob(\xi \leq h) \tag{3.13}$$

For a real random variable, the derivative of its distribution function (if it exists) gives the density of probability $P_\xi(h)$. The density of probability verifies the following properties:

$$Prob(\xi \in [a, b]) = \int_a^b P_\xi(h)dh \tag{3.14}$$

$$\forall h,\ P_\xi(h) > 0,\ \int_{-\infty}^{+\infty} P_\xi(h)dh = 1 \tag{3.15}$$

In practice, this function can be estimated by the normalized histogram of the values taken by the random variable $\xi(x, \mathcal{W})$ or $(x, y, \mathcal{W})$ for a given position. In the Gaussian case, the probability density is determined entirely by two parameters which are the statistical mean of the random variable which we denote as m_ξ and the standard deviation σ_ξ. σ_ξ measures the dispersion of the values of the random variable around the mean value m_ξ.

$$m_\xi = \mathbb{E}[\xi] = \int_{-\infty}^{+\infty} hP_\xi(h)dh \tag{3.16}$$

$$\sigma_\xi^2 = \mathbb{E}[\xi^2] - \mathbb{E}[\xi]^2 \tag{3.17}$$

$$\mathbb{E}[\xi^2] = \int_{-\infty}^{+\infty} h^2 P_\xi(h)dh \tag{3.18}$$

where $\mathbb{E}$ denotes the expectation which provides the statistical mean of the considered random variable. The analytical expression of a Gaussian probability density is given by:

$$P_\xi(h) = \frac{1}{\sigma_\xi\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{h-m_\xi}{\sigma_\xi})^2} \tag{3.19}$$

Now we consider a two-dimensional Gaussian random variable (ξ_1, ξ_2), with statistical average (m_{ξ_1}, m_{ξ_2}) and standard deviation $(\sigma_{\xi_1}, \sigma_{\xi_2})$. The associated joint probability

density is:

$$P_{\xi_1,\xi_2}(h_1,h_2) = \frac{1}{2\pi\sigma_{\xi_1}\sigma_{\xi_2}\sqrt{1-\rho_c^2}} exp(-\frac{\frac{(h_1-m_{\xi_1})^2}{\sigma_{\xi_1}^2} - \frac{2\rho_c(h_1-m_{\xi_1})(h_2-m_{\xi_2})}{\sigma_{\xi_1}\sigma_{\xi_2}} + \frac{(h_2-m_{\xi_2})^2}{\sigma_{\xi_2}^2}}{2(1-\rho_c^2)}) \tag{3.20}$$

with ρ_c the correlation coefficient:

$$\rho_c = \frac{\mathbb{E}[\xi_1\xi_2] - m_{\xi_1}m_{\xi_2}}{\sigma_{\xi_1}\sigma_{\xi_2}} \tag{3.21}$$

To analyze a random process $\xi(\vec{r},\mathcal{W})$, we can consider a study of a single realisation ($\mathcal{W}$ fixed). This allows us to know its spatial moments. In the general case, these moments may depend on the realisation, that is to say the random $\mathcal{W}$. The other way to do is to look at the values of the random variable associated with the process for a given position $\vec{r'}$ and several realizations of the process $\xi(\vec{r},\mathcal{W})$ (family of realisations). In this case, the process is described by using these statistical moments (statistical average and higher order moments).

Knowing the probability density of the random variable $\xi(\vec{r},\mathcal{W})$ for a given $\vec{r'}$, the statistical moment of order n, associated to this random variable is the expectation of order n. It is defined for a continuous random variable by:

$$m_{\xi^n} = \mathbb{E}[\xi^n(\vec{r'},\mathcal{W})] = \int_{-\infty}^{+\infty} h^n P_\xi(h)dh \tag{3.22}$$

To evaluate the correlation that may exist between two values taken by the random variable $\xi(\vec{r},W)$, at two different points $\vec{r'}$ and $\vec{r'}+\vec{r}$, we compute its statistical auto-correlation function defined by:

$$R(\vec{r'},\vec{r'}+\vec{r}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} hh' P_{(\xi_{\vec{r'}},\xi_{\vec{r'}+\vec{r}})}(h,h')dhdh' \tag{3.23}$$

$P_{\xi_{\vec{r'}},\xi_{\vec{r'}+\vec{r}}}(h,h')$ is the joint probability. This function depends on the position of two points, that is to say, $(\vec{r'},\vec{r}+\vec{r'})$.

If we work now on a single realisation, we have access to spatial moments. If we denote Δ the extent of $\xi(\vec{r}, \mathcal{W}_0)$, then the spatial average of $\xi(\vec{r}, \mathcal{W}_0)$ can be estimated by:

$$\overline{\xi(\vec{r}, \mathcal{W})} = \lim_{\Delta \to +\infty} \frac{1}{\Delta} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \xi(\vec{r}, \mathcal{W}_0) d\vec{r} \tag{3.24}$$

The spatial autocorrelation is defined as:

$$C_{\xi,\xi}(\vec{r}, \mathcal{W}_0) = \overline{\xi(\vec{r'}, \mathcal{W}_0)\xi(\vec{r'} + \vec{r}, \mathcal{W}_0)} = \lim_{\Delta \to +\infty} \frac{1}{\Delta} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \xi(\vec{r'}, \mathcal{W}_0)\xi(\vec{r'} + \vec{r}, \mathcal{W}_0) d\vec{r'} \tag{3.25}$$

Without prior assumptions about the random processes, the spatial moments depend on the realization. We will see what are the assumptions that allow the equality between spatial moments and statistical moments up to a certain order n. In general , the mean and spatial autocorrelation depend on the realisation. This means that these are random variables. If the spatial moments are independent of the random hazard until the order 2, then the spatial process is ergodic to order 2. Concerning the statistical moments, the calculated statistical average depends, in general, on the position $\vec{r}$ and the autocorrelation function of the two positions $\vec{r'}$ and $\vec{r}+\vec{r'}$. If now the autocorrelation function depends only on the distance between the two points and that the statistical average is constant when we change position, then the process is said to be stationary to order 2. If these last two properties (ergodicity and stationarity to order 2) are made simultaneously satisfied, then Birkoff's theorem allows to say that the statistical moments are equal to the spatial moments up to order 2. We can write:

$$m = \mathbb{E}[\xi(\vec{r_0}, \mathcal{W})] = \overline{\xi(\vec{r}, \mathcal{W}_0)} = Constant \tag{3.26}$$

$$R_{\xi\xi}(\vec{r}) = C_{\xi\xi}(\vec{r}) \tag{3.27}$$

For the surface, the random variable $a(\vec{r}, \mathcal{W})$ defines the height of the surface at all the points $\vec{r}$. The simulated surfaces are assumed to satisfy the two properties above. Moreover, we assume that the surface satisfies the following conditions:

- The probability density of the surface height $P(h)$ is Gaussian with mean $\mathbb{E}[a(\vec{r},\mathcal{W})] = 0$ and standard deviation σ_a. The density is given by:

$$P(h) = \frac{1}{\sigma_a\sqrt{2\pi}} exp(-\frac{h^2}{2\sigma^2}) \tag{3.28}$$

- The autocorrelation function is given by:

$$R(x,y) = \begin{cases} \sigma_a^2 exp(-\frac{x^2}{l_x^2}),\ one-dimensional\ case \\ \sigma_a^2 exp(-\frac{x^2}{l_x^2} - \frac{y^2}{l_y^2}),\ two-dimensional\ case \end{cases} \tag{3.29}$$

If the surface is isotropic, we have $l_x = l_y = l$.

We generate the random surface based on the principle of linear filtration of Gaussian white noise. In fact, the formula is as follows:

$$a(\vec{r}) = (h_f * B)(\vec{r}) = \int_{-\infty}^{+\infty} h_f(\vec{r} - \vec{r'})B(\vec{r'})d\vec{r'} \tag{3.30}$$

where B is the entry signal of the filter, it is related to a Gaussian white noise characterized by the Gaussian probability density and the autocorrelation function:

$$R_{BB}(\vec{r}) = \sigma_a^2\delta(\vec{r}) \tag{3.31}$$

with $\delta(\vec{r})$ the Dirac distribution. From the equation (3.30), we have the formula for $R_{aa}(\vec{r})$:

$$R_{aa}(\vec{r}) = (C_{h_f} * R_{BB})(\vec{r}) \tag{3.32}$$

where $C_{h_f}(\vec{r})$ is the spatial autocorrelation of the impulse response of filter:

$$C_{h_f}(\vec{r}) = \int_{-\infty}^{+\infty} h_f(\vec{r})h_f(\vec{r} - \vec{r'})d\vec{r'} \tag{3.33}$$

So we obtain that

$$R_{aa}(\vec{r}) = \sigma_a^2 C_{h_f}(\vec{r}) \tag{3.34}$$

Now suppose that $\hat{H}_f(\alpha,\beta)$ is the Fourier transformation (FT) of the impulse response:

$$\hat{H}_f(\alpha,\beta) = FT[h_f(x,y)] \tag{3.35}$$

Then we have

$$|\hat{H}_f(\alpha,\beta)|^2 = FT[C_{h_f}(x,y)] \tag{3.36}$$

The formula of filtration gives:

$$FT[R_{aa}(\vec{r})] = \sigma_a^2|\hat{H}_f(\alpha,\beta)|^2 \tag{3.37}$$

Given the function $R_{aa}(\vec{r})$ and suppose that $\hat{H}_f(\alpha,\beta) = |\hat{H}_f(\alpha,\beta)|$, we can calculate the impulse response:

$$h_f(x,y) = FT^{-1}[\hat{H}_f(\alpha,\beta)] = FT^{-1}[\sqrt{FT[R_{aa}(x,y)]}] \tag{3.38}$$

In particular, for the two-dimensional isotropic Gaussian surface, the impulse response is

$$h_f(x,y) = \frac{2}{l\sqrt{\pi}} exp(-2(\frac{x^2+y^2}{l})^2) \tag{3.39}$$

To implement this method, we need to discrete version of the formula. Suppose Δx and Δy are the length of the step along the direction $\vec{u}_x$ and $\vec{u}_y$. We have:

$$\Delta x = \frac{L_x}{N_x},\ \Delta y = \frac{L_y}{N_y} \tag{3.40}$$

and

$$a(x_i,y_j) = \Delta x \Delta y \sum_p \sum_q h_f(U_p,V_q)B(U_p - x_i, V_q - y_j) \tag{3.41}$$

with

$$x_i = i\Delta x,\ y_j = j\Delta y,\ U_p = p\Delta x,\ V_q = q\Delta y \tag{3.42}$$

$A = L_xL_y$ is the area of the generated surface. Thereafter, $L = L_x = L_y$ and $N_x = N_y = N_e$ and N_e^2 is the number of samples.

3.2.3 Beam of elementary plane waves

For the two dimensional interface, suppose the interface separating the air from a dielectric medium and described by the function $z = a(x,y)$. For the $E_{//}$ polarization, the component E_z will be zero and for the $H_{//}$ polarization, the component H_z will be zero.

	Nonzero components of $\vec{E}$	Nonzero components of $\vec{H}$
$E_{//}$	E_x, E_y	H_x, H_y, H_z
$H_{//}$	E_x, E_y, E_z	H_x, H_y

Outside the deformation, the diffracted field $(\vec{E}, \vec{H})$ could be represented by the Rayleigh integral when $z > \max a(x, y)$ or $z < \min a(x, y)$. The electric field $\vec{E}$ and the magnetic field $\vec{H}$ in $E_{//}$ could be represented as follows:

$$\vec{E}^{(1)}(x, y, z) = \frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C^{(1,E_{//})}(\alpha, \beta) \vec{h}(\alpha, \beta) e^{-j\vec{k}^{(1)}(\alpha,\beta)\cdot\vec{r}} d\alpha d\beta \tag{3.43}$$

$$Z^{(1)}\vec{H}^{(1)}(x, y, z) = \frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C^{(1,E_{//})}(\alpha, \beta) (\frac{\vec{k}^{(1)}(\alpha, \beta)}{k^{(1)}} \times \vec{h}(\alpha, \beta)) e^{-j\vec{k}^{(1)}(\alpha,\beta)\cdot\vec{r}} d\alpha d\beta \tag{3.44}$$

where $\vec{h}$ is the unit polarization vector:

$$\vec{h}(\alpha, \beta) = -\frac{\beta}{\sqrt{\alpha^2 + \beta^2}} \vec{u}_x + \frac{\alpha}{\sqrt{\alpha^2 + \beta^2}} \vec{u}_y \tag{3.45}$$

and

$$\vec{k}^{(1)} = \alpha \vec{u}_x + \beta \vec{u}_y + \gamma^{(1)} \vec{u}_z \tag{3.46}$$

with $\alpha^2 + \beta^2 + (\gamma^{(1)})^2 = (k^{(1)})^2$, $Im(\gamma^{(1)}) \leq 0$, $Re(\gamma^{(1)}) \geq 0$.

If $\alpha^2 + \beta^2 > (k^{(1)})^2$, the constant of propagation γ is pure imaginary and corresponds to the evanescent wave. If $\alpha^2 + \beta^2 \leq (k^{(1)})^2$, $\gamma^{(1)}$ is real and corresponds to the propagation wave. $C^{(1,E_{//})}(\alpha, \beta)$ is the amplitude of the elementary wave $e^{-j\vec{k}^{(1)}(\alpha,\beta)\cdot\vec{r}}$.

For the $H_{//}$, we have

$$Z^{(1)}\vec{H}^{(1)}(x, y, z) = \frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C^{(1,H_{//})}(\alpha, \beta) \vec{h}(\alpha, \beta) e^{-j\vec{k}(\alpha,\beta)\cdot\vec{r}} d\alpha d\beta \tag{3.47}$$

$$\vec{E}^{(1)}(x, y, z) = -\frac{1}{4\pi^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C^{(1,H_{//})}(\alpha, \beta) (\frac{\vec{k}^{(1)}(\alpha, \beta)}{k^{(1)}} \times \vec{h}(\alpha, \beta)) e^{-j\vec{k}^{(1)}(\alpha,\beta)\cdot\vec{r}} d\alpha d\beta \tag{3.48}$$

Equations (3.47) and (3.48) can be obtained from (3.43) and (3.44) by replace $\vec{E}^{(1)}$ with $Z^{(1)}\vec{H}^{(1)}$ and $Z^{(1)}\vec{H}^{(1)}$ with $-\vec{E}^{(1)}$.

For the one-dimensional case, suppose the interface is described by $y = a(x)$ which is invariant in the direction Oz. If the surface is illuminated by a plane wave with wave

vector $\vec{k}_0$ which is in the plane Oxy. As in the two-dimensional case, we have the following table:

	Nonzero components of $\vec{E}$	Nonzero components of $\vec{H}$
$E_{//}$	E_z	H_x, H_y
$H_{//}$	E_x, E_y	H_z

For $E_{//}$, the electromagnetic fields are:

$$\vec{E}^{(1)}(x,y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} C^{(1,E_{//})}(\alpha)e^{-j\vec{k}^{(1)}(\alpha)\cdot\vec{r}}d\alpha\vec{u}_z \tag{3.49}$$

$$Z^{(1)}\vec{H}^{(1)}(x,y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} C^{(1,E_{//})}(\alpha)(\frac{\vec{k}^{(1)}(\alpha)}{k^{(1)}} \times \vec{u}_z)e^{-j\vec{k}^{(1)}(\alpha)\cdot\vec{r}}d\alpha \tag{3.50}$$

with

$$\vec{k}^{(1)}(\alpha) = \alpha\vec{u}_x + \beta^{(1)}\vec{u}_y \tag{3.51}$$

and

$$\alpha^2 + (\beta^{(1)})^2 = (k^{(1)})^2,\ Im(\beta^{(1)}) \leq 0,\ Re(\beta^{(1)}) \geq 0 \tag{3.52}$$

For the $H_{//}$, the fields are:

$$Z^{(1)}\vec{H}^{(1)}(x,y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} C^{(1,H_{//})}(\alpha)e^{-j\vec{k}^{(1)}(\alpha)\cdot\vec{r}}d\alpha\vec{u}_z \tag{3.53}$$

$$\vec{E}^{(1)}(x,y) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} C^{(1,H_{//})}(\alpha)(\frac{\vec{k}^{(1)}(\alpha)}{k^{(1)}} \times \vec{u}_z)e^{-j\vec{k}^{(1)}(\alpha)\cdot\vec{r}}d\alpha \tag{3.54}$$

3.2.4 Scattering patterns

In the far-field zone, the Rayleigh expansion (3.43) and (3.44) is reduced to the only contribution of the propagation waves. For $E_{//}$, the method of stationary phase [26] leads to the asymptotic fields at the point $M(r,\theta,\varphi)$:

$$\vec{E}^{(1)}_{far}(r,\theta,\varphi) = C^{(1,E_{//})}(k^{(1)}\sin\theta\cos\varphi, k^{(1)}\sin\theta\sin\varphi)\cos\theta\frac{e^{-jk^{(1)}r}}{\lambda r}e^{-j\frac{\pi}{2}}\vec{u}_\varphi \tag{3.55}$$

$$Z^{(1)}\vec{H}^{(1)}_{far}(r,\theta,\varphi) = C^{(1,E_{//})}(k^{(1)}\sin\theta\cos\varphi, k^{(1)}\sin\theta\sin\varphi)\cos\theta\frac{e^{-jk^{(1)}r}}{\lambda r}e^{-j\frac{\pi}{2}}\vec{u}_\theta \tag{3.56}$$

Substituting $\vec{E}^{(1)}$ by $Z^{(1)}\vec{H}^{(1)}$ and $Z^{(1)}\vec{H}^{(1)}$ by $-\vec{E}^{(1)}$ in equation (3.55) and (3.56), we obtain the $H_{//}$ polarization components of magnetic and electric field vectors. For an incident wave in (a) polarization and a scattered wave in (b) polarization ($(a,b) \in \{E_{//}, H_{//}\}$), the normalized bistatic scattering coefficient $\sigma^{(ba)}$ is defined as follows:

$$\sigma^{(ba)}(\theta,\varphi) = \frac{1}{P_0^{(a)}}\frac{dP^{(ba)}}{d\Omega} = \frac{|C^{(1,ba)}(k^{(1)}\sin\theta\cos\varphi, k^{(1)}\sin\theta\sin\varphi)\cos\theta|^2}{\lambda^2 L^2\cos\theta_0} \tag{3.57}$$

where $\frac{dP^{(ba)}}{d\Omega}$ is the power scattered per unit solid angle $d\Omega = \sin\theta d\theta d\varphi$ with

$$dP^{(ba)} = \frac{1}{2}Re(\vec{E}^{(ba)}_{far} \times \vec{H}^{(ba)*}_{far} dS\vec{u}_r) \tag{3.58}$$

The symbol $*$ designates the complex conjugate. dS is the element surface with $dS = r^2 d\Omega$. The unit vectors $\vec{u}_r, \vec{u}_\theta, \vec{u}_\varphi$ are drawn in the direction of increasing r, θ and φ such as to constitute a right-hand base system. $P_0^{(a)}$ is the flux of incident power through the modulated region with L the modulated length (see equations 2.61 and 2.94):

$$P_0^{(a)} = \frac{1}{2}\int_{-L/2}^{+L/2}\int_{-L/2}^{L/2} Re(\vec{E}_i^{(a)} \times \vec{H}_i^{(a)*} dxdy\vec{u}_z) \tag{3.59}$$

For random rough surface, the average bistatic scattering coefficient is defined:

$$\mathbb{E}[\sigma^{(ba)}(\theta,\varphi)] = \frac{1}{\lambda^2 L^2}\frac{\cos^2\theta}{\cos\theta_0}\mathbb{E}[|C^{(1,ba)}(k^{(1)}\sin\theta\cos\varphi, k^{(1)}\sin\theta\sin\varphi)\cos\theta|^2] \tag{3.60}$$

For infinite extension surfaces, the three classical analytical methods (First-order perturbation method, first-order small slope approximation and the Kirchhoff method) lead to closed-form formula for the average bistatic scattering coefficient [13, 14, 17]. Exact methods require solutions for many realizations of two-dimensional rough surfaces. The Monte Carlo technique is applied to estimate the average bi-static coefficient from results over N_R different realizations [18].

$$\mathbb{E}[\sigma^{(ba)}(\theta,\varphi)] = \frac{1}{N_R}\sum_{j=1}^{N_R}\sigma_j^{(ba)}(\theta,\varphi) \tag{3.61}$$

Some authors prefer to use the radar cross section which is $4\pi\cos\theta_0 L^2\mathbb{E}[\sigma^{(ba)}(\theta,\varphi)]$.

3.3 Conclusion

In this chapter, we present the theory of diffraction gratings and the theory of scattering from random rough surfaces. We recall that the Rayleigh expansion is only valid outside the modulated zone. This is why we can not simply use Rayleigh expansion in the modulated zone and the C-method is needed here. For analyzing gratings, we present the concept of matrix $\boldsymbol{S}$, see equation (3.9), this concept gives rise to chapter 8, the new version of C-method: C-method as an initial value problem. We also define the average bistatic coefficient that is a quantity measured in remote sensing and in optics. In chapter 6, we compare the average bistatic coefficient estimated with the C-method and experimental data.

In the next chapter, we present the C-method and we show how this method leads to eigenvalue problem.

Chapter 4

The curvilinear coordinate method

4.1 Introduction

The resolution of Maxwell's equations requires to consider the continuity of certain components of the field on the interface. The continuity relation is simplified if the interface is one coordinate surface. If the interface is geometrically simple (plane, cylinder, sphere), we can use the corresponding coordinates. If it is not, in order to see clearly the continuity relations, we need to use nonorthogonal coordinates.

Several authors have adapted the Maxwell-Minkowsky equations to the three-dimensional space. By combining the electromagnetic field tensor, Minkowsky has generalized the Maxwell's equations to the space-time. E.J.Post has written the equations in the rationalized MKS system [27]. J.Chandezon *et al* have proposed to adapt this formulation to the three-dimensional space and valid for all the curvilinear coordinates [28]. Under this form, the Maxwell's equations are not affected by the coordinate system which is different from classical ways.

In order to use this formalism, J.Chandezon *et al* introduced two systems of nonorthogonal coordinates, the translation system to study the diffraction of a plane wave by a grating [28] and the revolution system to study the propagation of wave in the periodical cylindrical guides [29]. The principle of the proposed method is called the C-method.

This formalism has been extensively used in the theory of grating: grating with finite conductivity in conical incidence, multilayer gratings with parallel or non parallel interface, bi-crossed gratings [30–39]. The Maxwell's equations in covariant form leads to new pertubation methods [40] and to models of two roughness levels [41]. For these works, the medium is linear, homogeneous and isotropic. E. Popov and M. Nevière have extended the C-method to the grating containing materials with nonzero susceptibility $\chi^{(3)}$ [42]. Harris *et al* [43, 44], Inchaussandague and Depine [45, 46] have generalized the principle of resolution with anisotropic materials. G. Granet *et al* investigated the diffraction gratings with inhomogeneous materials [47]. L. Li *et al* have proposed a new formalism of C-method to study the interface with edges [48]. This formulation is based on the factorization rules of Fourier series and has a faster numerical convergence.

In most studies, the grating surface is described as a function and the study is done in the translation coordinate system. Plumey *et al* [49] have extended the C-method to study gratings that are not described by functions. Granet *et al* studied gratings given by parametric equations [50]. This formalism has given rise to some works in waveguide for the study of waveguide bends and power divider [51–54]. The C-method is also an efficient theoretical tool for analyzing rough surfaces illuminated by a plane wave [55–61] or a electromagnetic beam [62, 63]. Recently, D. Prémel *et al* have implemented an original formulation based on the field-potential vectors and applied to the domain of low frequencies [65, 75].

4.2 The Maxwell's equations in covariant form and the translation system

We will derive the Maxwell's equation in covariant form in this subsection. For the one-dimensional case, we have the surface function $y = a(x)$. We consider the translation system:

$$\begin{cases} x' = x \\ y' = u = y - a(x) \\ z' = z \end{cases} \tag{4.1}$$

Then the transformation matrix is:

$$A_{i'}^{i} = \frac{\partial x^{i}}{\partial x'^{i}} = \begin{pmatrix} A_{x'}^{x} & A_{y'}^{x} & A_{z'}^{x} \\ A_{x'}^{y} & A_{y'}^{y} & A_{z'}^{y} \\ A_{x'}^{z} & A_{y'}^{z} & A_{z'}^{z} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{da}{dx} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.2}$$

$$A_{i}^{i'} = \frac{\partial x^{i'}}{\partial x^{i}} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{da}{dx} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.3}$$

The covariant basis vectors can be expressed from the basis vectors of the Cartesian coordinate system $(\vec{u}_x, \vec{u}_y, \vec{u}_z)$:

$$\begin{cases} \vec{u}_{x'} = \vec{u}_x + \frac{da}{dx}\vec{u}_y \\ \vec{u}_{y'} = \vec{u}_y \\ \vec{u}_{z'} = \vec{u}_z \end{cases} \tag{4.4}$$

and the contravariant basis vectors can be written as:

$$\begin{cases} \vec{u}^{x'} = \vec{u}^{x} \\ \vec{u}^{y'} = -\frac{da}{dx}\vec{u}^{x} + \vec{u}^{y} \\ \vec{u}^{z'} = \vec{u}^{z} \end{cases} \tag{4.5}$$

The covariant and contravariant metric tensors are [3]:

$$g_{i'j'} = \sum_{i,j} A_{i'}^{i} A_{j'}^{j} g_{ij} = \begin{pmatrix} 1 + (\frac{da}{dx})^2 & \frac{da}{dx} & 0 \\ \frac{da}{dx} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.6}$$

$$g^{i'j'} = \sum_{i,j} A_{i}^{i'} A_{j}^{j'} g^{ij} = \begin{pmatrix} 1 & -\frac{da}{dx} & 0 \\ -\frac{da}{dx} & 1 + (\frac{da}{dx})^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.7}$$

So in the new coordinate system, the covariant components of a vector $\vec{v}$ can be written as,

$$\begin{cases} v_{x'} = v_x + \frac{da}{dx} v_y \\ v_{y'} = v_y \\ v_{z'} = v_z \end{cases} \tag{4.8}$$

and the contravariant components as,

$$\begin{cases} v^{x'} = v^x \\ v^{y'} = -\frac{da}{dx} v^x + v^y \\ v^{z'} = v^z \end{cases} \tag{4.9}$$

The covariant components $v_{y'}$ and $v_{z'}$ become identified with Cartesian ones v_y and v_z. Moreover, the covariant component $v_{x'}$ and $v_{y'}$ are parallel to the interface given by $u = 0$ (i.e. $y = a(x)$).

By a similar procedure, for the two-dimensional surface $z = a(x, y)$, we have by using the translation system $(x', y', z') = (x, y, z - a(x, y))$:

$$A^i_{i'} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{\partial a}{\partial x} & \frac{\partial a}{\partial y} & 1 \end{pmatrix} \tag{4.10}$$

$$A^{i'}_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{\partial a}{\partial x} & -\frac{\partial a}{\partial y} & 1 \end{pmatrix} \tag{4.11}$$

$$g_{i'j'} = \begin{pmatrix} 1 + (\frac{\partial a}{\partial x})^2 & \frac{\partial a}{\partial x}\frac{\partial a}{\partial y} & \frac{\partial a}{\partial x} \\ \frac{\partial a}{\partial x}\frac{\partial a}{\partial y} & 1 + (\frac{\partial a}{\partial y})^2 & \frac{\partial a}{\partial y} \\ \frac{\partial a}{\partial x} & \frac{\partial a}{\partial y} & 1 \end{pmatrix} \tag{4.12}$$

$$g^{i'j'} = \begin{pmatrix} 1 & 0 & -\frac{\partial a}{\partial x} \\ 0 & 1 & -\frac{\partial a}{\partial y} \\ -\frac{\partial a}{\partial x} & -\frac{\partial a}{\partial y} & 1 + (\frac{\partial a}{\partial x})^2 + (\frac{\partial a}{\partial y})^2 \end{pmatrix} \tag{4.13}$$

So, in the translation coordinate system, the covariant components of a vector $\vec{v}$ can be written as,

$$\begin{cases} v_{x'} = v_x + \frac{\partial a}{\partial x} v_z \\ v_{y'} = v_y + \frac{\partial a}{\partial y} v_z \\ v_{z'} = v_z \end{cases} \tag{4.14}$$

and the contravariant component as,

$$\begin{cases} v^{x'} = v^x \\ v^{y'} = v^y \\ v^{z'} = v^z - \frac{\partial a}{\partial x} v^x - \frac{\partial a}{\partial y} v^y \end{cases} \tag{4.15}$$

The covariant component $v_{z'}$ is simply the vertical component v_z. Moreover, the covariant components $v_{x'}$ and $v_{y'}$ are parallel to the surface coordinate $z' = 0$ (i.e. $z = a(x, y)$).

From the Ostrogradsky theorem and the Stokes' theorem expressed in a non orthogonal coordinate system $(x^{i'}, x^{j'}, x^{k'})$, the Maxwell's equations for time-harmonic fields can be written as [3]:

$$\frac{1}{\sqrt{g'}} \sum_{i'} \frac{\partial}{\partial x^{i'}} (\sqrt{g'} B^{x^{i'}}) = 0 \tag{4.16}$$

$$\frac{1}{\sqrt{g'}} (\frac{\partial E_{x^{k'}}}{\partial x^{j'}} - \frac{\partial E_{x^{j'}}}{\partial x^{k'}}) = -j\omega B^{x^{i'}} \tag{4.17}$$

$$\frac{1}{\sqrt{g'}} \sum_{i'} \frac{\partial}{\partial x^{i'}} (\sqrt{g'} D^{x^{i'}}) = 0 \tag{4.18}$$

$$\frac{1}{\sqrt{g'}} (\frac{\partial H_{x^{k'}}}{\partial x^{j'}} - \frac{\partial H_{x^{j'}}}{\partial x^{k'}}) = j\omega D^{x^{i'}} \tag{4.19}$$

where $g' = det(g^{i'j'})$. Here, we assume that there is no current density and no charge density. For a linear, homogeneous, isotropic and non magnetic medium, the constitutive relations for time-harmonic fields can be written as:

$$D^{x^{i'}} = \varepsilon_c E^{x^{i'}} = \varepsilon_c \sum_{j'=1}^{3} g^{i'j'} E_{x^{j'}} \tag{4.20}$$

$$B^{x^{i'}} = \mu_0 H^{x^{i'}} = \mu_0 \sum_{j'=0}^{3} g^{i'j'} E_{x^{j'}} \tag{4.21}$$

So the equations (4.17) and (4.19) associated with the constitutive relations (4.20) and (4.21) can be written as:

$$\frac{\partial E_{x^{k'}}}{\partial x^{j'}} - \frac{\partial E_{x^{j'}}}{\partial x^{k'}} = -j\omega\mu_0 \sum_{j'} g^{i'j'} H_{x^{j'}} \tag{4.22}$$

and

$$\frac{\partial H_{x^{k'}}}{\partial x^{j'}} - \frac{\partial H_{x^{j'}}}{\partial x^{k'}} = j\omega\varepsilon_c \sum_{j'} g^{i'j'} E_{x^{j'}} \tag{4.23}$$

more specifically, equations (4.22) and (4.23) are equivalent to the following six equations [28]:

$$\frac{\partial E_{z'}}{\partial y'} - \frac{\partial E_{y'}}{\partial z'} = -j\omega\mu_0(g^{x'x'}H_{x'} + g^{x'y'}H_{y'} + g^{x'z'}H_{z'}) \tag{4.24}$$

$$\frac{\partial E_{z'}}{\partial x'} - \frac{\partial E_{x'}}{\partial z'} = -j\omega\mu_0(g^{y'x'}H_{x'} + g^{y'y'}H_{y'} + g^{y'z'}H_{z'}) \tag{4.25}$$

$$\frac{\partial E_{y'}}{\partial x'} - \frac{\partial E_{x'}}{\partial y'} = -j\omega\mu_0(g^{z'x'}H_{x'} + g^{z'y'}H_{y'} + g^{z'z'}H_{z'}) \tag{4.26}$$

$$\frac{\partial H_{z'}}{\partial y'} - \frac{\partial H_{y'}}{\partial z'} = j\omega\varepsilon_c(g^{x'x'}E_{x'} + g^{x'y'}E_{y'} + g^{x'z'}E_{z'}) \tag{4.27}$$

$$\frac{\partial H_{z'}}{\partial x'} - \frac{\partial H_{x'}}{\partial z'} = j\omega\varepsilon_c(g^{y'x'}E_{x'} + g^{y'y'}E_{y'} + g^{y'z'}E_{z'}) \tag{4.28}$$

$$\frac{\partial H_{y'}}{\partial x'} - \frac{\partial H_{x'}}{\partial y'} = j\omega\varepsilon_c(g^{z'x'}E_{x'} + g^{z'y'}E_{y'} + g^{z'z'}E_{z'}) \tag{4.29}$$

4.3 Formulation for one-dimensional case

We first consider the one-dimensional case. As we have described before, a surface by equation $y = a(x)$ separates two different media. It is illuminated by a monochromatic plane wave with wavelength λ under incident angle θ_0. The incident wave vector $\vec{k}_0$ is defined by the incident angle θ_0.

$$\vec{k}_0 = \alpha_0\vec{u}_x + \beta_0\vec{u}_y \tag{4.30}$$

with $\alpha_0 = k\sin\theta_0, \beta_0 = k\cos\theta_0$. The surface could be periodic or non periodic. Here we consider periodic surfaces or periodic random surfaces.

We represent the vector function by its complex vector function and omit its time-dependence factor $exp(j\omega t)$. So for the horizontal ($E_{//}$) polarization and vertical ($H_{//}$)

polarization,

$$F_0(x,y) = exp(-j\alpha_0 x + j\beta_0 y) = \begin{cases} E_{0,z}(x,y) & E_{//} \\ Z_1 H_{0,z}(x,y) & H_{//} \end{cases} \tag{4.31}$$

and

$$Z_1 \vec{H} = \frac{\vec{k}_0}{k} \wedge \vec{E} \tag{4.32}$$

where $Z_1 = 120\pi \cdot Ohms$.

The reflected and transmitted plane waves can be written in a similar form. But, for rough surface, we have, in addition to the incident, reflected and transmitted plane waves, a scattered field $F(x,y)$ because of the deformation. The problem consists in working out the scattered field within the two media. The rough surface here is generated by simulation.

Equations (4.24) to (4.29) enable us to write Maxwell's equations associated with the constitutive relations:

$$\begin{cases} \frac{j}{k_1}\frac{\partial F(x,u)}{\partial u} = \frac{j}{k_1} b(x) \frac{\partial F(x,u)}{\partial x} + c(x) G(x,u) \\ \frac{j}{k_1}\frac{\partial G(x,u)}{\partial u} = \frac{1}{k_1^2}\frac{\partial}{\partial x}(c(x)\frac{\partial F(x,u)}{\partial x}) + \nu^2 F(x,u) \\ + \frac{j}{k_1}\frac{\partial}{\partial x}(b(x)G(x,u)) \end{cases} \tag{4.33}$$

with

$$b(x) = \frac{\frac{da}{dx}}{1 + (\frac{da}{dx})^2}, c(x) = \frac{1}{1 + (\frac{da}{dx})^2}$$

and in medium (m), $F(x,u) = F_m(x,u), G(x,u) = G_m(x,u), \nu = \nu_m, m = 1,2$. In $E_{//}$ polarization, $F(x,u) = E_{z'}(x,u), G(x,u) = Z_1 H_{x'}(x,u)$. In $H_{//}$ polarization, $F(x,u) = \frac{Z_1}{\nu} H_{z'}(x,u), G(x,u) = -\nu E_{x'}(x,u)$.

System (4.33) can be written in the form as follows:

$$\frac{j}{k_1}\frac{\partial \psi(x,u)}{\partial u} = \mathcal{L}\psi(x,u) \tag{4.34}$$

with

$$\mathcal{L} = \begin{pmatrix} \frac{j}{k_1} b(x) \frac{\partial \cdot}{\partial x} & c(x)\cdot \\ \frac{1}{k_1^2}\frac{\partial}{\partial x}(c(x)\frac{\partial \cdot}{\partial x}) + \nu^2 \cdot & \frac{j}{k_1}\frac{\partial b(x)\cdot}{\partial x} \end{pmatrix} \tag{4.35}$$

and

$$\psi(x,u) = \begin{pmatrix} F(x,u) \\ G(x,u) \end{pmatrix} \tag{4.36}$$

We separate the variables by writing $\psi(x,u) = \varphi(x)\kappa(u)$, then we get:

$$\frac{j}{k_1\kappa(u)}\frac{d\kappa(u)}{du} = \frac{\mathcal{L}\varphi(x)}{\varphi(x)} = r = constant \tag{4.37}$$

So we conclude that $\kappa(u) = Cexp(-jk_1ru)$, $\mathcal{L}\varphi(x) = r\varphi(x)$ and $\psi(x,u) = Aexp(-jk_1ru)\varphi(x)$. If the function $a(x)$ is a period function with D its period, then one has,

$$\begin{cases} a(x) = \sum\limits_m a_m exp(-j2\pi mx/D) \\ f(x) = \sum\limits_m f_m exp(-j\alpha_m x) \\ g(x) = \sum\limits_m g_m exp(-j\alpha_m x) \end{cases} \tag{4.38}$$

with $\varphi(x) = \begin{pmatrix} f(x) \\ g(x) \end{pmatrix}$ and $\alpha_m = k_1 sin\theta_0 + m\frac{2\pi}{D}$. Under this function decomposition, the eigenproblem $\mathcal{L}\varphi(x) = r\varphi(x)$ has a matrix form as follows:

$$[\mathcal{L}]\vec{\varphi} = r\vec{\varphi} \tag{4.39}$$

with

$$[\mathcal{L}] = \begin{pmatrix} [\mathcal{L}_{ff}] & [\mathcal{L}_{fg}] \\ [\mathcal{L}_{gf}] & [\mathcal{L}_{gg}] \end{pmatrix} and \quad \vec{\varphi} = \begin{pmatrix} \vec{f} \\ \vec{g} \end{pmatrix} \tag{4.40}$$

where $[\mathcal{L}_{ff}] = [C][\dot{A}][\tilde{\alpha}], [\mathcal{L}_{fg}] = [C], [\mathcal{L}_{gf}] = \nu^2[I] - [\tilde{\alpha}][C][\tilde{\alpha}], [\mathcal{L}_{gg}] = [\tilde{\alpha}][C][\dot{A}], [\dot{A}]_{pq} = \dot{a}_{p-q} = (p-q)\frac{2\pi}{d}a_{p-q}, [I]_{pq} = \delta_{p-q}, [C] = ([I]+[\dot{A}][\dot{A}])^{-1}, \tilde{\alpha}_p = \frac{\alpha_p}{k_1}, [\tilde{\alpha}]_{pq} = \delta_{p-q}\tilde{\alpha}_p, (\vec{f})_p = f_p, (\vec{g})_p = g_p, \forall(p,q) \in \mathbb{Z}^2$. Equations (4.39) and (4.40) give an eigenvalue system of infinite dimension. In a numerical computation, one can truncate it to a finite order problem with a truncation order M. Theoretically, increasing the truncation order M will increase the precision of results as well as increase the computational time. We want to ensure a certain precision and also keep M relatively small. For the lossless medium, i.e. the medium with optical index ν real, the power balance criterion (3.12) is checked to see if the truncation order M is large enough for a certain precision. The new system is similar to the original one except that now, we have $-M \leq p,q \leq M$.

By solving the truncated eigenvalue problem, with the eigenvalues r_n and eigenvectors $\vec{\varphi}_n, 1 \neq n \neq 2M+1$ one gets:

$$\begin{cases} F_n(x,u) = f_n(x)exp(-jk_1r_nu) \\ = \sum\limits_{-M\leq m\leq M} f_{mn}exp(-j\alpha_m x)exp(-jk_1r_nu) \\ G_n(x,u) = g_n(x)exp(-jk_1r_nu) \\ = \sum\limits_{-M\leq m\leq M} g_{mn}exp(-j\alpha_m x)exp(-jk_1r_nu) \end{cases} \tag{4.41}$$

So we are left with the eigenproblem of order $4M+2$. The signs of the real and imaginary parts of the eigenvalues r_n define the nature of the wave corresponding to the elementary wavefunction. In particular, the associated expression represent an outgoing wave propagating with no attenuation if $Re(r_n) > 0$ and $Im(r_n) = 0$. For an evanescent wave, $Im(r_n) < 0$. Finally, the field scattered in the air can be represented as a linear combination of all the solutions that verifies the outgoing conditions.

$$\psi^{(i)}(x,u) = \sum_{n=1}^{2M+1} C_n^{(i)}\psi_n^{(i)}(x,u), i = 1,2 \tag{4.42}$$

and the amplitudes $C_n^{(i)}$ are determined by solving the boundary conditions at $u = 0$ (i.e., at $y = a(x)$). The boundary conditions stipulate the continuity of the electric and magnetic components parallel to the surface. These components are $(H_{x'}, E_{z'})$ in $E_{//}$ polarization and $(E_{x'}, H_{z'})$ in $H_{//}$ polarization.

4.4 Formulation for the two-dimensional case

Now we consider the two-dimensional case. We assume the surface $z = a(x,y)$ is illuminated by a monochromatic plane wave with wavelength λ. $a(x,y)$ is a local function with L denotes the deformation length with respect to the Ox and Oy axis. For the formulation applied to crossed gratings, we refer the readers to [38, 39]. The incident wave vector $\vec{k}_0$ is defined by the zenith angle θ_0 and the azimuth angle φ_0.

$$\vec{k}_0 = \alpha_0\vec{u}_x + \beta_0\vec{u}_y - \gamma_0\vec{u}_z \tag{4.43}$$

with

$$\alpha_0 = k\sin\theta_0\cos\varphi_0, \beta_0 = k\sin\theta_0\sin\varphi_0, \gamma_0 = k\cos\theta_0 \tag{4.44}$$

and

$$k = \frac{2\pi}{\lambda} \tag{4.45}$$

For the $E_{//}$ polarization, the incident field can be expressed as:

$$\vec{E}_0(x,y,z) = \vec{h}e^{-j\vec{k}_0\vec{r}} \; and \; Z\vec{H}_0 = \frac{\vec{k}_0}{k} \times \vec{E}_0 \tag{4.46}$$

For the $H_{//}$ polarization, the incident field can be expressed as:

$$Z\vec{H}_0(x,y,z) = \vec{h}e^{-j\vec{k}_0\vec{r}} \; and \; Z\vec{H}_0 = \frac{\vec{k}_0}{k} \times \vec{E}_0 \tag{4.47}$$

Here

$$\vec{h} = -\sin\varphi_0\vec{u}_x + \cos\varphi_0\vec{u}_y \tag{4.48}$$

and

$$\vec{r} = x\vec{u}_x + y\vec{u}_y + z\vec{u}_z \tag{4.49}$$

We want to know the scattered field, but it cannot be expressed by the Rayleigh integral (3.43) in the modulated zone if the perturbation amplitude is too large. We can obtain an expression of field that is valid over the surface by solving Maxwell's equation in the translation coordinate system:

$$\begin{cases} x' = x \\ y' = y \\ z' = z - a(x,y) \end{cases} \tag{4.50}$$

In a source-free medium, from equations (4.24)-(4.29), we can obtain that the longitudinal components $E_{z'}$ and $ZH_{z'}$ obey to the propagation equation [39]:

$$\begin{aligned} &-\frac{\partial}{\partial z'}(g^{x'z'}\frac{\partial\psi}{\partial x'} + \frac{\partial g^{x'z'}\psi}{\partial x'}) - \frac{\partial}{\partial z'}(g^{y'z'}\frac{\partial\psi}{\partial y'} + \frac{\partial g^{y'z'}\psi}{\partial y'}) + jkg^{z'z'}\frac{\partial\psi'}{\partial z'} \\ &= \frac{\partial^2\psi}{\partial x'^2} + \frac{\partial^2\psi}{\partial y'^2} + k^2\psi \end{aligned} \tag{4.51}$$

with

$$\psi' = \frac{j}{k}\frac{\partial \psi}{\partial z'} \tag{4.52}$$

and $\psi(x', y', z') = E_{z'}(x', y', z')$ or $ZH_{z'}(x', y', z')$. And $g^{x'z'}$, $g^{y'z'}$ and $g^{z'z'}$ are elements of metric tensor which depend on the derivatives of function $a(x', y')$ with respect to x' and y'. From equation (4.13), we have:

$$\begin{cases} g^{x'z'} = -\frac{\partial a}{\partial x'} \\ g^{y'z'} = -\frac{\partial a}{\partial y'} \\ g^{z'z'} = 1 + (\frac{\partial a}{\partial x'})^2 + (\frac{\partial a}{\partial y'})^2 \end{cases} \tag{4.53}$$

Again from (4.24)-(4.29), we obtain expression of components $E_{x'}, E_{y'}, H_{x'}$ and $H_{y'}$ in terms of longitudinal components $E_{z'}$ and $ZH_{z'}$ only.

$$\frac{\partial^2 E_{x'}}{\partial z'^2} + k^2 E_{x'} = \frac{\partial^2 E_{z'}}{\partial x' \partial z'} - k^2 g^{x'z'} E_{z'} - jk g^{y'z'} \frac{\partial ZH_{z'}}{\partial z'} - jk \frac{\partial ZH_{z'}}{\partial y'} \tag{4.54}$$

$$\frac{\partial^2 E_{y'}}{\partial z'^2} + k^2 E_{y'} = \frac{\partial^2 E_{z'}}{\partial y' \partial z'} - k^2 g^{y'z'} E_{z'} - jk g^{x'z'} \frac{\partial ZH_{z'}}{\partial z'} + jk \frac{\partial ZH_{z'}}{\partial x'} \tag{4.55}$$

$$\frac{\partial^2 ZH_{x'}}{\partial z'^2} + k^2 ZH_{x'} = \frac{\partial^2 ZH_{z'}}{\partial x' \partial z'} - k^2 g^{x'z'} ZH_{z'} - jk g^{y'z'} \frac{\partial E_{z'}}{\partial z'} - jk \frac{\partial E_{z'}}{\partial y'} \tag{4.56}$$

$$\frac{\partial^2 ZH_{y'}}{\partial z'^2} + k^2 ZH_{y'} = \frac{\partial^2 ZH_{z'}}{\partial y' \partial z'} - k^2 g^{y'z'} ZH_{z'} - jk g^{x'z'} \frac{\partial E_{z'}}{\partial z'} + jk \frac{\partial E_{z'}}{\partial x'} \tag{4.57}$$

The covariant components $E_{x'}$ and $E_{y'}$ are parallel to the interface. Consequently, for instance, for a perfectly conducting surface, we have $E_{x'} = E_{y'} = 0$ at $z' = 0$. We need to solve the propagation equation (4.51).

To solve equation (4.51), we use a Fourier transform with respect to x' and y', then the equations take the following form

$$\begin{aligned} &\frac{\partial}{\partial z'}[j\alpha(\hat{g}^{x'z'} * \hat{\psi}) + j\hat{g}^{x'z'} * (\alpha\hat{\psi}) + j\beta(\hat{g}^{y'z'} * \hat{\psi}) + j\hat{g}^{y'z'} * (\beta\hat{\psi})] \\ &+ jk\hat{g}^{z'z'} * \frac{\partial \hat{\psi}'}{\partial z'} = \gamma^2 \hat{\psi} \end{aligned} \tag{4.58}$$

with

$$\frac{j}{k}\frac{\partial \hat{\psi}}{\partial z'} = \hat{\psi}' \tag{4.59}$$

and

$$\hat{\psi} = \int \psi(x,y) e^{j\alpha x} e^{j\beta y} d\alpha d\beta \tag{4.60}$$

Here, $\hat{K} * \hat{L}$ is the convolution product of two Fourier transforms $\hat{K}(\alpha,\beta,z')$ and $\hat{L}(\alpha,\beta,z')$. Now, convolution products are approximated as follows:

$$\begin{aligned}(\hat{K} * \hat{L})(\alpha,\beta,z') &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{K}(\alpha',\beta',z')\hat{L}(\alpha-\alpha',\beta-\beta')d\alpha' d\beta' \\ &\approx \frac{\Delta\alpha^2}{4\pi^2}\sum_p\sum_q \hat{K}(\alpha_p,\beta_q,z')\hat{L}(\alpha-\alpha_p,\beta-\beta_q)\end{aligned} \tag{4.61}$$

where

$$\alpha_p = k\sin\theta_0\cos\varphi_0 + p\Delta\alpha, \beta_q = k\sin\theta_0\sin\varphi_0 + q\Delta\alpha \tag{4.62}$$

and $\Delta\alpha = \Delta\beta = \frac{2\pi}{L}$ is the spectral resolution. Using this approximation and applying the point matching method at discrete values (α_s,β_t) to equation (4.58), we obtain

$$\begin{aligned}&\frac{j}{k}\frac{\partial}{\partial z'}\Big(\sum_{p,q}\Big(\frac{\alpha_s}{k}\hat{g}^{x'z'}_{s-p,t-q} + \hat{g}^{x'z'}_{s-p,t-q}\frac{\alpha_a}{k} + \frac{\beta_t}{k}\hat{g}^{y'z'}_{s-p,t-q} + \hat{g}^{y'z'}_{s-p,t-q}\frac{\beta_q}{k}\Big)\hat{\psi}(\alpha_p,\beta_q,z')\Big) \\ &+ \frac{j}{k}\frac{\partial}{\partial z'}\Big(\sum_{p,q}\hat{g}^{z'z'}_{s-p,t-q}\hat{\psi}'(\alpha_p,\beta_q,z')\Big) = \frac{\gamma_{st}^2}{k^2}\hat{\psi}(\alpha_s,\beta_t,z')\end{aligned} \tag{4.63}$$

$$\frac{j}{k}\frac{\partial\hat{\psi}(\alpha_s,\beta_t,z')}{\partial z'} = \hat{\psi}'(\alpha_s,\beta_t,z') \tag{4.64}$$

with

$$\hat{g}^{x'z'}_{p,q} = \frac{\Delta\alpha^2}{4\pi^2}\hat{g}^{x'z'}(\alpha_p,\beta_q) \tag{4.65}$$

$$\hat{g}^{y'z'}_{p,q} = \frac{\Delta\alpha^2}{4\pi^2}\hat{g}^{y'z'}(\alpha_p,\beta_q) \tag{4.66}$$

$$\hat{g}^{z'z'}_{p,q} = \delta_{pq} + \sum_{u,v}\hat{g}^{x'z'}_{p-u,q-v}\hat{g}^{x'z'}_{u,v} + \sum_{u,v}\hat{g}^{y'z'}_{p-u,q-v}\hat{g}^{y'z'}_{u,v} \tag{4.67}$$

Equation (4.63) can be written in matrix form

$$\frac{j}{k}[\mathcal{L}_l]\frac{\partial}{\partial z'}\begin{pmatrix}\vec{\psi}\\ \vec{\psi}'\end{pmatrix} = [\mathcal{L}_r]\begin{pmatrix}\vec{\psi}\\ \vec{\psi}'\end{pmatrix} \tag{4.68}$$

With a M-th order truncated approximation, the matrices $[\mathcal{L}_l]$ and $[\mathcal{L}_r]$ are $2M_s$-dimensional ones with $M_s = (2M+1)^2$.

The elementary solution of equation (4.68) is defined as follows

$$\begin{pmatrix} \vec{\psi}_{mn} \\ \vec{\psi}'_{mn} \end{pmatrix} = \begin{pmatrix} \vec{\phi}_{mn} \\ \vec{\phi}'_{nm} \end{pmatrix} exp(-jkr_{mn}z') \tag{4.69}$$

with

$$r_{mn}[\mathcal{L}_l] \begin{pmatrix} \vec{\phi}_{mn} \\ \vec{\phi}'_{mn} \end{pmatrix} = [\mathcal{L}_r] \begin{pmatrix} \vec{\phi}_{mn} \\ \vec{\phi}'_{mn} \end{pmatrix} \tag{4.70}$$

System (4.70) represents an eigenvalue problem, the size of which is $N = 2M_s$. Then, according to the sample theorem [76], the elementary wave functions $\hat{\psi}_{mn}(\alpha, \beta, z')$ and $\hat{\psi}'_{mn}(\alpha, \beta, z')$ can be constructed from ϕ_{mn} and ϕ'_{mn}

$$\begin{aligned} \hat{\psi}_{mn}(\alpha, \beta, z') =& exp(-jkr_{mn}z') \\ & \times \sum_{s=-M}^{s=M} \sum_{t=-M}^{t=M} \phi_{mn}(\alpha_s, \beta_t) sinc(\frac{\pi}{\Delta\alpha}(\alpha - \alpha_s)) sinc(\frac{\pi}{\Delta\alpha}(\beta - \beta_t)) \end{aligned} \tag{4.71}$$

$$\begin{aligned} \hat{\psi}'_{mn}(\alpha, \beta, z') =& exp(-jkr_{mn}z') \\ & \times \sum_{s=-M}^{s=M} \sum_{t=-M}^{t=M} \phi'_{mn}(\alpha_s, \beta_t) sinc(\frac{\pi}{\Delta\alpha}(\alpha - \alpha_s)) sinc(\frac{\pi}{\Delta\alpha}(\beta - \beta_t)) \end{aligned} \tag{4.72}$$

Finally, the Fourier transform of Oz-component is defined as a linear combination of M_s eigensolutions satisfying the outgoing wave condition:

$$\hat{\psi}_d(\alpha, \beta, z') = \sum_{(m,n)\in D_s} A_{mn} \hat{\psi}_{mn}(\alpha, \beta, z') \tag{4.73}$$

$$\hat{\psi}'_d(\alpha, \beta, z') = \sum_{(m,n)\in D_s} A_{mn} \hat{\psi}'_{mn}(\alpha, \beta, z') \tag{4.74}$$

Substituting $E_{z'} = 0$ and applying the same procedure in the spectral domain, we obtain the Fourier transforms of horizontal polarized transverse components:

$$\hat{\psi}^{(ha)}_{dT}(\alpha, \beta, z') = \sum_{(m,n)\in D_s} A^{(ha)}_{mn} \hat{\psi}^{(ha)}_{T,mn}(\alpha, \beta) exp(-jkr_{mn}z') \tag{4.75}$$

with

$$\hat{\psi}_{dT}^{(ha)}(\alpha,\beta,z') = \begin{pmatrix} \hat{E}_{dx'}^{(ha)}(\alpha,\beta) \\ \hat{E}_{dy'}^{(ha)}(\alpha,\beta) \\ Z\hat{H}_{dx'}^{(ha)}(\alpha,\beta) \\ Z\hat{H}_{dy'}^{(ha)}(\alpha,\beta) \end{pmatrix} \quad and \quad \hat{\psi}_{T,mn}^{(ha)}(\alpha,\beta) = \begin{pmatrix} \hat{E}_{x',mn}^{(ha)}(\alpha,\beta) \\ \hat{E}_{y',mn}^{(ha)}(\alpha,\beta) \\ Z\hat{H}_{x',mn}^{(ha)}(\alpha,\beta) \\ Z\hat{H}_{y',mn}^{(ha)}(\alpha,\beta) \end{pmatrix} \tag{4.76}$$

According to the sampling theorem [76], we write

$$\hat{\psi}_{T,mn}^{(ha)}(\alpha,\beta) = \sum_{s=-M}^{M}\sum_{t=-M}^{M} \hat{\psi}_{T,mn}^{(ha)}(\alpha_s,\beta_t) sinc(\frac{\pi}{\Delta\alpha}(\alpha-\alpha_s)) sinc(\frac{\pi}{\Delta\alpha}(\beta-\beta_t)) \tag{4.77}$$

where

$$\hat{E}_{x',mn}^{(ha)}(\alpha_s,\beta_t) = -k^2 \sum_{p=-M}^{M}\sum_{q=-M}^{M} \hat{g}_{s-p,t-q}^{y'z'} \phi'_{mn}(\alpha_p,\beta_q) - k\beta_t \phi_{mn}(\alpha_s,\beta_t) \tag{4.78}$$

$$\hat{E}_{y',mn}^{(ha)}(\alpha_s,\beta_t) = k^2 \sum_{p=-M}^{M}\sum_{q=-M}^{M} \hat{g}_{s-p,t-q}^{x'z'} \phi'_{mn}(\alpha_p,\beta_q) + k\alpha_s \phi_{mn}(\alpha_s,\beta_t) \tag{4.79}$$

$$Z\hat{H}_{x',mn}^{(ha)}(\alpha_s,\beta_t) = -k^2 \sum_{p=-M}^{M}\sum_{q=-M}^{M} \hat{g}_{s-p,t-q}^{x'z'} \phi_{mn}(\alpha_p,\beta_q) - k\alpha_s \phi'_{mn}(\alpha_s,\beta_t) \tag{4.80}$$

$$Z\hat{H}_{y',mn}^{(ha)}(\alpha_s,\beta_t) = -k^2 \sum_{p=-M}^{M}\sum_{q=-M}^{M} \hat{g}_{s-p,t-q}^{y'z'} \phi_{mn}(\alpha_p,\beta_q) - k\beta_t \phi'_{mn}(\alpha_s,\beta_t) \tag{4.81}$$

Taking $H_{z'} = 0$ and substituting $\hat{E}^{(ha)}$ by $Z\hat{H}^{(va)}$ and $Z\hat{H}^{(ha)}$ by $-\hat{E}^{(va)}$, we obtain the vertical components of magnetic and electric fields.

The scattering amplitudes $A_{mn}^{(ha)}$ and $A_{mn}^{(va)}$ are found by solving the boundary conditions. So, for an incident wave in (a) polarization, we can write

$$\begin{aligned} &E_{dx'}^{(1,ha)}(x',y',z') + E_{dx'}^{(1,va)}(x',y',z') - E_{dx'}^{(2,ha)}(x',y',z') - E_{dx'}^{(2,va)}(x',y',z') \\ &= -(E_{0x'}^{(a)}(x',y',z') + \rho_r^{(a)} E_{rx'}^{(a)}(x',y',z')) - \rho_t^{(a)} E_{tx'}^{(a)}(x',y',z') \end{aligned} \tag{4.82}$$

$$\begin{aligned} &E_{dy'}^{(1,ha)}(x',y',z') + E_{dy'}^{(1,va)}(x',y',z') - E_{dy'}^{(2,ha)}(x',y',z') - E_{dy'}^{(2,va)}(x',y',z') \\ &= -(E_{0y'}^{(a)}(x',y',z') + \rho_r^{(a)} E_{ry'}^{(a)}(x',y',z')) - \rho_t^{(a)} E_{ty'}^{(a)}(x',y',z') \end{aligned} \tag{4.83}$$

$$\begin{aligned} &H_{dx'}^{(1,ha)}(x',y',z') + H_{dx'}^{(1,va)}(x',y',z') - H_{dx'}^{(2,ha)}(x',y',z') - H_{dx'}^{(2,va)}(x',y',z') \\ &= -(H_{0x'}^{(a)}(x',y',z') + \rho_r^{(a)} H_{rx'}^{(a)}(x',y',z')) - \rho_t^{(a)} H_{tx'}^{(a)}(x',y',z') \end{aligned} \tag{4.84}$$

$$\begin{aligned} &H_{dy'}^{(1,ha)}(x',y',z') + H_{dy'}^{(1,va)}(x',y',z') - H_{dy'}^{(2,ha)}(x',y',z') - H_{dy'}^{(2,va)}(x',y',z') \\ &= -(H_{0y'}^{(a)}(x',y',z') + \rho_r^{(a)} H_{ry'}^{(a)}(x',y',z')) - \rho_t^{(a)} H_{ty'}^{(a)}(x',y',z') \end{aligned} \tag{4.85}$$

where $\rho_r^{(a)}$ and $\rho_t^{(a)}$ are the Fresnel reflection and transmission coefficients. After a Fourier transform, the point matching method is applied, then a $4M_s$-dimensional matrix system is obtained, the inversion of which leads to scattering amplitude $A_{mn}^{(ha)}$ and $A_{mn}^{(va)}$. These scattering amplitudes lead to the bistatic coefficients as defined in equation (3.60).

4.5 Conclusion

In this chapter, we show the C-method and how this C-method leads to eigenvalue problem. We present the formulations for both one-dimensional case and two-dimensional case. The computational time of the C-method is a key topic of our research.

The computational time of the C-method is mainly spent on the computation of eigenvalues and eigenvectors. We give a figure here to show the computation time of the C-method. The figure 4.1 shows the computational time of numerical experiment of one realisation. The perfectly conducting surface we consider is of 64 square wavelengths and $\sigma_a = \lambda$ and $l_x = l_y = 1.41\lambda$. The incident angle is chosen as $\theta_0 = 30°$ and $\varphi_0 = 0°$. The computational time varies as N^3 where $N = 2(2M+1)^2$ is the order of the considered matrix. In fact, from figure 4.1, we can see that we have approximately the computational time $t = a(2(2M+1)^2)^\alpha$, with $\alpha \approx 3.1$ a good fit. In terms of computational time, the C-method is not competitive with respect to fast integral method whose complexity is $\mathcal{O}(N \log N)$ [20–25]. The computational time is a weak point of the C-method, in particular, for analyzing rough random surfaces insofar as the average scattered intensity is estimated over results of several surface realizations (Monte-Carlo method). However, the strength of the C-method is that it leads to the eigensolutions of the scattering problem. It is an accurate method and it can be used as a reference for the analytical methods [77].

In the next chapter, we propose a parallel QR algorithm adapted to the C-method for reducing the computational time. The proposed method keeps the strength of C-method and improves the weak point of C-method.

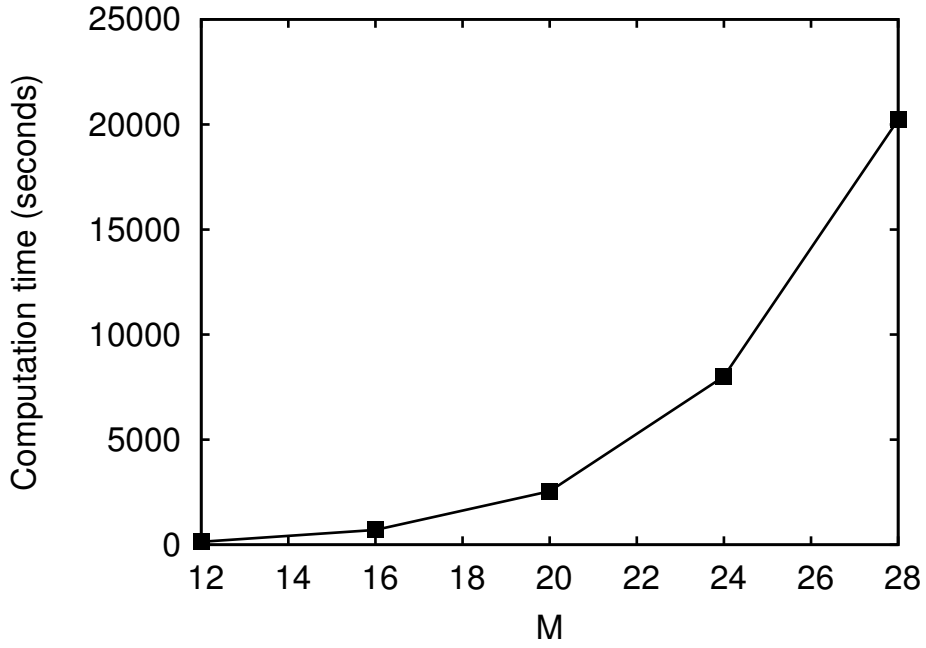

FIGURE 4.1: Computational time relative to truncation order

Chapter 5

Parallel QR algorithm for the C-method

5.1 Introduction

The most time-consuming part of C-method is to compute the eigenvalues and eigenvector of the scattering matrix A [58, 59]. Here, for the one-dimensional case, from equation (4.39), A equals the truncated version of $[\mathcal{L}]$, for the two-dimensional case, from equation (4.68), A equals the truncated version of $[\mathcal{L}_r]^{-1}[\mathcal{L}_l]$. Moreover, A is a large size, complex, dense and non-Hermitian matrix. Iterative eigensolvers, such as Krylov subspace methods or Jacobi-Davidson methods [66] have been developed to deal with large-scale eigenvalue problems. However, they have the possibility of missing some eigenvalues. So these iterative methods are ineffective for the C-method because all the eigenvalues and eigenvectors are needed. In contrast, the QR algorithm, which is based on similarity transformations, calculates all the eigenvalues and eigenvectors with very little danger, and only with a warning of missing some eigensolutions. We propose a specifically designed parallel QR algorithm for the C-method to save the computation time. Three techniques are used in the implementation of parallel QR algorithm: early shift, parallel bulge chasing and parallel aggressive early deflation (AED). The early shifts are introduced in parallel algorithm to give approximation of a part of the eigenvalues of the matrix. The early shifts are based on physical interpretation and observation based on the C-method and they are specifically designed and first introduced in our work. They

provide the possibility of quick deflation. For the bulge chasing, instead of only a single bulge, containing two shifts, a chain of several tightly coupled bulges, each containing two shifts, is chased in the course of one multishift QR iteration. As described in [78], this idea allows performing most of the computational work in terms of matrix-matrix multiplications to benefit from highly efficient level 3 BLAS. The idea of AED allows to detect converged eigenvalues much earlier than conventional deflation strategies. We will first present QR sequential algorithm and the shift strategy to accelerate the convergence, then we present all the parallel techniques for this specifically designed QR algorithm.

5.2 The basic QR algorithm

The QR algorithm computes a Schur decomposition of a matrix. It is certainly one of the most important algorithms in eigenvalue computations. As QR seems to be the only method that can provide us all the eigenvalues and eigenvectors, we choose to use the QR algorithm.

The QR algorithm consists of two separate stages. First, by means of a similarity transformation, the original matrix is transformed in a finite number of steps to Hessenberg form. This first stage of algorithm prepares its second stage, the actual QR iterations that are applied to the Hessenberg matrix. The overall complexity (number of floating points) of the algorithm is $\mathcal{O}(N^3)$ where the matrix A is assumed to be of the order $N \times N$.

We start with a basic iteration, given by algorithm 1. We notice that:

Algorithm 1 Basic QR algorithm

Input: $A \in \mathbb{C}^{N \times N}$ Output: An upper triangular matrix T and a unitary matrix U such that $A = UTU^*$ is the Schur decomposition of A.

1: Set $A_0 = A$ and $U_0 = I$.
2: **for** $k = 1, 2, 3, ...$ **do**
3: $\quad A_{k-1} = Q_k R_k$
4: $\quad A_k = R_k Q_k$
5: $\quad U_k = U_{k-1} Q_k$
6: **end for**
7: Set $T = A_\infty$ and $U = U_\infty$

$$A_k = R_k Q_k = Q_k^* A_{k-1} Q_k \tag{5.1}$$

and hence A_k and A_{k-1} are unitary similar. The matrix sequence $\{A_k\}$ converges (under certain assumptions) towards an upper triangular matrix [79]. Let us assume that the eigenvalues are pairwise different in magnitude and we can therefore number the eigenvalues such that $|\lambda_1| > |\lambda_2| > ... > |\lambda_N|$. Then the elements of A_k below the diagonal converge to zero like [79]:

$$|a_{ij}^{(k)}| = \mathcal{O}(|\frac{\lambda_i}{\lambda_j}|^k), i > j \tag{5.2}$$

From equation (5.1), we have:

$$A_k = Q_k^* A_{k-1} Q_k = Q_k^* Q_{k-1}^* A_{k-2} Q_{k-1} Q_k = Q_k^* ... Q_1^* A_0 Q_1 ... Q_k \tag{5.3}$$

With the same assumption on the eigenvalues, A_k tends to an upper triangular matrix and $U_k = Q_1 ... Q_k$ converges to the matrix of Schur vectors.

The convergence of the basic QR algorithm is slow and expensive. We want to:

- find a matrix structure that is preserved by the QR algorithm and that lowers the cost of a single iteration step.
- improve the convergence properties of the algorithm.

The desired matrix structure is a Hessenberg matrix: a matrix H is a Hessenberg matrix if its elements below the lower off-diagonal are zero, $h_{ij} = 0$ for $i > j+1$. The Hessenberg form is preserved by the QR algorithm and this form can lower the cost of a single iteration step [79].

There are several means of Hessenberg reduction such as Gram-Schmidt transformation, Householder reduction and Givens rotations [79]. An Efficient parallel algorithm for this Hessenberg reduction is implemented in the ScaLAPACK [80] (Scalable Linear Algebra PACKage) routine PZGEHRD. So, we will focus on the iterative part that comes after this Hessenberg reduction and try to improve the convergence properties of the algorithm.

5.3 QR algorithm with shift

We will show how the convergence of the Hessenberg QR algorithm can be improved dramatically by introducing spectral shifts into the algorithm.

Lemma 1. Let H be an irreducible Hessenberg matrix, i.e., $h_{i+1,i} \neq 0$ for all $i = 1, ..., N-1$. Let $H = QR$ be the factorization of H. Then for the diagonal elements of R, we have $|r_{kk}| > 0$, for all $k < N$. Thus, if H is singular then $r_{NN} = 0$.

This lemma gives the motivation of shift strategy to speed up the convergence of the QR algorithm. To see this, assume that λ is an eigenvalue of the irreducible Hessenberg matrix H. We perform:

Algorithm 2 The single shift QR algorithm (one iteration)

1: $H - \lambda I = QR$
2: $\overline{H} = RQ + \lambda I$

We can see that $\overline{H}$ is similar to H:

$$\overline{H} = Q^*(H - \lambda I)Q + \lambda I = Q^*HQ \tag{5.4}$$

By the lemma, we have:

$$H - \lambda I = QR, \ with \ R_{NN} = 0 \tag{5.5}$$

So,

$$\overline{H} = RQ + \lambda I = \begin{pmatrix} \overline{H}_1 & h_1 \\ 0 & \lambda \end{pmatrix} \tag{5.6}$$

So if we apply a QR step with a perfect shift to a Hessenberg matrix, the eigenvalue drops out. We then have a deflation, i.e. we can proceed the algorithm with a smaller matrix $\overline{H}_1$ of size $(N-1) \times (N-1)$.

For the single shift, when the item $h_{N-1,N-1}$ is $\mathcal{O}(h_{N,N-1})$, the convergence could be slow even the Rayleigh quotient shift gives a very good approximation (e.g. $h_{N,N-1}$ is very small). In practice, the double shift QR algorithm is very commonly used for real matrices and can be extended to complex matrices [81]. The algorithm is characterized by a "bulge chasing" procedure.

Suppose σ_1 and σ_2 are two shifts of the Hessenberg matrix H. The algorithm proceeds as follows:

1. Calculate the first column of the shift polynomial

$$v = (H - \sigma_1 I)(H - \sigma_2 I)e_1 = \begin{pmatrix} * \\ * \\ * \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{5.7}$$

2. Construct a 3×3 Householder transformation Q_1 such that the second and third entries of v are transformed to zero. The similarity transformation gives the updated matrix H_1:

$$H_1 = Q_1^* H Q_1 = \begin{pmatrix} * & * & * & * & * & * & \cdots \\ X & X & X & * & * & * & \cdots \\ X & X & X & * & * & * & \cdots \\ X & X & X & * & * & * & \cdots \\ 0 & 0 & 0 & * & * & * & \cdots \\ 0 & 0 & 0 & 0 & * & * & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \tag{5.8}$$

   The Hessenberg structure is damaged by the bulge that we denote with symbol "X".

3. Construct a 3×3 Householder transformation Q_2 such that the third and fourth entries of the first column of H_1 reduce to zero. The similarity transformation

gives the updated matrix H_2:

$$H_2 = Q_2^* H_1 Q_2 = \begin{pmatrix} * & * & * & * & * & * & \cdots \\ * & * & * & * & * & * & \cdots \\ 0 & X & X & X & * & * & \cdots \\ 0 & X & X & X & * & * & \cdots \\ 0 & X & X & X & * & * & \cdots \\ 0 & 0 & 0 & 0 & * & * & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \tag{5.9}$$

4. Continue similar operations to chase the bulge. In general, construct a 3×3 Householder transformation Q_k such that the $(k+1)$th and $(k+2)$th entries of the $(k-1)$th column of H_k are mapped to zero. Applying the corresponding similarity transformation to H_k results the updated matrix H_{k+1}, where $k = 2, 3, ..., N-1$. The bulge will be chased to vanish at the bottom right corner and lead to zeros, thus deflations. For example, H_3 will be look like as follows:

$$H_3 = Q_3^* H_2 Q_3 = \begin{pmatrix} * & * & * & * & * & * & \cdots \\ * & * & * & * & * & * & \cdots \\ 0 & * & * & * & * & * & \cdots \\ 0 & 0 & X & X & X & * & \cdots \\ 0 & 0 & X & X & X & * & \cdots \\ 0 & 0 & X & X & X & * & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \tag{5.10}$$

5.4 Early shift

Some of the eigenvalues of the scattering matrix can be approximated as follows: for the one-dimensional case,

$$r_n^{(m)} = \pm\sqrt{(\nu^{(m)})^2 - \tilde{\alpha}_n^2} \tag{5.11}$$

with $\tilde{\alpha}_n = \alpha_n / k^{(1)}$, and for the two-dimensional case,

$$r_{pq}^{(m)} = \pm\sqrt{(\nu^{(m)})^2 - \tilde{\alpha}_p^2 - \tilde{\beta}_q^2} \tag{5.12}$$

with $\tilde{\alpha}_p = \alpha_p/k^{(1)}, \tilde{\beta}_q = \beta_q/k^{(1)}$. The values $r_n^{(m)}$ or $r_{pq}^{(m)}$ constitute very good approximations when the index n or p, q are small relative to the matrix size. In general if $\nu^{(m)}$ is real or has a very small imaginary part, for the n that satisfies $\tilde{\alpha}_n^2 < Re(\nu^{(m)})^2$, or for the pair (p, q) that satisfies $\tilde{\alpha}_p^2 + \tilde{\beta}_q^2 < Re(\nu^{(m)})^2$, the approximation can be quite accurate. For a lossless medium (m), a real value $r_n^{(m)}$ can be associated with the angle defining a propagation direction with $\tilde{\alpha}_n = \sin\theta_n^{(m)}$, and $r_n^{(m)} = \pm\cos\theta_n^{(m)}$. The real eigenvalue $r_n^{(m)}$ defines the propagation direction of the associated eigenfunction. A similar interpretation exists for the two-dimensional case.

We take the one-dimensional case as an example to show why $r_n^{(m)}$ can be a good approximation. In fact, we can check that the functions

$$\tilde{F}_{\pm}(x, u) = exp(-jk^{(m)}\alpha_n x \pm jk^{(m)}r_n^{(m)}u) \tag{5.13}$$

satisfy the differential system (4.33). If we follow the equations from (4.33) to (4.37), we easily see that the set of $r_n^{(m)}$ is just the set of eigenvalues of (4.37). And the problem under consideration (4.39) is just a truncated form of the infinite dimensional eigenvalue problem (4.37). It is therefore attempting to take advantage of this analytical solution to represent the solution as a linear combination of $\tilde{F}_{\pm}$, however, if we operate in such a way, we can show that this method is equivalent to the well known Rayleigh expansion method, which leads to a numerical failure [28].

We therefore propose to use (5.11) and (5.4) as shifts. Moreover, according to our observation, we use only these approximations when $\nu^{(m)}$ is real or has a very small imaginary part, and n satisfies $\tilde{\alpha}_n^2 < Re(\nu^{(m)})^2$, or for the pair (p, q) that satisfies $\tilde{\alpha}_p^2 + \tilde{\beta}_q^2 < Re(\nu^{(m)})^2$. The approximations come in pair and each used pair will be used only once to create a 3×3 bulge.

We include an example for the one-dimensional case, where $M = 15, \nu = 1, \lambda = 1, d = 10.5, \theta = \frac{2\pi}{9}$. The surface used is a generated Gaussian rough surface with correlation length $l = \lambda$ and standard deviation of height is $\sigma = 0.2\lambda$. The eigenvalues of the scattering matrix are listed in the first column and the second column of the table 5.1. The eigenvalues in the first column correspond to the outgoing wave and the eigenvalues in the second column correspond to the incoming wave. The values of the $r_n = \pm\sqrt{\nu^2 - \tilde{\alpha}_n^2}$, where $-M < n < M$ are listed in the third column of the table 5.1:

TABLE 5.1: Comparison of eigenvalues and early shift

outgoing waves	incoming waves	$r_n, -M < n < M$
-0.1792 - 1.4278i	-0.1792 + 1.4278i	$\pm 1.8140i$
0.1933 - 1.4471i	0.1933 + 1.4471i	$\pm 1.7044i$
0.1126 - 1.4278i	0.1126 + 1.4278i	$\pm 1.5930i$
-0.1296 - 1.3333i	-0.1296 + 1.3333i	$\pm 1.4794i$
0.0582 - 1.2890i	0.0582 + 1.2890i	$\pm 1.3629i$
-0.0944 - 1.2226i	-0.0944 + 1.2226i	$\pm 1.2428i$
0.0360 - 1.1072i	0.0360 + 1.1072i	$\pm 1.1179i$
0.0113 - 0.9843i	0.0113 + 0.9843i	$\pm 0.9865i$
0.0006 - 0.8504i	0.0006 + 0.8504i	$\pm 0.8454i$
-0.0002 - 0.6893i	-0.0002 + 0.6893i	$\pm 0.6887i$
-0.0000 - 0.5021i	-0.0000 + 0.5021i	$\pm 0.5021i$
0.0000 - 0.2192i	0.0000 + 0.2192i	$\pm 0.2192i$
0.3713 - 0.0000i	-0.3713 + 0.0000i	± 0.3713
0.5529 + 0.0000i	-0.5529 - 0.0000i	± 0.5529
0.6748 - 0.0000i	-0.6748 - 0.0000i	± 0.6748
0.7660 + 0.0000i	-0.7660 - 0.0000i	± 0.7660
0.8368 - 0.0000i	-0.8368 - 0.0000i	± 0.8368
0.8919 - 0.0000i	-0.8591 - 0.0000i	± 0.8919
0.9341 - 0.0000i	-0.9342 - 0.0000i	± 0.9341
0.9652 - 0.0000i	-0.9652 + 0.0000i	± 0.9651
0.9861 + 0.0000i	-0.9862 - 0.0000i	± 0.9860
0.9975 + 0.0000i	-0.9976 - 0.0000i	± 0.9975
0.9996 - 0.0000i	-0.9996 + 0.0000i	± 0.9997
0.9924 + 0.0000i	-0.9922 - 0.0000i	± 0.9929
0.9758 - 0.0000i	-0.9749 + 0.0000i	± 0.9768
0.9478 + 0.0000i	-0.9474 + 0.0000i	± 0.9509
0.9044 - 0.0000i	-0.9094 - 0.0000i	± 0.9144
0.8532 + 0.0000i	-0.8919 - 0.0000i	± 0.8660
0.7613 + 0.0000i	-0.7572 - 0.0000i	± 0.8035
0.6781 - 0.0000i	-0.6757 + 0.0000i	± 0.7233
0.5713 + 0.0000i	-0.5711 + 0.0000i	± 0.6185

Based on this observation, the expression of $r_n^{(m)}$ or $r_{pq^{(m)}}$ can be used as shifts to approximate eigenvalues of the scattering matrix. In fact, if we increase the truncation order M, we can see that the approximations are better for relatively small n and large M [28]. The convergence is very quick due to the good approximation. We call this the "early shift". The "early shift" is used in pairs corresponding to the double shift QR algorithm. One pair of the "early shift" will create a bulge to be chased. Wilkinson's shift can be used after the "early shift".

5.5 Parallel QR with tightly coupled bulge chasing

The parallel bulge chasing algorithm was proposed by Bai, Demmel [78] and Braman et al. [67]. In order to benefit from the level 3 BLAS, they parallelize the bulge chasing procedure by performing the chasing of multiple chains of tightly coupled bulges. With the delay and accumulate technique, the main computation work become the matrix-matrix multiplications. The procedure of intrablock chasing and interblock chasing are described below.

To describe the parallel algorithm, we first introduce the data layout mapping in a distributed memory environment as follows:

- The $p = p_r p_c$ processors are arranged into a $p_r \times p_c$ grid. Usually the values of p_r and p_c are set to be as close as possible.
- The $N \times N$ matrix A is partitioned in $2D$ block cyclic scheme [82] and is mapped on $p_r \times p_c$ grid as shown in table 5.2. The table shows a 4 grid with $p_r = p_c = 2$. The four processors are denoted as $(0,0), (0,1), (1,0), (1,1)$. The block size is $M_b \times N_b$ and we require the block to be square $M_b = N_b$. Generally, a processor will store a collection of non-contiguous blocks. In table 5.2, if the size of matrix $N = 16$, then the block size is 4×4 and processor $(0,0)$ will store the elements $A(1:4, 1:4), A(1:4, 9:12), A(9:12, 1:4), A(9:12, 9:12)$. An array descriptor stores the details of data layout. The mapping between entries of the global matrix and their corresponding locations in the memory can be established from the array descriptor.

TABLE 5.2: 2D block cyclic scheme

(0,0)	(0,1)	(0,0)	(0,1)
(1,0)	(1,1)	(1,0)	(1,1)
(0,0)	(0,1)	(0,0)	(0,1)
(1,0)	(1,1)	(1,0)	(1,1)

Locally, each processor in the mesh may also utilize multithreading. This can be seen as adding another level of explicit parallelization by organizing all the $p = p_r \times p_c$ processors into a three-dimension mesh.

We use the shifts that are mentioned earlier to introduce the chain of bulges into diagonal blocks. Each of the chains reside on a different diagonal block. We choose the number of shifts such that each chain covers at most half of the data layout block. The "early shift" is distributed from left-upper diagonal blocks to right-lower diagonal blocks. Each "early shift" is used once. When there are no "early shift" to distribute, we use Wilkinson's shift.

For the intrablock chasing where the chain is chased from the top left corner to the lower right corner within a contiguous diagonal block. We may use a sequence of 3×3 Householder transformations to chase the chain of bulges down some rows to the down right-hand corner of the contiguous diagonal block. We start from the lowest bulge of the block and chase one bulge at a time. The intrablock chasing can be performed locally on the process that own this chain and simultaneously between different diagonal contiguous blocks which saves computation time. Figure 5.1 shows how the intrablock chasing are performed, the grey bulges are chased to the black bulges. After the bulge chasing within the diagonal block, the accumulated unitary matrices are sent to the corresponding processors in order to update the off-diagonal blocks. The off-diagonal blocks are then updated by matrix-matrix multiplications which uses level 3 BLAS. The broadcasts are sent in parallel. In order to avoid conflicts in the intersecting parts, they are performed first in the row direction and then in the column direction. See figure 5.1 for the intrablock chasing, here $(p_r, p_c) = (2, 2)$ and $M_b = N_b = 20$.

For the interblock chasing where a chain of bulges from one contiguous diagonal block is chased to a different one on another processor, for each contiguous diagonal blocks in which the bulge chains reside, we create copies of its neighbors and it becomes similar to the case of intrablock chasing. Figure 5.2 illustrates the procedure with $(p_r, p_c) = (2, 2)$ and $M_b = N_b = 20$, the grey bulges are chased to the black bulges. More precisely, the processor that stores the grey bulges create a copy of the block on each side of the border. Then we can perform the chasing locally, just as in the intrablock chasing and broadcast the corresponding orthogonal factors to the blocks on both sides of the cross border. The updated neighboring block are sent to its owner. To update the corresponding off-diagonal blocks, we broadcast orthogonal matrix accumulated in the diagonal chasing stage to the corresponding rows/columns of processors which are involved in off-diagonal updating. Then each involved processor exchanges data blocks

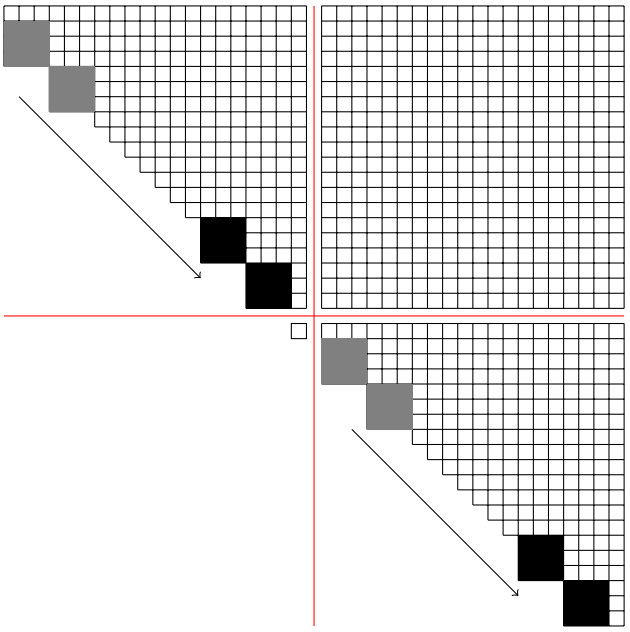

FIGURE 5.1: Intrablock parallel bulge chasing. The grey bulges are chased to the black bulges.

with its neighbor as illustrated in figure 5.2 in the two large gray blocks. The off-diagonal blocks are then updated by multiplication with the accumulated orthogonal matrix. We perform interblock chasing first for the odd-numbered blocks and then for the even-numbered blocks. This odd-even manner avoids conflicts between different tightly coupled chains [68].

For each diagonal block, the corresponding orthogonal transformations are accumulated into an orthogonal factor. Each orthogonal has the following shape:

$$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \tag{5.14}$$

where U_{12} is a lower triangular matrix, and U_{21} is a upper triangular matrix. So matrix multiplication by U will be broken into two dense by dense matrix multiplications and two triangular by dense matrix multiplications. Computation time is saved because of the triangular structure.

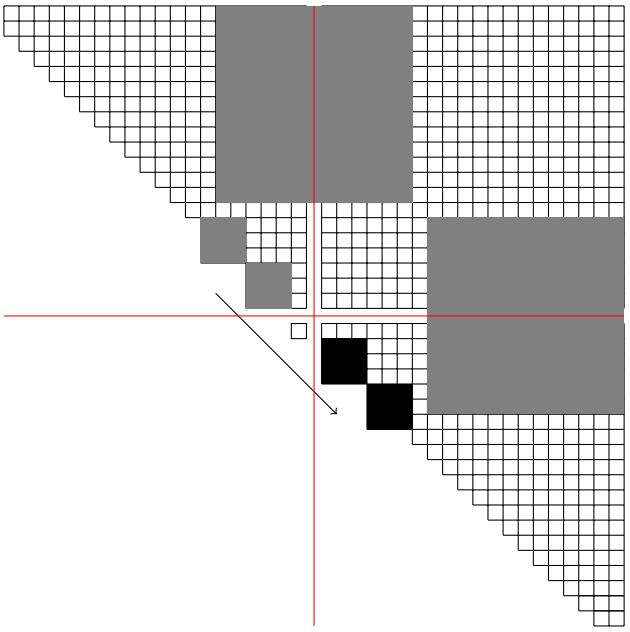

FIGURE 5.2: Interblock parallel bulge chasing. The gray bulges are chased to the black bulges.

5.6 Parallel AED

The parallel aggressive early deflation (AED) algorithm was proposed in [69]. We divide the Hessenberg matrix H as follows:

$$H = \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ 0 & H_{32} & H_{33} \end{pmatrix} \tag{5.15}$$

where H_{11} is of size $(n-k-1) \times (n-k-1)$ and H_{33} is of size $k \times k$. We use the pipeline parallel QR algorithm to find the Schur decomposition of H_{33}: $H_{33} = VTV^*$ and perform the following similarity transformation:

$$\begin{aligned} \begin{pmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & V \end{pmatrix}^* \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ 0 & H_{32} & H_{33} \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & V \end{pmatrix} \\ = \begin{pmatrix} H_{11} & H_{12} & H_{13}V \\ H_{21} & H_{22} & H_{23}V \\ 0 & s & T \end{pmatrix} \end{aligned} \tag{5.16}$$

Now the matrix looks like as in figure 5.3. The spike s is denoted as the gray part as in the figure 5.3.

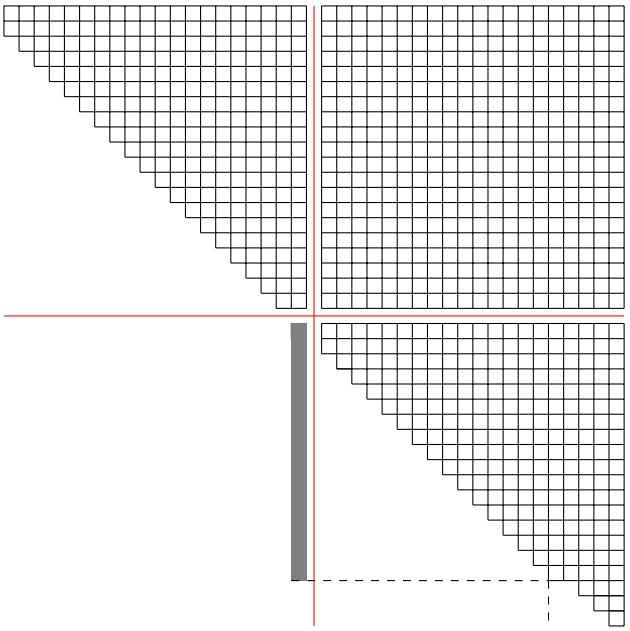

FIGURE 5.3: Aggressive early deflation. The gray spike contains the vector s

It has been proved that it is often the case that the some of the last components of s are very small [69]. If it is the case that the trailing several components of s are negligible, they can set to be zero. The matrix is deflated. This technique often detects convergence much earlier. If it is not the case, we move the eigenvalues to the top left corner of the block H_{33}.

5.7 Conclusion

In this chapter, we propose the specifically designed parallel QR algorithm for the C-method. We present why we propose the "early shift" and how it can be used to accelerate the convergence. We also present the techniques of parallel QR with tightly coupled bulge chasing and parallel AED. These techniques are used to reduce the computational time of the C-method.

In the next chapter, we will use these techniques including "early shift", parallel QR with tightly coupled bulge chasing and parallel AED to analyze gratings, one-dimensional and two-dimensional surfaces. It is a real novelty in the context of the C-method.

Chapter 6

Numerical experiments with parallel QR algorithm

6.1 Hardware and software platforms

In this chapter, we implement the algorithm described in the previous chapter. The numerical results are presented. The experiments are performed on the machine *Poincare*, hosted by IDRIS national computing center in France (http://www.idris.fr). This machine is an IBM computer, composed by mainly iDataPlex dx360 M4 servers:

- 92 nodes ("$poincare[001-092]$") equipped with:
  - 2 Sandy Bridge E5-2670 processors (2.60GHz, 8 cores every processor, 16 cores every node)
  - 32 GB memory every node
- 4 nodes GPU ("poincaregup[001-004]", separate from the above 92 nodes) equipped with:
  - 2 Sandy Bridge E5-2670 processors
  - 64 GB memory every node
  - 2 Tesla K20 GPU (Cuda Capability 3.5, 4.8 GB memory every GPU)
- 4 interactive login nodes equipped with:

- 2 Sandy Bridge E5-2670 processors
- 32 GB memory every node

We make use of the 92 nodes and login nodes. We make use of the following libraries: mkl 11.0, intelmpi 4.0.3, lapack 3.5_*gnu*47. The program is written in Fortran and compiled with the following settings:

$FC = mpif90$, $CC = mpicc$, $NOOPT = -00$, $FCFLAGS = -03$, $CCFLAGS = -03$, $FCLOADER = \$(FC)$, $CCLOADER = \$(CC)$, $FCLOADFLAGS = \$(FCFLAGS)$, $CCLOADFLAGS = \$(CCFLAGS)$

Our implementation is based on an simple imitation of the ScaLAPACK routine PDHSEQR we try to use this to a complex implementation with our own shift strategy. We adopt the recommended values such as the size of the deflation window and the tuning parameter $NIBBLE$ which determines when to skip a QR sweep and perform AED in [68]. For all the following experiments, we use block factor $M_b = N_b = 50$.

6.2 Numerical results for one-dimensional case

For the one-dimensional case, we compare the parallel algorithm with the pipeline parallel algorithm which is implemented in ScaLAPACK routine PZLAHQR. The routine PZLAHQR is also compiled and built on the machine *Poincare* with the same compiler and flags as our program. We also compare the parallel algorithm with or without the "early shift". In the following experiment, we have the physical model parameters (see section 3.1.2 for details): $\nu = 1, D = 200\lambda, \theta = 40°$. The surface used is a generated Gaussian rough surface with correlation length $l = 3\lambda$ and standard derivation of height λ. We set $M = 1000$, so the matrix size is 4002×4002. In fact, we have performed experiments with different M to check the error on the power balance. If we denote $error = 1 - \sum \epsilon_n$, see (3.12), figure 6.1 shows how the function $-\log_{10}(|error|)$ changes with the truncation order M. It shows if we require a precision of 10^{-2}, it should be enough to set $M = 1000$.

The accuracy of the early shift can be seen in Table 5.1, page 78, as an example. A plot will be difficult for recognition. The actual eigenvalues that are real are very well

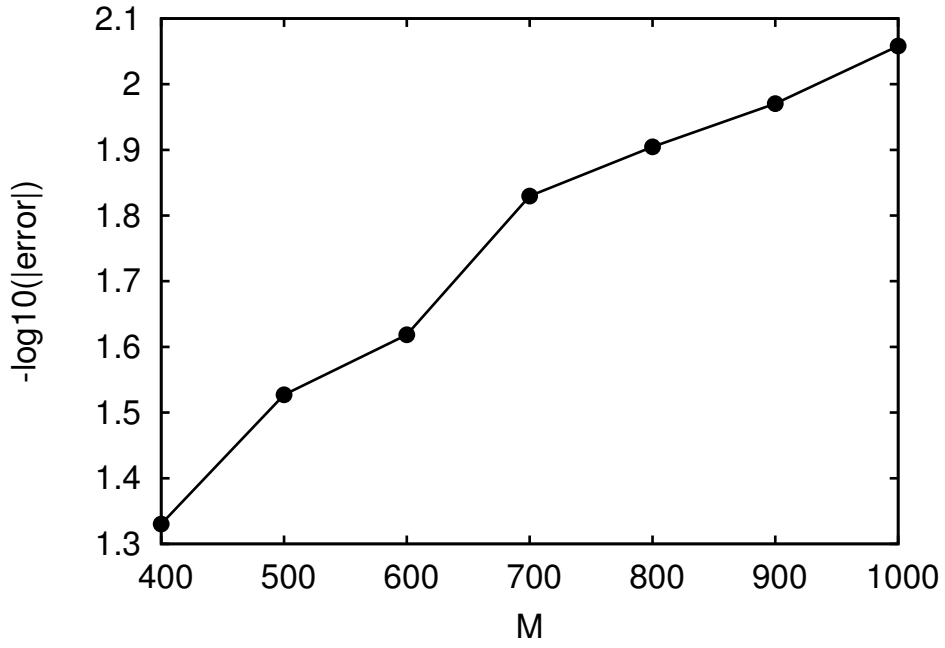

FIGURE 6.1: Function $-\log_{10}(|error|)$ relative to M.

approximated, but in the picture, all one can see is blue points on the real axis. We present the Figure 6.2 and 6.3.

Figure 6.4 shows the comparison between the pipeline parallelization and the parallelization with parallel multishift and AED techniques. The item Pipeline parallel represents the performance of PZLAHQR. For our parallel algorithm, the update stage takes approximately 40% of the total time, the Hessenberg translation stage takes approximately 30% to 40% of the total time, the AED stage takes approximately 15% to 25% of the total time and the chasing stage takes approximately 5% to 10% of the total time.

Figure 6.5 shows that the early shift does speed up the convergence and save the computation time. With early shift, the computational time of parallel multishift and AED decreases approximately 16%.

We also performed an experiment where M and thus the size of the matrix is changed. The number of cores used here is 16 (i.e. 4×4). Figure 6.6 shows the comparison between the pipeline parallelization and the parallelization with parallel multishift and AED techniques. Figure 6.6 shows that the new version of parallel QR algorithm with parallel multishift and AED is much faster than the existing parallel QR algorithm.

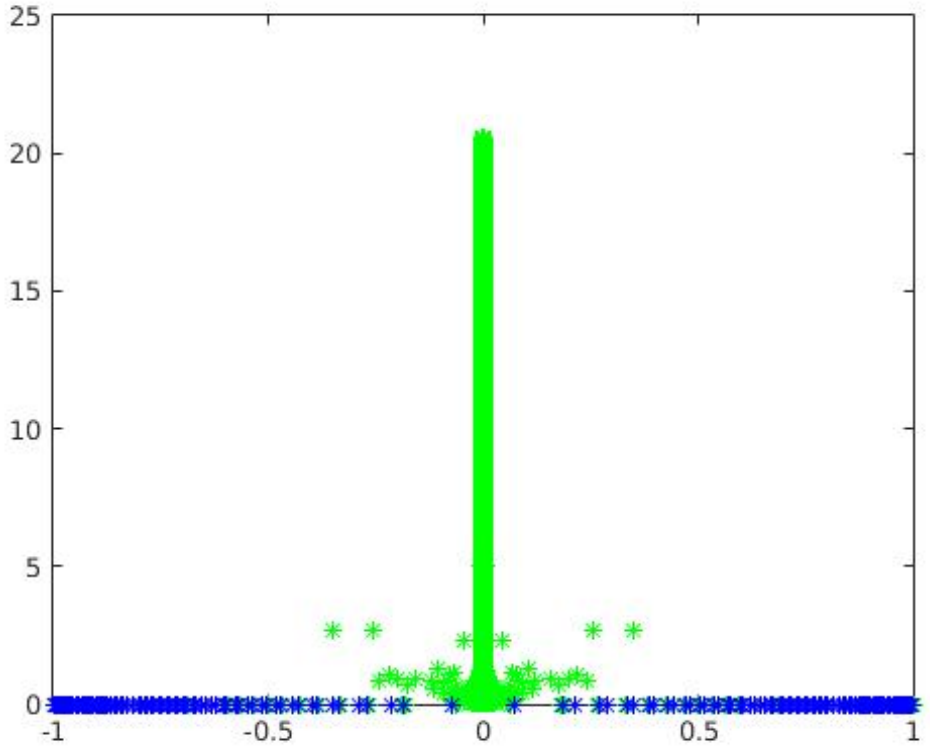

FIGURE 6.2: The green points represent actual eigenvalues, the blue points represent the used early shifts

Figure 6.7 shows that the early shift does speed up the convergence and save the computation time.

6.3 Numerical results for two-dimensional case

For the two-dimensional case, we performed the following experiments. To compare the computation time of sequential algorithm and the specifically parallel algorithm, we present figure 6.8. This figure is based on experimental result of one realisation. The surface we consider is perfectly conducting. The area of the surface is 64 square wavelengths and $\sigma = \lambda^{(1)}$ and $l_x = l_y = 1.41\lambda^{(1)}$. The incident angle is chosen as $\theta_0 = 30°, \varphi_0 = 0°$. The truncation order is $M = 28$, so the matrix size is $N = 6498$. In term of power balance criterion, the truncation order $M = 28$ gives good enough results (the error is smaller than 1%).

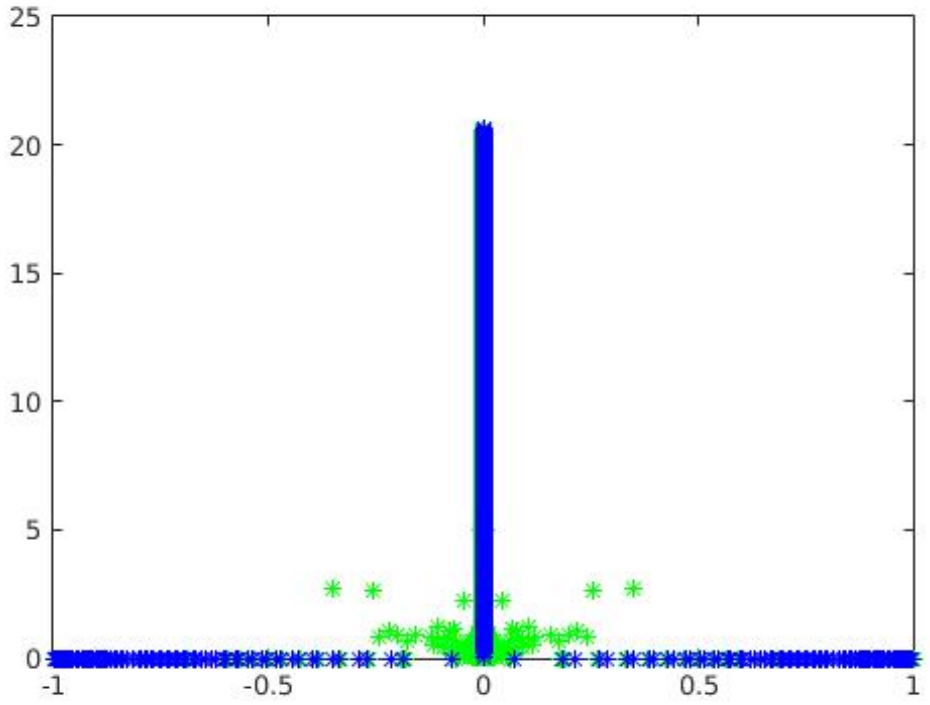

FIGURE 6.3: The green points represent actually eigenvalues, the blue points represent the values from equation (5.11)

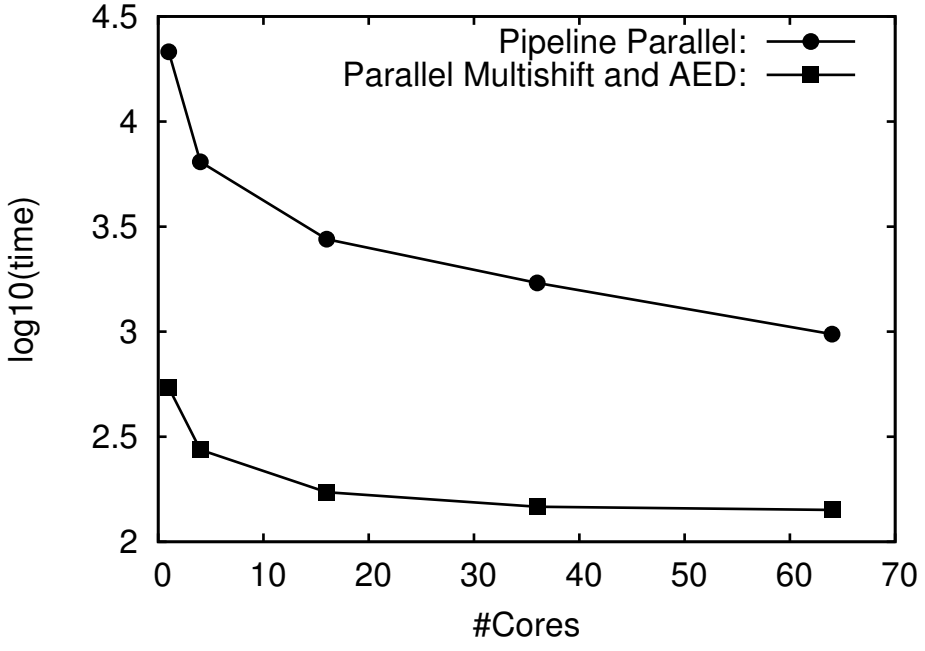

FIGURE 6.4: Computation time of two different parallelizations relative to number of cores

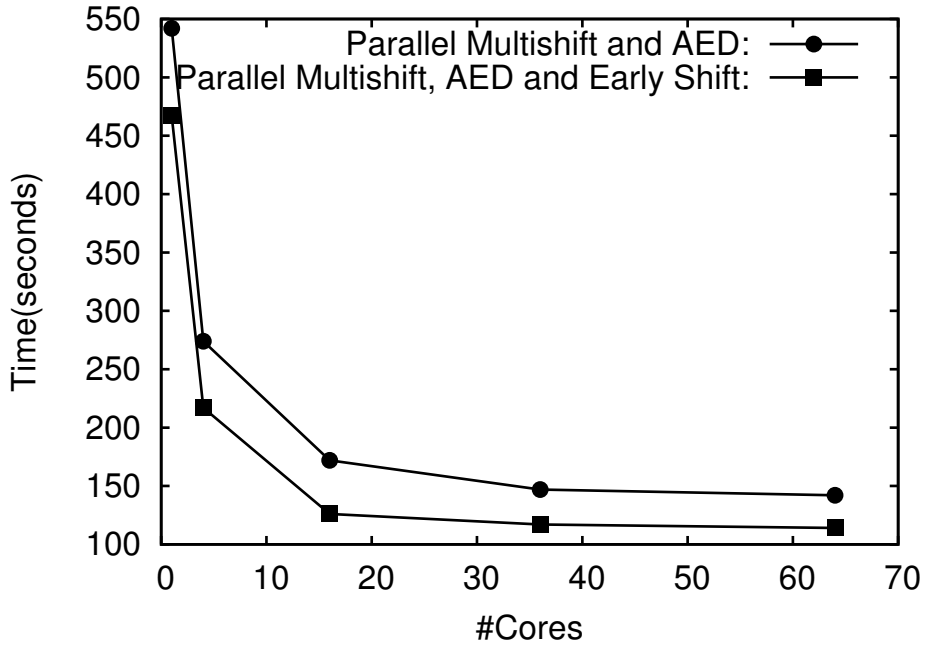

FIGURE 6.5: Computation time with or without early shift relative to number of cores

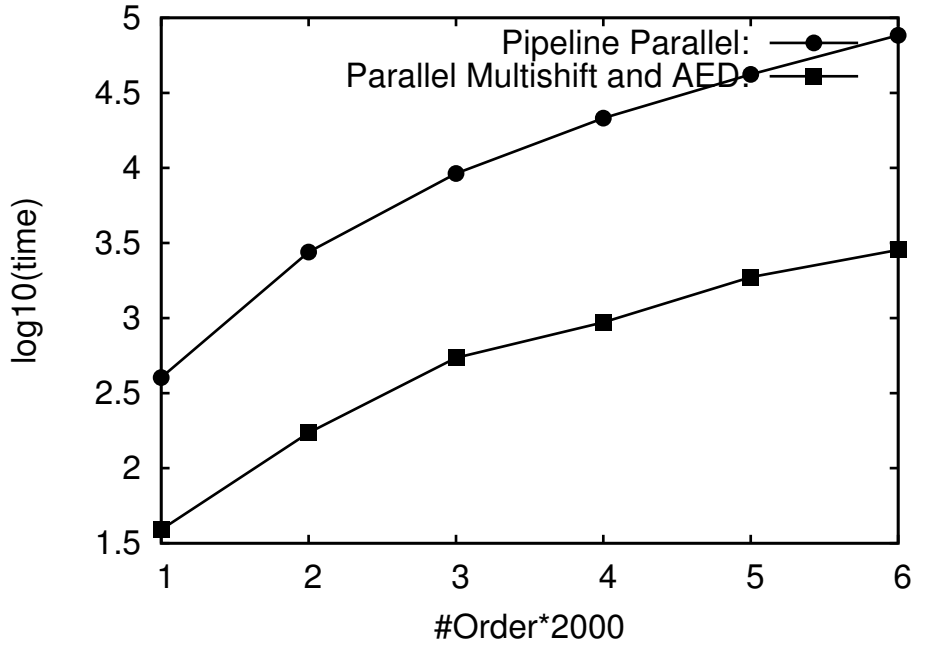

FIGURE 6.6: Computation time of two different parallelizations relative to order of matrix

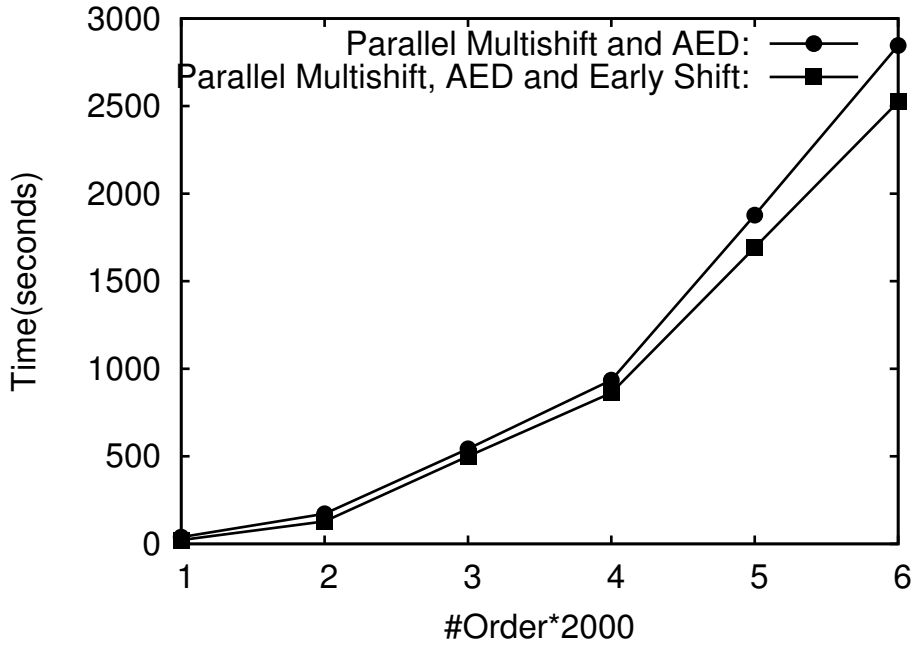

FIGURE 6.7: Computation time with or without early shift relative to order of matrix

Figure 6.8 gives the results of the pairs $(p_r, p_c) = (1,1), (2,2), (4,4), (6,6), (8,8)$. The left most point on the curve for the parallel performance represents that of one core. From this figure, we can see if we have 12800 cores, and we simulate the problem for $N_R = 200$ times. With the naive parallel strategy (we choose N_R cores and perform a simulation on each chosen core), it will cost us approximately 5.5 hours. With the new version of parallel strategy (we use 64 cores for one simulation), it will cost us approximately only 8 minutes. This shows when we have many cores, the new version of parallel strategy can be significantly more efficient than the naive parallel strategy and if we only have a few cores, the difference may not so great.

We also compare the computation cost of sequential algorithm and parallel algorithm when the truncation order M is varying. Figure 6.9 shows this comparison based on one realisation. The parameters of surface are the same as above. For the parallel realisation, the number of cores is fixed to be 16, $(p_r, p_c) = (4,4)$. These curves show that the reduction of computation time is important. For instance, the ratio is close to 25 when $M = 28$. We can see from figure 6.9, the computational time of the sequential code is approximately $t = a(2(2M+1)^2)^{\alpha_1}$ with $\alpha_1 \approx 3.2$ and the computational time of the parallel code is approximately $t = b(2(2M+1)^2)^{\alpha_2}$ with $\alpha_2 \approx 2.2$. While these two relations are only approximation from observing the data, they show that how the

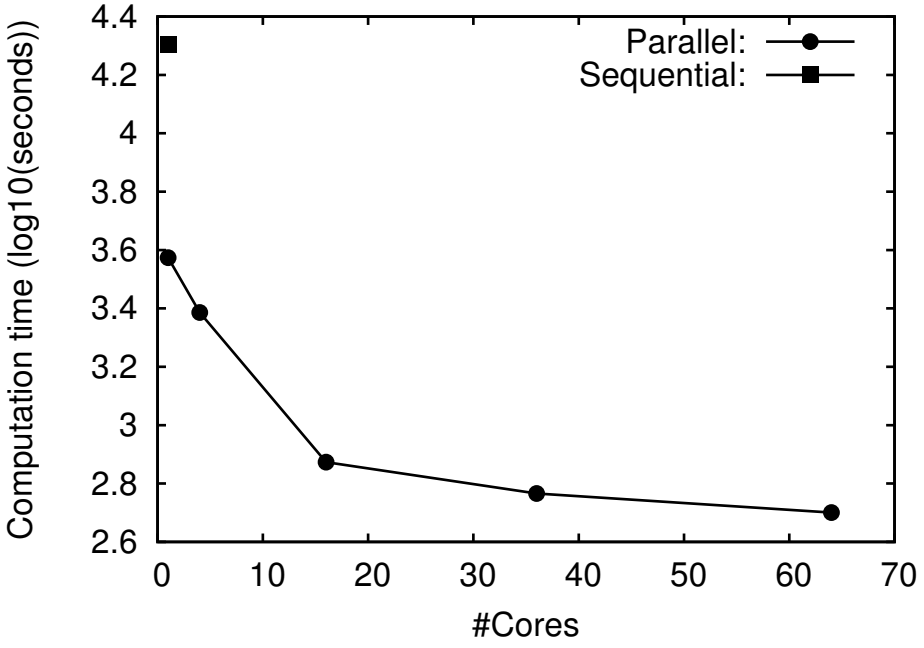

FIGURE 6.8: Comparison of computing time, sequential code relative to parallel code with respect to number of cores

computational time is changed using the parallel code within a certain range of matrix order.

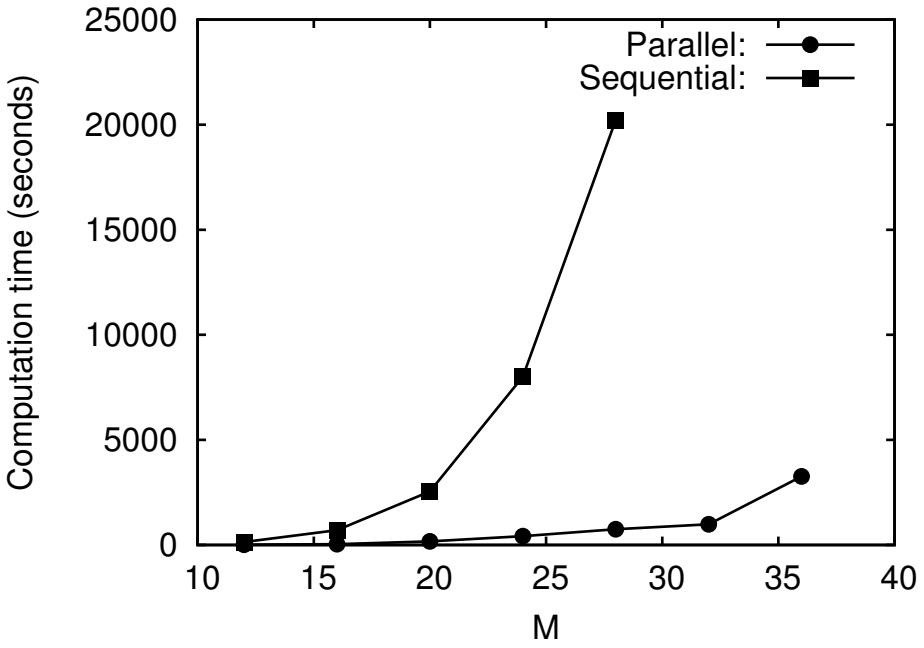

FIGURE 6.9: Comparison of computing time, sequential code relative to parallel code with respect to truncation order

We compare the results of numerical experiments with experimental data from literature.

6.4 Comparison with experimental data for random rough surfaces

We consider an isotropic surface with $\sigma_a = 0.352\lambda^{(1)}$ and $l_x = 2.21\lambda^{(1)}$. The optical index of the lower medium is $\nu^{(2)} = 1.62 - 0.001i$. The other simulation parameters are: $D = 8\lambda^{(1)}, \theta_0 = 35°$ or $\theta_0 = 55°, \varphi_0 = 0°, M = 28$ and the number of realizations is $N_R = 200$. The following figures 6.10-6.13 show results of the implementation of parallel C-methods compared with the experimental data which come from [83]. In these figures, the DRC which stands for differential reflection coefficient is plotted versus the scattering angle. It is noteworthy that the figure 6.13 presents a minimum similar to the Brewster angle for a planar surface. (By analogy with reflection from a smooth surface, a lossless dielectric with a refractive index equal to 1.62 provides a Brewster angle close to 58°.) The comparison with experimental data is excellent.

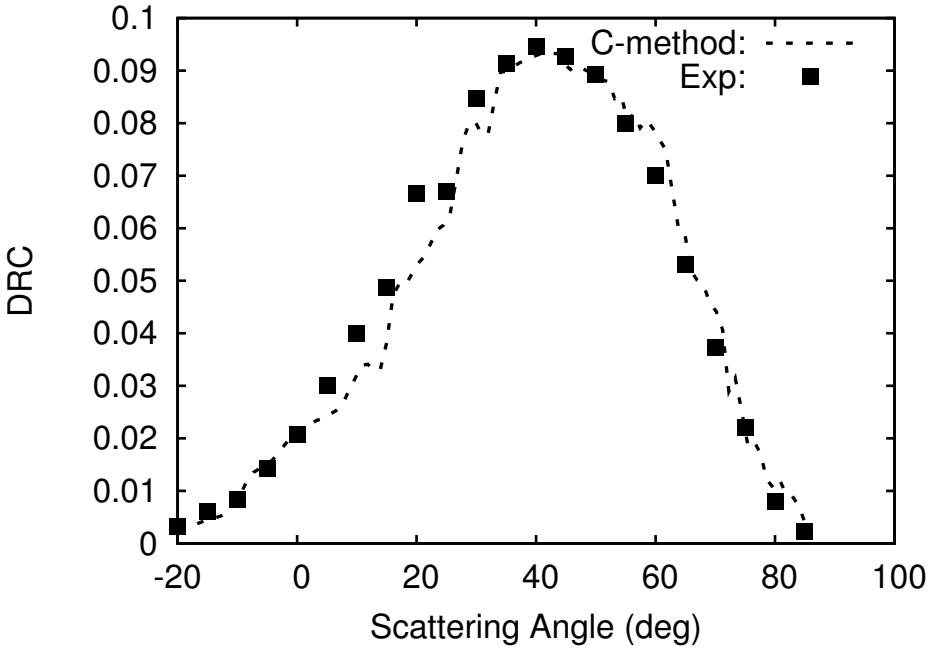

FIGURE 6.10: Differential reflection coefficient versus observation angle in the incidence plane, $\theta_0 = 35°$, polarization(hh)

We then present the figures 6.14- 6.17 which show results of the implementation of parallel C-methods compared with the experimental data which come from [84]. In these figures, the bistatic coefficient is plotted versus the observation angle. The perfectly conducting surface under consideration is a very rough surface with $\sigma_a = \lambda^{(1)}$ and $l_x = 1.41\lambda^{(1)}$. The other simulation parameters are: $D = 8\lambda^{(1)}$, $\theta_0 = 20°, \varphi_0 = 0°, M = 28$

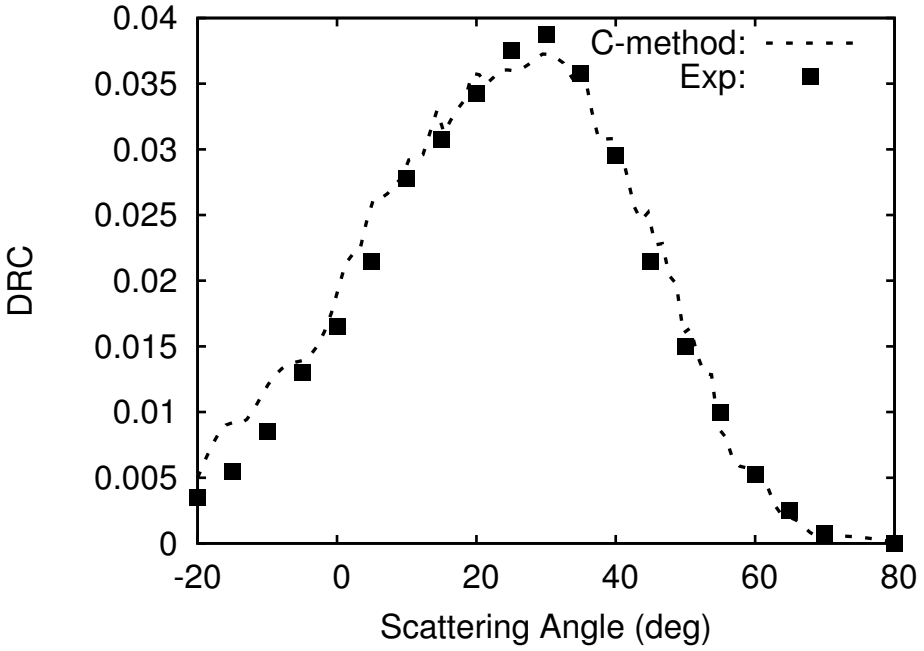

FIGURE 6.11: Differential reflection coefficient versus observation angle in the incidence plane, $\theta_0 = 35°$, polarization(vv)

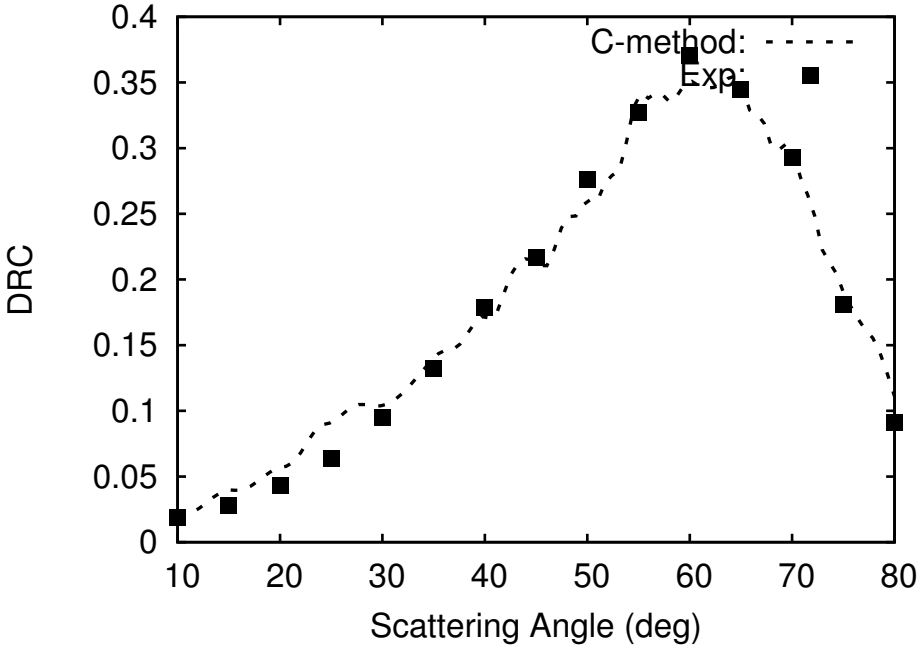

FIGURE 6.12: Differential reflection coefficient versus observation angle in the incidence plane, $\theta_0 = 55°$, polarization(hh)

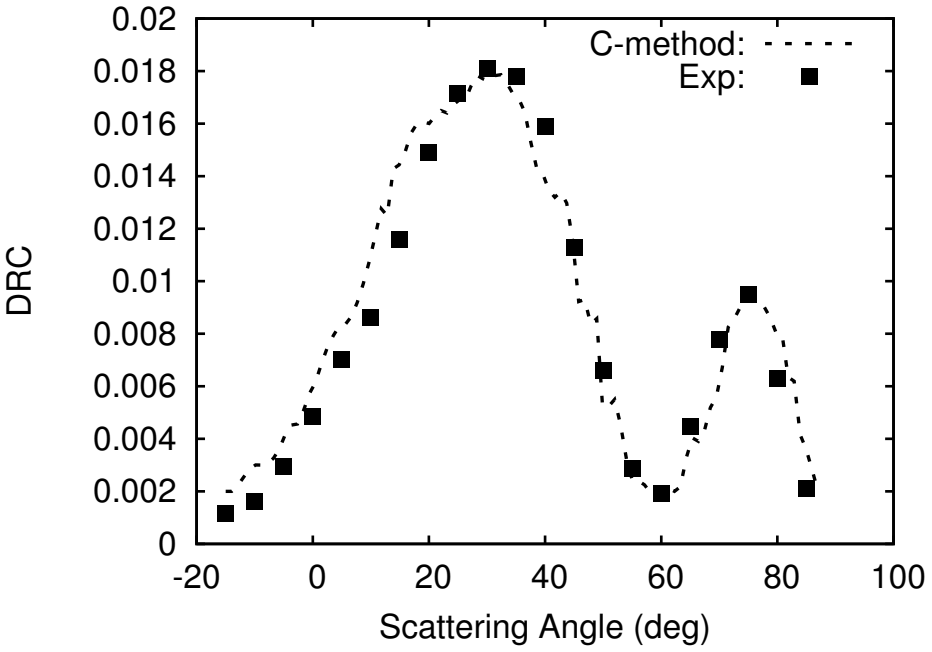

FIGURE 6.13: Differential reflection coefficient versus observation angle in the incidence plane, $\theta_0 = 55°$, polarization(vv)

and $N_R = 200$. It is noteworthy that the surface exhibits backscattering enhancement in both co-polarized and cross-polarized returns. The comparison with experimental data is very good. The backscattering peaks coincide well.

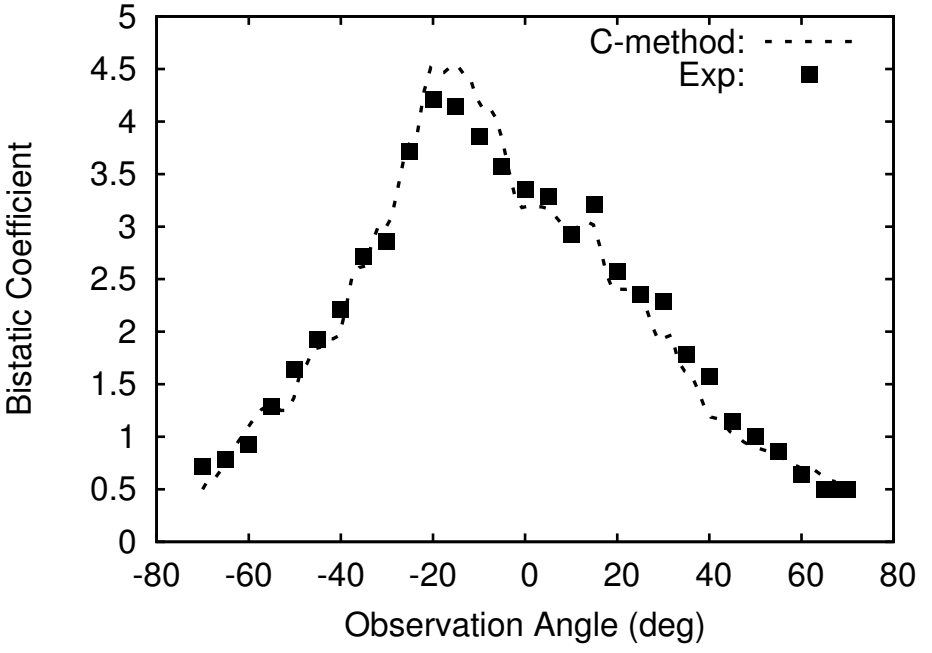

FIGURE 6.14: Average bistatic coefficient versus observation angle in the incidence plane, very rough isotropic surface, polarization(hh)

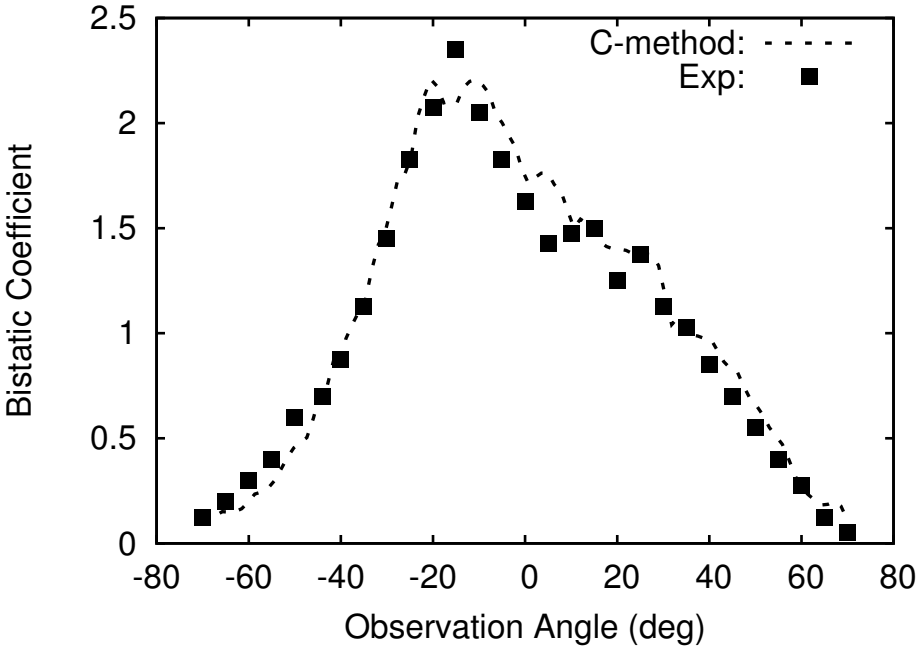

FIGURE 6.15: Average bistatic coefficient versus observation angle in the incidence plane, very rough isotropic surface, polarization(vh)

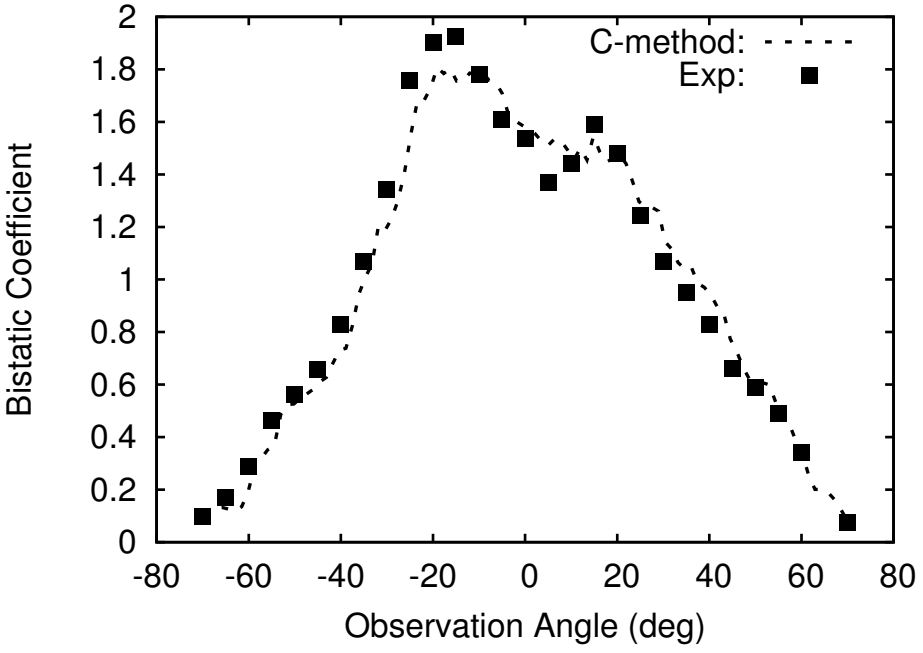

FIGURE 6.16: Average bistatic coefficient versus observation angle in the incidence plane, very rough isotropic surface, polarization(hv)

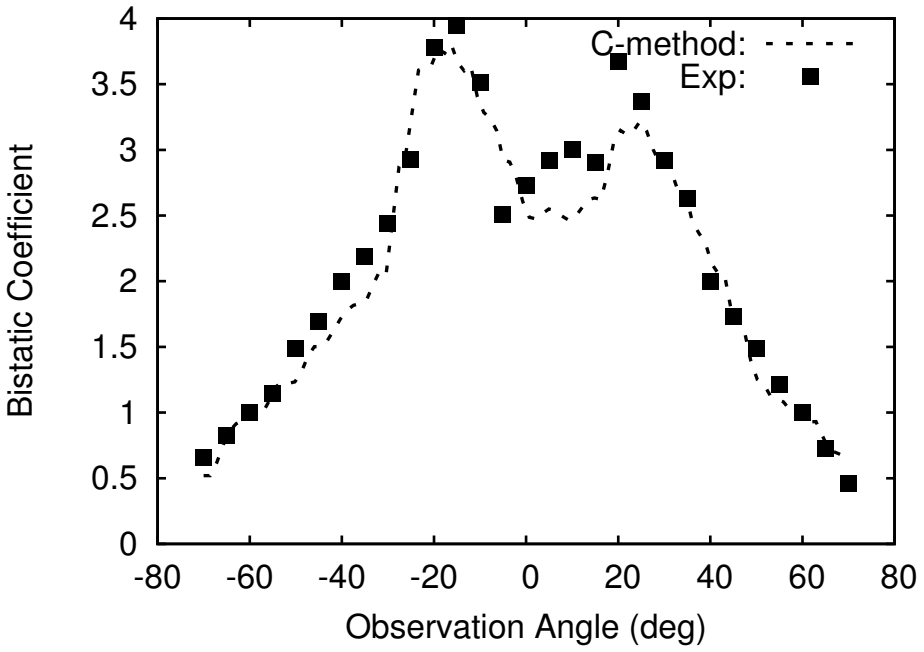

FIGURE 6.17: Average bistatic coefficient versus observation angle in the incidence plane, very rough isotropic surface, polarization(vv)

We also consider an anisotropic surface with an $\sigma = \lambda^{(1)}$, $l_x = 2\lambda^{(1)}, l_y = 4\lambda^{(1)}$. Figure 6.18 gives the co-polarized bistatic coefficient in the incidence plane for a perfectly conducting surface illuminated under $\theta_0 = 20°$ and $\varphi_0 = 90°$. The other simulation parameters are: $D = 8\lambda^{(1)}, M = 28$ and $N_R = 200$. Figure 6.19 gives the co-polarized return when the incidence angles are $\theta_0 = 20°$ and $\varphi_0 = 0°$. Although the elementary cell area is reduced to $8l_x l_y$, the comparison with experimental data which come from [85] is satisfactory. The comparison is also conclusive for other polarizations.

6.5 Conclusion

In this chapter, from numerical experiments, we have observed that some eigenvalues of the scattering matrix can be approximated efficiently by a certain formula. We designed the "early shift" algorithm to take advantage of this property. We plug this "early shift" method, together with Wilkinson's shift and exceptional shift [86], into a new parallel QR algorithm. This new QR algorithm uses multiple chains of tightly coupled bulges chasing technique to parallelize the conventional bulge chasing and the aggressive early deflation technique to detect deflation quickly. We apply this specifically designed parallel QR algorithm to the scattering matrix. We also compare the computation time with that of

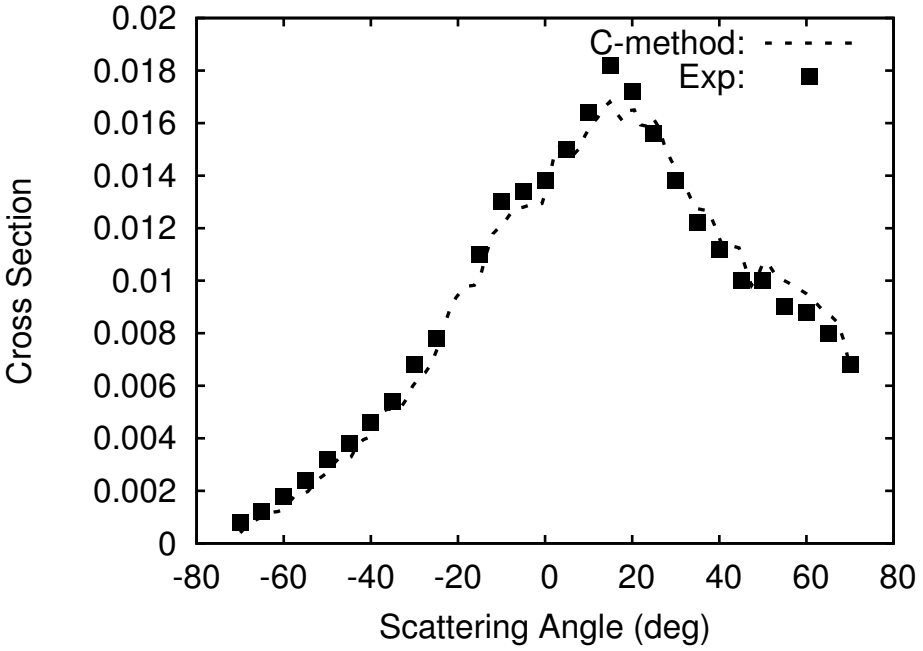

FIGURE 6.18: Average bistatic coefficient versus observation angle in the Oyz incidence plane, very rough anisotropic surface, polarization(hh)

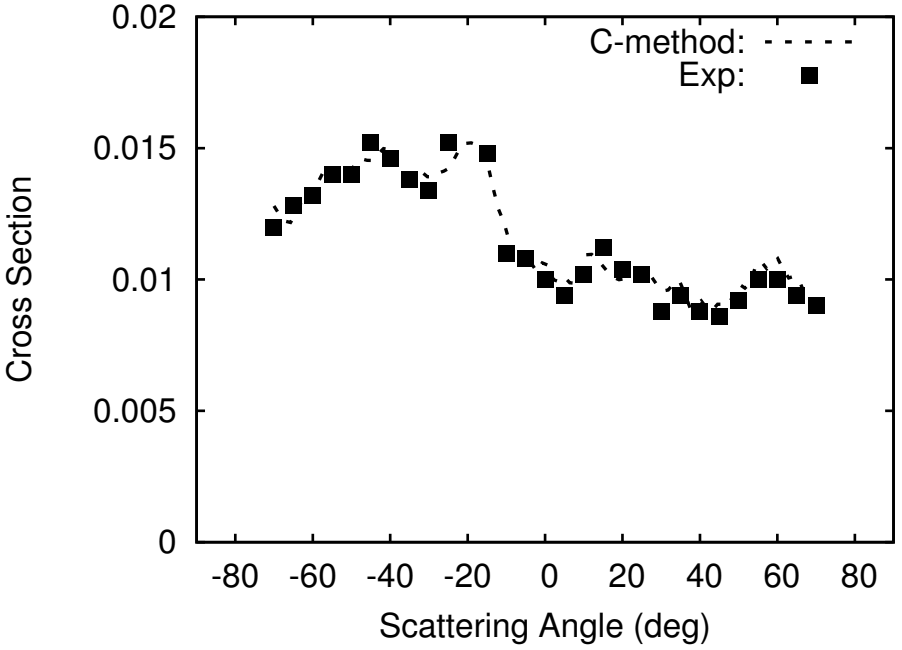

FIGURE 6.19: Average bistatic coefficient versus observation angle in the Oxz incidence plane, very rough anisotropic surface, polarization(hh)

the sequential code. The results show a significant speed up to approximately 40 for 64 cores with our new QR algorithm. This combination of "early shift" and other shifts can also be used in the problems such as linear-quadratic optimal control problem where a large number of eigenvalues and eigenvectors are needed and background of the original problem can provides very good initial approximations.

This parallel QR algorithm can be used for analyzing crossed gratings or two-dimensional random surfaces. Comparisons with experimental data for moderate roughness and isotropic or anisotropic very rough surfaces are conclusive in both co-polarized and cross-polarized components. Comparisons allow the validity of our approach.

In the next chapter, we propose an alternative to the QR algorithm for solving the eigenvalue problem. The proposed method has better scalability than the QR algorithm.

Chapter 7

A proposal: spectral projection method as a global eigensolver

It has been shown that theoretical it is impossible for the standard QR algorithm to be scalable [87]. We want a spectral divide-and-conquer algorithms that can provide us all the eigenvalues and eigenvectors which has a very good scalability. Two related work on spectral divide-and-conquer algorithms are [88] and [89]. In [88], the authors present four versions of divide-and-conquer algorithms and present eigenvalue problem that attain lower bounds, and analyze their convergence and communication costs. Paper [89] shows that all linear algebra operations can also be done stably in $\mathcal{O}(n^{\omega+\eta})$ operations. The authors of [89] consider known divide-and-conquer algorithms for reducing the complexity of matrix inversion to the complexity of matrix multiplication and show that these algorithm can achieve the same forward error bound (bound on the norm of the error in the output) as a conventional backward stable algorithm. We consider a spectral divide-and-conquer algorithm which essentially transform the eigenvalue to linear system problems.

In this chapter, we propose an alternative to QR algorithm for solving the eigenvalue problem. We propose a global eigensolver by a combination of the Sakuria and Sugiura method (SSM) and multiple implicitly restarted Arnolid method with nested subspaces (MIRAMns). The first method allows the computation of interior eigenvalues while the second permits to compute the eigenvalues in the extremities of spectrum. This proposed global eigensolver allows us to calculate all or a large number of the eigenvalues of a

generalized matrix. The MIRAMns [72] is a variant of the IRAM [90] that is based on the projection of the eigenproblem on several nested subspaces instead of a single one. It can therefore use the eigen-information of interest obtained in all subspaces to update the restarting vector. We will take the real matrix as an example, but this proposed global eigensolver can also be applied to a complex matrix in a similar way. Given a matrix $A \in \mathbb{R}^{n\times n}$, MIRAMns uses the Arnoldi method to compute the Ritz elements of A in t nested Krylov subspaces $\mathbb{K}_{m_i,v}$, $1 \leq i \leq t$ with $\mathbb{K}_{m_i,v} \subset \mathbb{K}_{m_{i+1},v}$. MIRAMns selects the "best subspace" by finding which subspace contains the "best" current Ritz elements. We denote the size of the "best" subspace m_{best}. The next step of MIRAMns is to apply the shifted QR procedure to the $m_{best} \times m_{best}$ matrix. By choosing the undesired eigenvalues as shifts, the information related to the desired eigenvalues are concentrated in the leading submatrix. MIRAMns then completes Arnoldi projection of t nested Krylov subspaces by restarting with this submatrix whose size is the number of wanted eigenvalues. The MIRAMns can be used to provide the extremes of the spectrum of A with good convergence properties.

The rest of the spectrum of A is thus located in a finite domain $\mathcal{D}$ described by the extremes provided by MIRAMns. The domain $\mathcal{D}$ is then divided into several sub-domains. For each sub-domain, the contour integral based projection method projects the matrix pencil $(A - I)$ onto the subspace associated with eigenvalues that are located in the sub-domain via numerical integration. A moment-based approach can be used to find the eigenvalues in each sub-domain independently. To avoid the numerically unstable problem of the computation using explicit moments, we often use a Rayleigh-Ritz procedure instead. For the computation of the contour integral, we solve a certain number of linear systems derived from the matrices A and I. When A is large, the computational costs for solving linear systems are dominant.

7.1 Algorithms

This spectral projection method as a global eigensolver is a combination of MIRAMns and SS method. We present these methods and how they can be combined to form a global solver in this section.

7.1.1 MIRAMns

The MIRAMns is a variant of the implicitly restarted Arnoldi method (IRAM). Recall that IRAM allows us to compute a few eigenvalues in the extremes of the spectrum of a large matrix. For that, IRAM combines the implicitly shifted QR algorithm with a k-step Arnoldi factorization to obtain a truncated form of implicitly shifted QR iteration. This approach offers a more efficient and numerically stable formulation than explicitly restarted Arnoldi method (ERAM) [91]. By using IRAM instead of ERAM, the numerical difficulties and storage problems normally associated with the Arnoldi process are avoided. The algorithm is capable of computing a few (k) eigenvalues with user specified features such as largest real part, largest magnitude, smallest real part or smallest magnitude.

We start from the Arnoldi factorization of length $m = p + k$,

$$AV_m = V_m H_m + f_m e_m^T \tag{7.1}$$

where e_m^T means the transpose of the vector e_m. We apply p shifts $\mu_1, ..., \mu_p$ implicitly

$$AV_m^+ = V_m^+ H_m^+ + f_m e_m^T Q \tag{7.2}$$

where $Q = Q_1 Q_2 \cdots Q_p$ the product of the orthogonal matrices related to $\mu_1, ..., \mu_p$ and $V_m^+ = V_m Q, H_m^+ = Q^T H_m Q$. From the fact that Q is the product of p (unitary) Hessenberg matrices, it is easy to see that Q has p non-zero off-diagonals below its main diagonal. So the first $k-1$ elements of the vector $e_m^T Q$ are zeros. So if we discard the last p columns of equation (7.2), we will have

$$\begin{aligned} AV_m^+(:, 1:k) &= V_m^+(:, 1:k+1) H_m^+(:, 1:k) + f_m e_m^T Q(:, 1:k) \\ &= V_m^+(:, 1:k) H_m^+(:, 1:k) + h_{k+1,k}^+ v_{k+1}^+ e_k^T + q_{m,k}^+ f_m e_k^T \\ &= V_m^+(:, 1:k) H_m^+(:, 1:k) + (v_{k+1}^+ \hat{\beta}_k + f_m \sigma_k) e_k^T \end{aligned} \tag{7.3}$$

where we denote $\hat{\beta}_k = h_{k+1,k}^+, \sigma_k = q_{m,k}^+$. Equation (7.3) can be also written as

$$AV_k^+ = V_k^+ H_k^+ + f_k^+ e_k^T \tag{7.4}$$

A description of this IRAM algorithm can be found in Algorithm 3.

Algorithm 3 The implicitly restarted Arnoldi process

Input: (A, V, k, m) with $AV_m = V_m H_m + f_m e_m^T$, an m-step Arnoldi factorization, with $m = p + k$

Output: k eigenvalues with user specified features and their corresponding eigenvectors.

1: **for** $l = 1, 2, 3, ...$ until convergence **do**
2: Compute the spectrum of H_m: $\sigma(H_m)$, if convergence, stop. Otherwise, select set of p shifts $\mu_1, \mu_2, ..., \mu_p$;
3: $q^T = e_m^T$;
4: **for** $j = 1, 2, ..., p$ **do**
5: Factor $[Q_j, R_j] = qr(H_m - \mu_j I)$;
6: $H_m = Q_j^T H_m Q_j$, $V_m = V_m Q_j$, $q^T = q^T Q_j$;
7: **end for**
8: $f_k = v_{k+1} H_m(k+1, k) + f_m q^T(k)$; $V_k = V_{m(1:n,1:k)}$; $H_k = H_{m(1:k,1:k)}$;
9: Begining with the k-step Arnoldi factorization, $AV_k = V_k H_k + f_k e_k^T$, apply p additional steps of the Arnoldi process to obtain a new m-step Arnoldi factorization, $AV_m = V_m H_m + f_m e_m^T$;
10: **end for**

The MIRAMns takes advantage of IRAM by choosing a set of initial Krylov subspaces that differ only by their sizes. MIRAMns chooses a set of t different subspace sizes $M = (m_1, ..., m_t)$ with a strict order $m_1 < ... < m_t$. MIRAMns then performs t Arnoldi projections on the subspaces $\mathbb{K}_{m_i,v}$, for $1 \leq i \leq t$, with $\mathbb{K}_{m_1,v} \subset \mathbb{K}_{m_t,v}$ and initial vector v. MIRAMns then chooses the subspace size m_{best} by finding the Arnoldi factorization which offers the best Ritz estimation for k desired eigenpairs.

$$AV_{m_{best}} = V_{m_{best}} H_{m_{best}} + f_{m_{best}} e_{m_{best}}^T \tag{7.5}$$

From this k-step Arnoldi factorization, $p_i = m_i - k, 1 \leq i \leq t$ additional steps of Arnoldi factorizations are applied to obtain t new projections onto the updated subspaces. This procedure can go on until convergence.

A description of this MIRAMns algorithm is presented in algorithm 4.

In order to select the best results in the above algorithm, we consider that $(V_{m_i}, H_{m_i}, f_{m_i})$ is "better" than $(V_{m_j}, H_{m_j}, f_{m_j})$ if $r_k^{m_i} < r_k^{m_j}$ where $r_k^m = \max(\rho_{1,m}, ..., \rho_{k,m})$ is defined by Ritz estimates $\rho_{i,m} = |\beta_m e_m^T y_i^{(m)}|$.

One advantage of the MIRAMns is that it overcomes the problem of sensitivity of convergence with respect to small perturbation of the subspace size that occurs in the normal restarted Arnoldi methods. It achieves this by choosing the "best" size among

Algorithm 4 Multiple IRAM with nested subspaces

Input: $(A, V_{m_i}, H_{m_i}, f_{m_i})$ with $AV_{m_i} = V_{m_i}H_{m_i} + f_{m_i}e_{m_i}^T$, an m_i-step Arnoldi factorization, with $m_i = p_i + k$, where $1 \leq i \leq t$.
Output: k eigenvalues with user specified features and their corresponding eigenvectors.

1: **for** $l = 1, 2, 3, ...$ until convergence **do**
2: Compute the spectrum of H_{m_i}: $\sigma(H_{m_i})$, if convergence, stop. Otherwise compute their associated eigenvectors and residuals for $1 \leq i \leq t$.
3: Select the best results in these subspaces and the associated best subspace size m_{best}. Set $m = m_{best}, H_m = H_{m_{best}}, V_m = V_{m_{best}}, f_m = f_{m_{best}}$.
4: Select a set of $p = m - k$ shifts $(\mu_1^{(m)}, ..., \mu_p^{(m)})$ based on $\sigma(H_m)$ or other information.
5: $q^T = e_m^T$;
6: **for** $j = 1, 2, ..., p$ **do**
7: Factor $[Q_j, R_j] = qr(H_m - \mu_j^{(m)} I)$;
8: $H_m = Q_j^H H_m Q_j$, $V_m = V_m Q_j$, $q = q^H Q_j$;
9: **end for**
10: $f_k = v_{k+1}\hat{\beta}_k + f_m\sigma_k$; $V_k = V_{m(1:n,1:k)}$; $H_k = H_{m(1:k,1:k)}$;
11: Begining with the k-step Arnoldi factorization, $AV_k = V_k H_k + f_k e_k^T$, apply $p_i = m_i - k$ additional steps of the Arnoldi process to obtain t new m_i-step Arnoldi factorization, $AV_{m_i} = V_{m_i}H_{m_i} + f_{m_i}e_{m_i}^T$;
12: **end for**

the different sizes. Another advantage of MIRAMns is that it has better property of convergence with almost the same time complexity compared with IRAM.

7.1.2 SS method

The SS method was introduced in [71], [92]. Given a finite domain $\mathcal{D}$, we want to calculate the eigenvalues of A that lie in it. Suppose that we cover the domain $\mathcal{D}$ with s subdomains $\mathcal{D}_i, 1 \leq i \leq s$, such that $\mathcal{D} = \cup_i \bar{\mathcal{D}}_i$. Now we only need to calculate the eigenvalues that lie in each sub-domain $\mathcal{D}_i, 1 \leq i \leq s$ and these tasks can be performed in parallel.

For each subdomain $\mathcal{D}_i, 1 \leq i \leq s$, define

$$f(z) = u^H(zI - A)^{-1}v \tag{7.6}$$

with non-zero vectors $u, v \in \mathbb{R}^n$. Define

$$\mu_k = \frac{1}{2\pi i}\int_{\partial\mathcal{D}_i}(z - z_0)^k f(z)dz, k = 0, 1, ... \tag{7.7}$$

where z_0 is located inside $\mathcal{D}_i$. Suppose there are m eigenvalues lie in $\mathcal{D}_i$, then these eigenvalues are exactly the same as the eigenvalues of pencil $H_m^< - \lambda H_m$, with the $m \times m$ Hankel matrices $H_m = [\mu_{i+j-2}]$ and $H_m^< = [\mu_{i+j-1}]$, $1 \leq i, j \leq m$. See [71] for a proof, here u, v are any non-zero vectors.

The number m can be calculated from the following formula [93]:

$$m = \frac{1}{2\pi i} \int_{\partial \mathcal{D}_i} tr(F(z)^{-1}) dz \tag{7.8}$$

For the eigenvectors, if we define

$$s_k = \frac{1}{2\pi i} \int_{\partial \mathcal{D}_i} (z - z_0)^k (zI - A)^{-1} v dz, k = 0, 1, ... \tag{7.9}$$

and V_m is the Vandermonde matrix

$$V_m = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 - z_0 & \lambda_2 - z_0 & \cdots & \lambda_m - z_0 \\ \vdots & \vdots & & \vdots \\ (\lambda_1 - z_0)^{m-1} & (\lambda_2 - z_0)^{m-1} & \cdots & (\lambda_m - z_0)^{m-1} \end{pmatrix} \tag{7.10}$$

then the associated eigenvectors are given by the formula:

$$[q_1, ..., q_m] = [s_0, ..., s_{m-1}] V_m^{-T} \tag{7.11}$$

Suppose that $\mathcal{D}_i$ is a circle centered at point γ with radius ρ, a description of the idea of contour integral based projection (explicit moments) can be found in algorithm 5.

Algorithm 5 The contour integral based projection method (explicit moments)

Input: $(u, v \in \mathbb{R}^n, N, m, \gamma, \rho)$

Output: approximated eigenvalues of A that lie in the finite domain $\mathcal{D}_i$: $\hat{\lambda}_1, ..., \hat{\lambda}_m$ and their associated eigenvectors: $\hat{q}_1, ..., \hat{q}_m$

1: Set $\omega_j = \gamma + \rho exp(2\pi\sqrt{-1}j/N)$, $j = 0, ..., N-1$;
2: Form $y_j = (\omega_j I - A)^{-1} v$, $j = 0, ..., N-1$;
3: Set $f_j = u^H y_j, j = 0, ..., N-1$;
4: Compute $\hat{\mu}_k = \frac{1}{N} \sum_{j=0}^{N-1} (\omega_j - \gamma)^{k+1} f(\omega_j), k = 0, ..., 2m-1$;
5: Compute $\hat{s}_k = \frac{1}{N} \sum_{j=0}^{N-1} (\omega_j - \gamma)^{k+1} y_j, k = 0, ..., m-1$
6: Compute the eigenvalues $\zeta_1, ..., \zeta_m$ of the pencil $H_m^< - \lambda H_m$;
7: Compute $\hat{q}_1, ..., \hat{q}_m$ given by $[\hat{q}_1, ..., \hat{q}_m] = [\hat{s}_0, ..., \hat{s}_{m-1}] \hat{V}_m^{-T}$;
8: Set $\hat{\lambda}_j = \gamma + \zeta_j, j = 1, ..., m$.

IF some eigenvalues are very close to each in the contour, the Hankel matrices $H_m^<$ and H_m are very ill-conditioned. The Rayleigh-Ritz method can be used to avoid the explicit use of moments and improve numerical accuracy. We apply a Rayleigh-Ritz procedure by projecting the matrix A to $\tilde{A} = \Pi^T A \Pi$ with an unitary basis $\Pi \in \mathbb{C}^{n \times m}$. The eigenvalues of A can be approximated by the Ritz values of the projected pencil $(\tilde{A}, I)$. In practice, the numerical value of m from equation (7.8) is not always an integer, so it is more convenient and more efficient to choose a number $M(\geq m)$ as the size of Hankel matrices . This choice can decrease the influence of the quadrature error suffered from eigenvalues located outside the boundary.

A description of the contour integral based projection (Rayleigh-Ritz) can be found in algorithm 6.

Algorithm 6 The contour integral based projection method (Rayleigh-Ritz)

Input: $(v \in \mathbb{R}^n, N, M, \gamma, \rho)$
Output: approximated eigenvalues of A that lie in the finite domain $\mathcal{D}_i$: $\hat{\lambda}_1, ..., \hat{\lambda}_m$ and their associated eigenvectors: $\hat{x}_1, ..., \hat{x}_m$

1: Set $\omega_j = \gamma + \rho exp(2\pi\sqrt{-1}(j + 1/2)/N)$, $j = 0, ..., N - 1$;
2: Solve $(\omega_j I - A)y_j = v$, for $y_j, j = 0, ..., N - 1$;
3: Compute $\hat{s}_k = \frac{1}{N}\sum_{j=0}^{N-1}(\omega_j - \gamma)^{k+1} y_j, k = 0, ..., M - 1$;
4: Compute construct an unitary basis Π from $(\hat{s}_0, ..., \hat{s}_{M-1})$;
5: Form $\tilde{A} = \Pi^T A \Pi$;
6: Compute eigenpairs (θ_j, w_j) with $j = 1, ..., M$ of $(\tilde{A}, I)$;
7: Set $p_j = \Pi w_j, j = 1, ..., M$;
8: Select the approximated eigenpairs $(\hat{\lambda}_1, \hat{x}_1), ..., (\hat{\lambda}_m, \hat{x}_m)$ from $(\theta_j, p_j), j = 1, ..., M$;

7.1.3 A global eigensolver

For the given matrix A, we first apply the MIRAMns algorithm to calculate k_s eigenvalues that have the smallest magnitude and k_l eigenvalues that have the largest magnitude. Thus we get the extremes of the spectrum of A: $|\lambda_1| \leq ... \leq |\lambda_{k_s}| \leq |\lambda_{n+1-k_l}| \leq ... \leq |\lambda_n|$. Denote $R = |\lambda_{n+1-k_l}|, r = |\lambda_{k_s}|$, then the rest of the spectrum lies in the finite domain $\mathcal{D} = B(0, R) \setminus B(0, r)$, where $B(x, y)$ represents the open ball centered at x with radius y in the complex plane. Cover the finite domain $\mathcal{D}$ by s subdomains $\mathcal{D}_i, 1 \leq i \leq s$ such that $\mathcal{D} = \cup_i \bar{\mathcal{D}}_i$. Use the contour integral based projection method to calculate the eigenvalues in each sub-domain $\mathcal{D}_i$ to get all the eigenvalues in $\mathcal{D}$. Together with the extremes of the spectrum, we get all the eigenvalues of A.

Assuming that the eigenvalues of A are arranged as $|\lambda_1| \leq |\lambda_2| \leq ...|\lambda_{n-1}| \leq |\lambda_n|$, the algorithm can be described as in Algorithm 7:

Algorithm 7 Spectral projection method as global eigensolver

Input: A, k_s, k_l, N

Output: approximated eigenvalues of A and their associated eigenvectors.

1: Apply MIRAMns to calculate $\lambda_1, ..., \lambda_{k_s}$, $\lambda_{n+1-k_l}, ..., \lambda_n$ and their associated eigenvectors;
2: Divide the domain $\mathcal{D} = B(0, R) \setminus B(0, r)$ by $\mathcal{D} = \cup_i \bar{\mathcal{D}}_i$;
3: Apply the SS method to calculate the eigenvalues that lie in $\mathcal{D}_i$ in parallel and their corresponding eigenvectors;
4: Collect the information from each domain $\mathcal{D}_i$ and obtain all the eigenvalues and eigenvectors;

7.1.4 Parallelism analysis

The proposed method has a good parallel nature. For the MIRAMns, the computation in different subspace of multiple implicitly restarted Arnoldi method (MIRAM) can be performed in parallel. In addition to this coarse grain parallelism, the communication of the eigen-information of interest among processes can also be made asynchronously. For the SS method, the solving of each sub-problem associated with the sub-domain can be done independently. Besides, for each sub-problem, one needs to use the trapezoidal rule to calculate the contour integral. The calculation related to each quadrature points can be done in parallel. Finally, for each quadrature point, we need to solve the associated linear system, this can also be done using parallel linear solvers.

7.2 Numerical experiments

We test the matrix A_n with $n = 1000$, $A_n(i, i) = 2, A_n(i, i+1) = A_n(i+1, i) = 1$, the other entries are 0.

$$A_n = \begin{pmatrix} 2 & 1 & & & & & & \\ 1 & 2 & 1 & & & & & \\ & 1 & 2 & 1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1 & 2 & 1 & & \\ & & & & 1 & 2 & 1 & \\ & & & & & 1 & 2 & \end{pmatrix} \tag{7.12}$$

The eigenvalues of A_n are proved to be

$$\lambda_k(A_n) = 2 + 2\cos(\frac{k\pi}{n+1}), k = 1, 2, ..., n \tag{7.13}$$

We choose this matrix because the eigenvalues are known and all of them are real, so we can check our algorithm easily with this matrix.

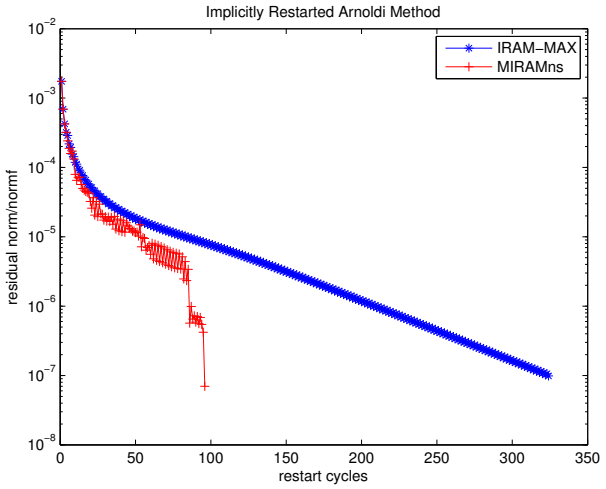

FIGURE 7.1: $MIRAMns(9, 12, 16)$ to calculate 3 eigenvalues with largest magnitude, the result is compared with $IRAM(16)$

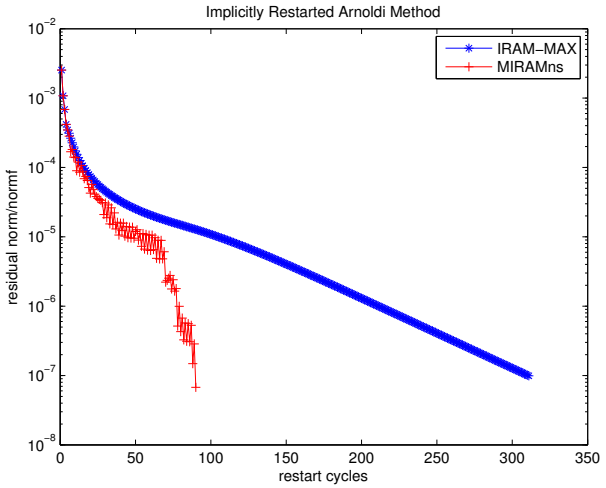

FIGURE 7.2: $MIRAMns(9, 12, 16)$ to calculate 3 eigenvalues with smallest magnitude, the result is compared with $IRAM(16)$

Figure 7.1 and figure 7.2 gives the results of MIRAMns. The subspaces' size are chosen as $(9, 12, 16)$ and we denote this as $MIRAMns(9, 12, 16)$. We observed that MIRAMns converges with fewer iterations than IRAM. For the testing matrix A_n, the 3 eigenvalues

with largest magnitude calculated from $MIRAMns(9, 12, 16)$ are:

$$\begin{pmatrix} 3.999990150111810 \\ 3.999911351284587 \\ 3.999753743363125 \end{pmatrix}$$

The results have an average precision (absolute error computed using Ritz estimate [72]) of 6.993000×10^{-8}. The 3 eigenvalues with smallest magnitude calculated from $MIRAMns(9, 12, 16)$ are:

$$\begin{pmatrix} 3.939947166722805 \times 10^{-5} \\ 1.575965946128510 \times 10^{-4} \\ 3.545973886132522 \times 10^{-4} \end{pmatrix}$$

The results have an average precision of 6.757897×10^{-8}.

So we know now that the rest of the spectrum of matrix A lies in the interval $(3.545973886132522 \times 10^{-4}, 3.999753743363125)$. To have certain tolerance of error, we will search the eigenvalues in the interval $[0, 4]$.

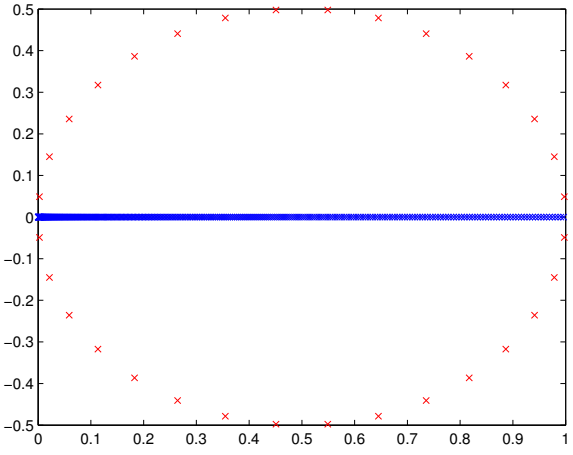

FIGURE 7.3: SS-RR method with $Center = 0.5, Radius = 0.5$, 333 eigenvalues are calculated

Figure 7.3 to figure 7.6 are the results of SS-RR method applied to the interval $[0, 4]$. The horizontal line represents the x-axis and the vertical line represents the y-axis. The red points represent the quadrature points and the blue points represent the eigenvalues that are calculated in the domain described by the red points. The total number of

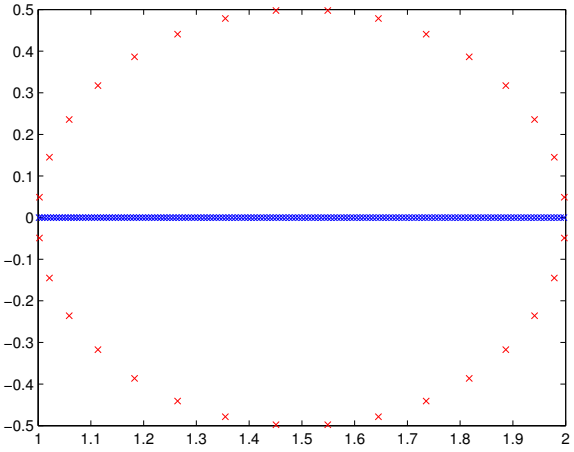

FIGURE 7.4: SS-RR method with $Center = 1.5, Radius = 0.5$, 167 eigenvalues are calculated

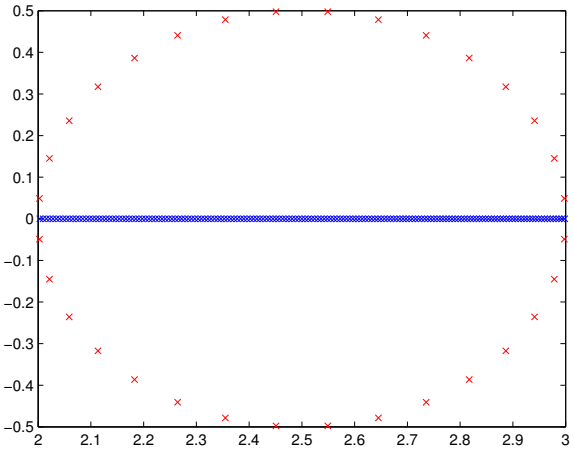

FIGURE 7.5: SS-RR method with $Center = 2.5, Radius = 0.5$, 167 eigenvalues are calculated

quadrature points is 32 for each circle, $N = 32$. We get all 1000 eigenvalues correctly. The average residual is 1.59×10^{-5}. Numerical experiments of the parallel version of this global eigensolver are also performed for the test matrix A_n. Figure 7.7 shows the scalability for the matrix A_n when we increase the number of cores. Here we use the MPI techniques to parallelize the algorithm. Because of the independence of the computation related to contour paths and quadrature points, we have an almost linear speed up. Figure 7.7 shows a good scalability of the proposed global eigensolver. The computation time is quite long for our method. A sequential MATLAB function eig() return the value in 0.020527 second, which is more than 100 faster than our method with 20 cores. However, the idea is that our method may still work when some existing

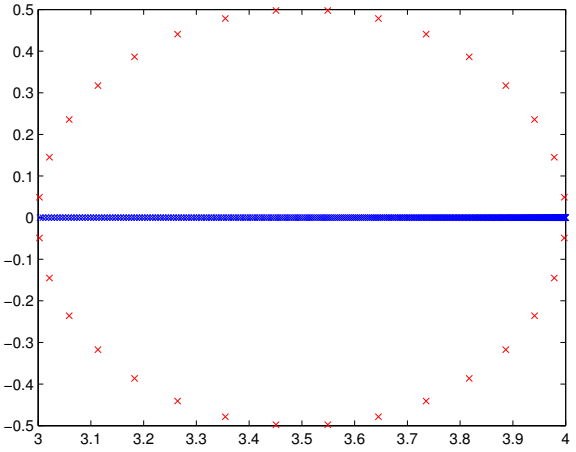

FIGURE 7.6: SS-RR method with $Center = 3.5, Radius = 0.5$, 333 eigenvalues are calculated

algorithms fail.

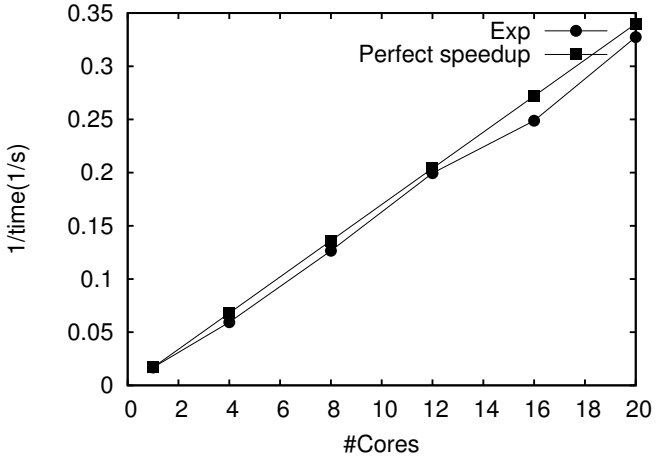

FIGURE 7.7: Scalability for test matrix A

We next test the matrix B_n with $n = 1000$

$$B_n = \begin{pmatrix} 5 & -4 & 1 & & & & \\ -4 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 6 & -4 \\ & & & & 1 & -4 & 5 \end{pmatrix} \tag{7.14}$$

This matrix is just the square of the previous test matrix, with diagonal similarity by $diag(1, -1, 1, -1, ...)$, so it should behave the same. Also both matrices are symmetric, so the conditioning is as good as possible. The exact eigenvalues of this matrix are given by

$$\lambda_k(B_n) = 16\cos^4(k\pi/(2n+2)), k = 1, ..., n \tag{7.15}$$

Figure 7.8 and figure 7.9 show the results by MIRAMns with the subspace size equals $9, 12$ and 16. The three eigenvalues with largest magnitude calculated are

$$\begin{pmatrix} 15.999921200819772 \\ 15.999684803750737 \\ 15.999290795479558 \end{pmatrix}$$

The results have an average precision of 7.991445×10^{-8}. The three eigenvalues with smallest magnitude calculated are

$$\begin{pmatrix} 4.473114840983656 \times 10^{-8} \\ 1.762265816198667 \times 10^{-5} \\ 6.482147601621269 \times 10^{-5} \end{pmatrix}$$

The results have an average precision of 5.393488×10^{-7}. We are going to search eigenvalues in the interval $[0, 16]$.

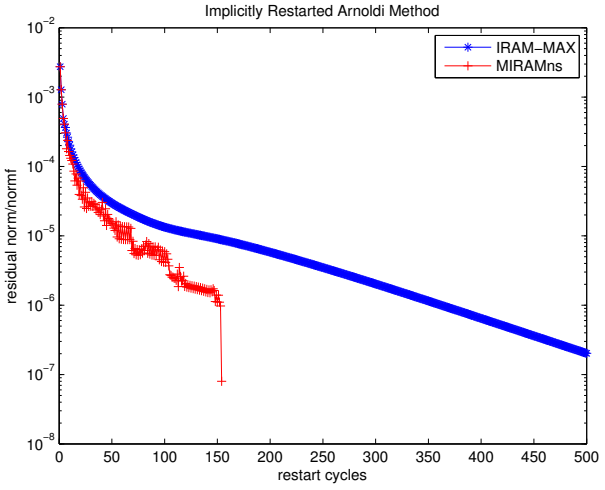

FIGURE 7.8: $MIRAMns(9, 12, 16)$ to calculate 3 eigenvalues with largest magnitude, the result is compared with $IRAM(16)$

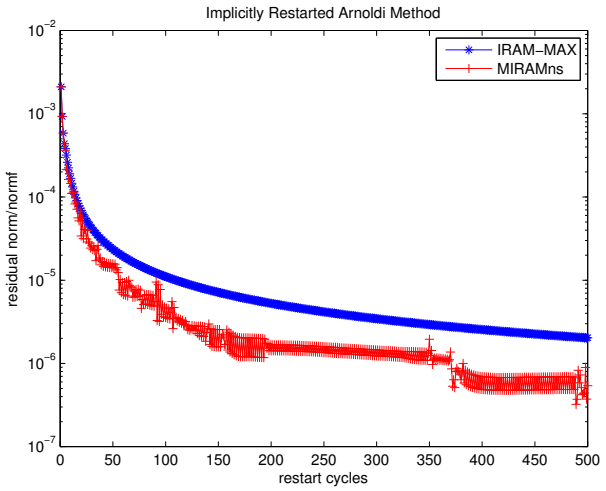

FIGURE 7.9: $MIRAMns(9, 12, 16)$ to calculate 3 eigenvalues with smallest magnitude, the result is compared with $IRAM(16)$

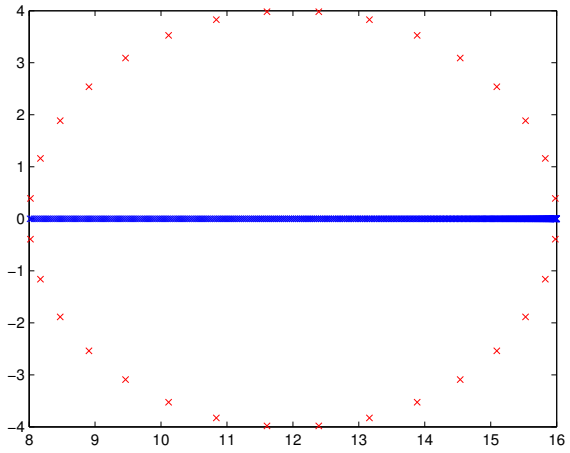

FIGURE 7.10: SS-RR method with $Center = 12, Radius = 4$, 364 eigenvalues are calculated, average residual is $3.53698402454027090 \times 10^{-6}$

Figure 7.10 to figure 7.19 show how the interval $[0, 16]$ is divided and 1000 eigenvalues are found. It can be noticed that when some of the eigenvalues are very close, the precision will decrease. To increase the precision, one can decrease the radius of the circle, thus, there will be a demand of more circles to be addressed. Similar to the test matrix A_n, we also run the parallel version of this global eigensolver for the test matrix B_n. Figure 7.20 shows the scalability for the matrix B_n wiht $n = 1000$ when we increase the number of cores. We can observed that the line is not as "straight" as that for the matrix A_n. The slope of the approximated line is smaller than that for the matrix A_n. This decrease of performance can be explained if one observes the location of the spectra of the two matrices. For matrix B_n, there are some eigenvalues

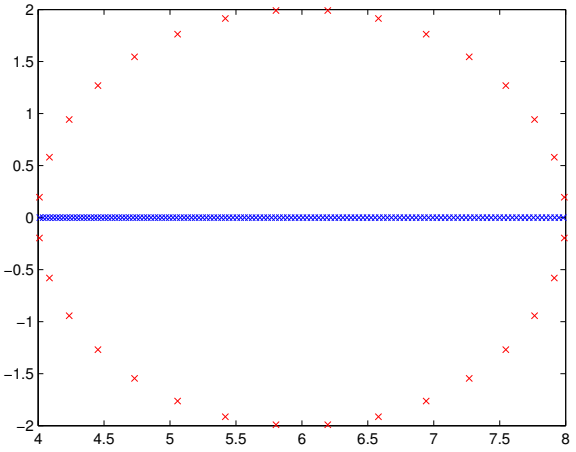

FIGURE 7.11: SS-RR method with $Center = 6, Radius = 2$, 136 eigenvalues are calculated, average residual is $5.78977861683748897 \times 10^{-6}$

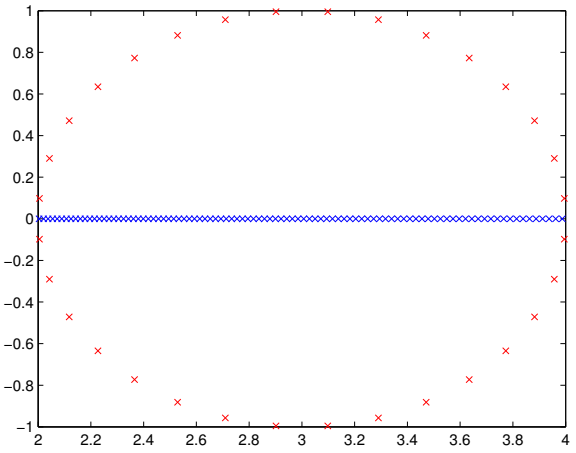

FIGURE 7.12: SS-RR method with $Center = 3, Radius = 1$, 95 eigenvalues are calculated, average residual is $2.18255949228685363 \times 10^{-4}$

that are very close to the point zero. We can say that there is a clustered point. In fact, the smallest theoretical eigenvalue is $16\cos^4(1000\pi/2002) = 9.7020 \times 10^{-11}$ and $|16\cos^4(1000\pi/2002) - 16\cos^4(999\pi/2002)| = 1.4553 \times 10^{-9}$. These eigenvalues are so close that it is difficult to tell them apart from each other. Thus a very small sub-domain is required to keep a good precision. The initial covering strategy around this clustered point will fail which leads to a poorer performance than that for the matrix A_n. Even in this case, figure 7.20 shows a good scalability of the proposed global eigensolver.

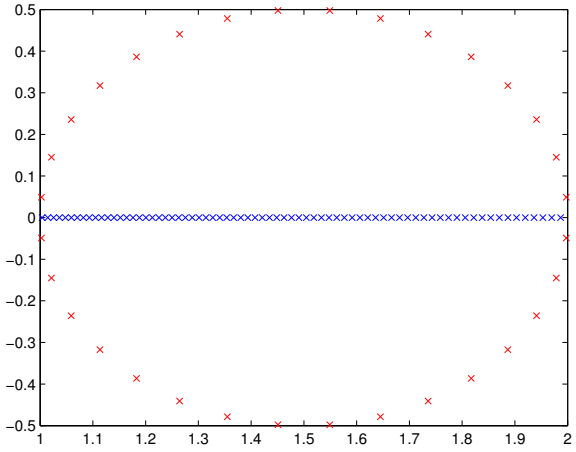

FIGURE 7.13: SS-RR method with $Center = 1.5, Radius = 0.5$, 72 eigenvalues are calculated, average residual is $2.59612185434167264 \times 10^{-6}$

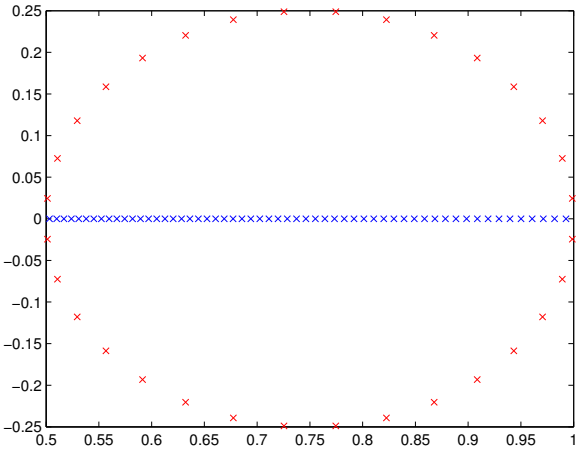

FIGURE 7.14: SS-RR method with $Center = 0.75, Radius = 0.25$, 57 eigenvalues are calculated, average residual is $1.45660089462261519 \times 10^{-6}$

7.3 Conclusion

In this chapter, we propose a global eigensolver by combination of the contour integral based projection method (SS method) and the multiple implicitly restarted Arnoldi method with nested subspaces (MIRAMns). This proposed global eigensolver allows us to calculate a large number of (or all) eigenvalues and eigenvectors of a generalized matrix.

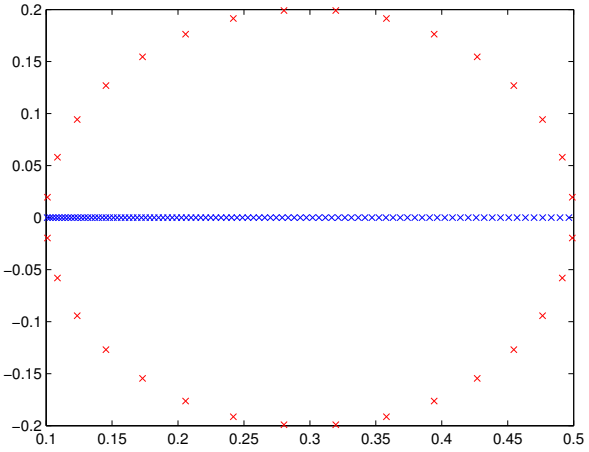

FIGURE 7.15: SS-RR method with $Center = 0.3, Radius = 0.2$, 95 eigenvalues are calculated, average residual is $2.59080105213133636 \times 10^{-4}$

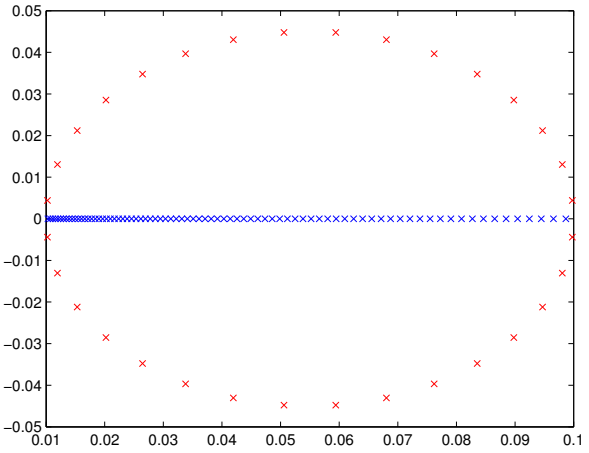

FIGURE 7.16: SS-RR method with $Center = 0.055, Radius = 0.045$, 80 eigenvalues are calculated, average residual is $1.40601714058967572 \times 10^{-5}$

This is the first attempt to combine MIRAMns and SS method to form a global eigensolver. Numerical experiments show this combination allows us to get all the eigenvalues and their corresponding eigenvectors. MIRAMns converges with less iterations than IRAM, and the SS method is very suitable for parallelization. The scalability of the global eigensolver is very good, we get almost linear speed up. The complexity of computation can be varying with the precision that is required. The precision can be increased with smaller sub-domain.

More precise error analysis of this method and the optimal division strategy, especially when the whole domain $\mathcal{D}$ is of two dimension, is a part of our future work. We may

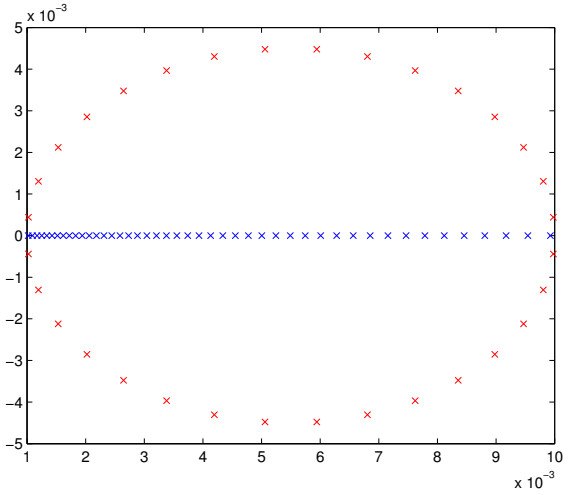

FIGURE 7.17: SS-RR method with $Center = 0.0055, Radius = 0.0045$, 45 eigenvalues are calculated, average residual is $2.61507800766467365 \times 10^{-6}$

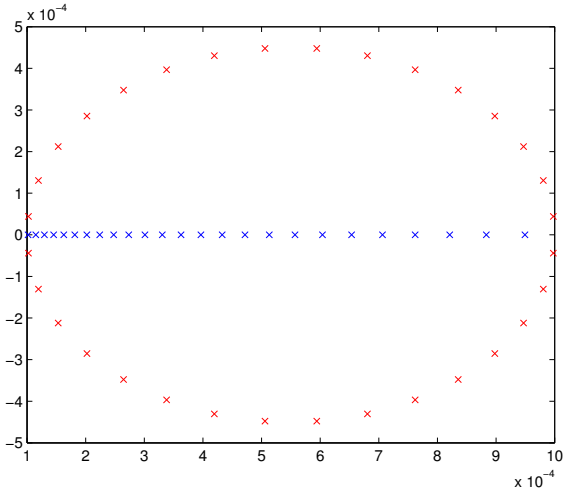

FIGURE 7.18: SS-RR method with $Center = 0.00055, Radius = 0.00045$, 25 eigenvalues are calculated, average residual is $4.74249918677096298 \times 10^{-10}$

want to extend this method to apply it to the C-method in the future.

In the next chapter, we will propose a new approach of the curvilinear coordinate method where we do not use the translation coordinate system. The proposed method allows analyzing the complex phenomenon of incident energy absorption. The new version of the C-method could be an attractive alternative to analyze multilayered grating having parallel or non-parallel interfaces.

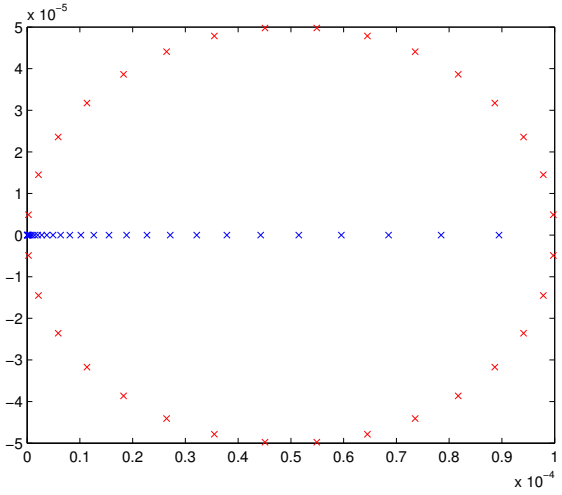

FIGURE 7.19: SS-RR method with $Center = 0.00005, Radius = 0.00005$, 31 eigenvalues are calculated, average residual is $4.34840427877808130 \times 10^{-6}$

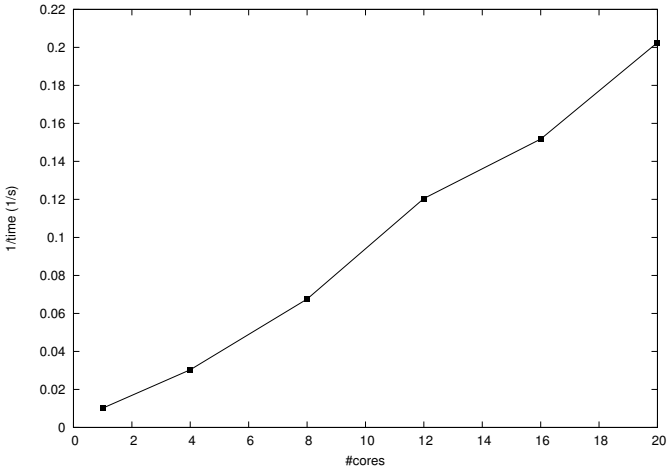

FIGURE 7.20: Scalability for test matrix B

Chapter 8

The C-method as an initial value problem

The C-method in the previous chapter is not very efficient when we are dealing with multi-layer gratings. We want to find other solutions. Especially, we want to explore the potential parallelization of multi-layer gratings. That is we want to find a way to deal with each layer independently and then combine them. In this chapter we propose a new version of the C-method.

8.1 Eigenvalue problem and initial value problem

We will propose a new approach of the curvilinear coordinate method where we don't use the translation coordinate system. We consider two horizontal plane above and below the grating. We define a coordinate system that the grating surface and both horizontal planes correspond to surface coordinate. Similar coordinate systems have been defined for analyzing discontinuities in rectangular waveguides [51–54] and radiation loss of optical waveguides [94]. Inside the area A delimited by the two horizontal planes, in the air and the low medium, the covariant formalism of Maxwell's equations lead to a differential equation system with non-constant coefficients. This system represents an initial value problem. The curvilinear coordinate method expressed in the translation coordinate system leads to an eigenvalue problem. It is the fundamental difference with this new approach. The scattering matrix ($\boldsymbol{S}-$matrix) relates the amplitudes of

outgoing planes waves to those of incoming waves. We show how to determine the $\boldsymbol{S}-$matrix by solving the initial value problem, by satisfying the boundary conditions on the grating interface and using the continuity relations on the two horizontal planes between covariant components of fields and Cartesian ones.

8.2 From Maxwell's equations in covariant form to an initial value problem

For simplicity, we consider only the one-dimensional case. In the Cartesian referential $Oxyz$, the grating is represented by a periodic cylindrical surface $y = a(x)$ (figure 8.1). This surface separates the air (medium 1) from the medium with a real or complex refractive index (medium 2). The grating of period D is illuminated by a monochromatic plane wave under the incidence θ_0. The incident wave vector lies in the xOy plane. For $E_{//}$ polarization, the electric vector is parallel to the grooves. For $H_{//}$ polarization, it is the case of magnetic vector. The letter m denotes indifferently the upper medium ($m = 1$) or the lower medium ($m = 2$). Henceforth, $\nu^{(m)}, Z^{(m)}$ and $k^{(m)}$ indicate the optical index. the inpendance and the wave number of medium (m).

As show by 8.1, the space is divided into four regions, Within the regions $y \geq y_1$ and $y \leq y_2$, we consider the Cartesian coordinates (x, y, z). Outside the grooves, i.e. when $y > max(a(x))$ and $y < min(a(x))$, the diffracted field can be represented by a combination of elementary plane waves, the Rayleigh expansion (3.7).

Within the regions A_1 and A_2 defined by $a(x) \leq y \leq y_1$ and $y_2 \leq y \leq a(x)$, we consider the non-orthogonal coordinate system defined as follows:

$$\begin{cases} x' = x \\ u = y_m \frac{y - a(x)}{y_m - a(x)} \\ z' = z \end{cases} \tag{8.1}$$

The grating surface $y = a_(x)$ coincides with the coordinate surface $u = 0$ and the horizontal plane $y = y_m$ with $u = y_m$. The problem consists in determining the $\boldsymbol{S}-$matrix by solving Maxwell's equations under covariant form expressed in the coordinate system

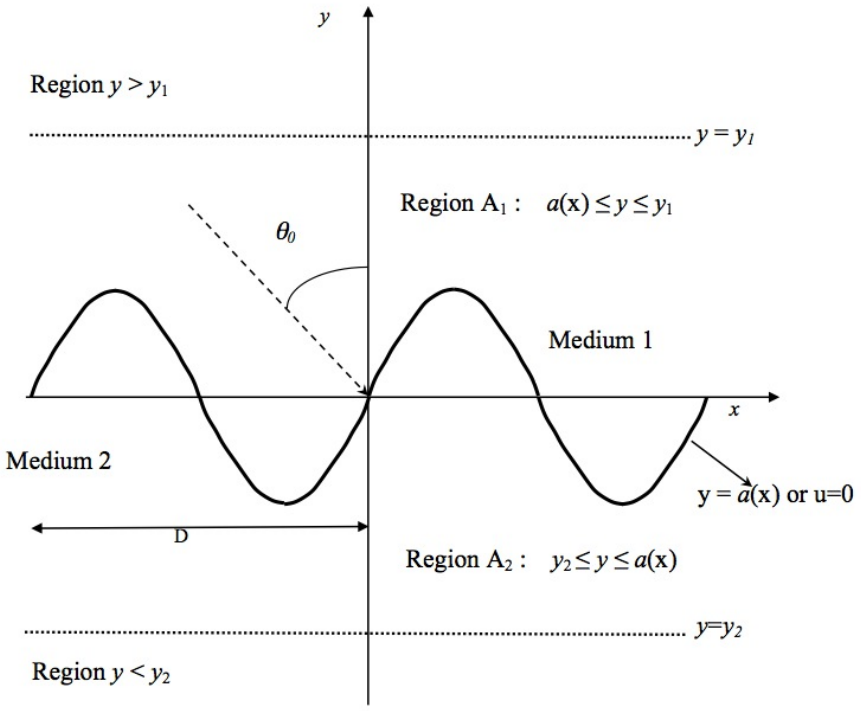

FIGURE 8.1: Grating illuminated by a plane wave under incidence θ_0 . The space is divided in four regions.

(3.9) and by using continuity relations in planes $y = y_m$ between covariant components of fields and Cartesian ones.

The covariant components $(v_{x'}, v_u, v_{z'})$ of a vector v are obtained from the Cartesian coordinate (v_x, v_y, v_z) as follows:

$$\begin{pmatrix} v_{x'} \\ v_u \\ v_{z'} \end{pmatrix} = A \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \tag{8.2}$$

Here A is the transformation matrix:

$$A = A^i_{i'} = \begin{pmatrix} \frac{\partial x}{\partial x'} & \frac{\partial y}{\partial x'} & \frac{\partial z}{\partial x'} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial z'} & \frac{\partial y}{\partial z'} & \frac{\partial z}{\partial z'} \end{pmatrix} = \begin{pmatrix} 1 & (y_m - u)\frac{\dot{a}(x)}{y_m} & 0 \\ 0 & \frac{y_m - a(x)}{y_m} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{8.3}$$

where $\dot{a}(x) = \frac{da(x)}{dx}$.

From the above equation, we can make several observations:

- the covariant component $v_{z'}(x,u,z)$ is equal to the Cartesian component $v_z(x,y,z)$ and it is parallel to the surface $u=0$ and $u=y_m$.
- The covariant component $v_u(x,y,z)$ is proportional to the Cartesian component $v_y(x,y,z)$.
- The covariant component $v_{x'}(x,u,z)$ is tangential to the grating interface $u=0$ and can be identified with the Cartesian component $v_x(x,y,z)$ on the plane $u=y_m$.

The passage of covariant components $(v_{x'}, v_u, v_{z'})$ to the contravariant components $(v^{x'}, v^u, v^{z'})$ is obtained by the metric tensor G

$$\begin{pmatrix} v^{x'} \\ v^u \\ v^{z'} \end{pmatrix} = G^{-1} \begin{pmatrix} v_{x'} \\ v_u \\ v_{z'} \end{pmatrix} \tag{8.4}$$

The tensor G is defined by (see equation 4.6):

$$G = g_{i'j'} = AG_cA^t = \begin{pmatrix} g_{x'x'} & g_{x'u} & g_{x'z'} \\ g_{ux'} & g_{uu} & g_{uz'} \\ g_{z'x'} & g_{z'u} & g_{z'z'} \end{pmatrix} \tag{8.5}$$

where G_c is the Cartesian system that is equal to the identity matrix and A^t denotes the transpose matrix of A.

So, we find:

$$G = \begin{pmatrix} 1+\frac{(y_m-u)^2}{y_m^2}\dot{a}^2(x) & \frac{(y_m-a(x))(y_m-u)}{y_m^2}\dot{a}(x) & 0 \\ \frac{(y_m-a(x))(y_m-u)}{y_m^2}\dot{a}(x) & \frac{(y_m-a(x))^2}{y_m^2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{8.6}$$

and

$$G^{-1} = g^{i'j'} = \begin{pmatrix} g^{x'x'} & g^{x'u} & g^{x'z'} \\ g^{ux'} & g^{uu} & g^{uz'} \\ g^{z'x'} & g^{z'u} & g^{z'z'} \end{pmatrix} = \begin{pmatrix} 1 & \frac{u-y_m}{y_m-a(x)}\dot{a}(x) & 0 \\ \frac{u-y_m}{y_m-a(x)}\dot{a}(x) & \frac{(u-y_m)^2\dot{a}^2(x)+y_m^2}{(y_m-a(x))^2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{8.7}$$

$$g = det(G) = \frac{(y_m-a(x))^2}{y_m^2} \tag{8.8}$$

From the Maxwell's equations written in the coordinate system, we find:

$$\begin{cases} \frac{\partial F^{(m)}(x,u)}{\partial u} = -jk^{(m)}\frac{y_m - a(x)}{y_m}G^{(m)}(x,u) + jk^{(m)}\frac{y_m-u}{y_m}\dot{a}(x)G_u^{(m)}(x,u) \\ \frac{\partial G^{(m)}(x,u)}{\partial u} = \frac{\partial G_u^{(m)}(x,u)}{\partial x} - jk^{(m)}\frac{y_m-a(x)}{y_m}F^{(m)}(x,u) \\ \frac{\partial F^{(m)}(x,u)}{\partial x} = -jk^{(m)}\frac{y_m-u}{y_m}\dot{a}(x)G^{(m)}(x,u) + j\frac{y_m k^{(m)}}{y_m-a(x)}(1+\frac{(y_m-u)^2}{y_m^2}\dot{a}^2(x))G_u^{(m)}(x,u) \end{cases} \tag{8.9}$$

In $E_{//}$ polarization, $F^{(m)} = E_{z'}^{(m)}, G^{(m)} = Z^{(m)}H_{x'}^{(m)}$ and $G_u^{(m)} = Z^{(m)}H_u^{(m)}$. In $H_{//}$ polarization, $F^{(m)} = Z^{(m)}H_{z'}^{(m)}, G^{(m)} = -E_{x'}^{(m)}$ and $G_u^{(m)} = -E_u^{(m)}$. The covariant components $F^{(m)}(x,u)$ and $G^{(m)}(x,u)$ are tangential to the grating interface and they appear in the boundary conditions at $u = 0$.

The periodic functions $a(x)$ and $\dot{a}(x)$ are expanded in Fourier series. The periodicity with respect to the variable x, as well as the excitation by a plane wave, leads to an expansion of founctions $F^{(m)}(x,u), G^{(m)}(x,u)$ and $G_u^{(m)}(x,u)$ in terms of the quasi-periodic functions $exp(-j\alpha_n x)$.

$$\begin{cases} F^{(m)}(x,u) = \sum_{n=-\infty}^{+\infty} f_n^{(m)}(u)exp(-j\alpha_n x) \\ G^{(m)}(x,u) = \sum_{n=-\infty}^{+\infty} g_n^{(m)}(u)exp(-j\alpha_n x) \\ G_u^{(m)}(x,u) = \sum_{n=-\infty}^{+\infty} g_{u,n}^{(m)}(u)exp(-j\alpha_n x) \end{cases} \tag{8.10}$$

Substituting these expansions into (8.9) and projecting on basis functions $exp(-j\alpha_n x)$ leads to a set of partial differential equations relating $f_n^{(m)}(u)$ and $g_n^{(m)}(u)$:

$$\begin{cases} \frac{d\vec{f}^{(m)}(u)}{du} = -j\frac{y_m-u}{y_m}\boldsymbol{D}(u)\dot{\boldsymbol{A}}\boldsymbol{\alpha}\vec{f}^{(m)}(u) - jk^{(m)}\boldsymbol{D}(u)\vec{g}^{(m)}(u) \\ \frac{d\vec{g}^{(m)}(u)}{du} = \frac{j}{k^{(m)}}(\boldsymbol{\alpha}\boldsymbol{D}(u)\boldsymbol{\alpha} - k^{(m)2}\boldsymbol{B})\vec{f}^{(m)}(u) - j\frac{y_m-u}{y_m}\boldsymbol{\alpha}\boldsymbol{D}(\boldsymbol{u})\dot{\boldsymbol{A}}\vec{g}^{(m)}(u) \end{cases} \tag{8.11}$$

where

$$\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{A}/y_m, \qquad \boldsymbol{D} = \boldsymbol{B}(\boldsymbol{I} + \frac{(y_m-u)^2}{y_m^2}\dot{\boldsymbol{A}}\dot{\boldsymbol{A}})^{-1} \tag{8.12}$$

Vector $\vec{f}^{(m)}$ and $\vec{g}^{(m)}$ contains the coefficients $f_n^{(m)}(u)$ and $g_n^{(m)}(u)$ respectively. $\boldsymbol{\alpha}$ is a diagonal matrix with the propagation coefficients α_n along the diagonal and $\boldsymbol{I}$ is the identity matrix. $\boldsymbol{A}$ is the Toeplitz matrix generated by the Fourier coefficients a_n of functions $a(x)$, such that its (p,q) element is a_{p-q}. $\dot{\boldsymbol{A}}$ is the Toeplitz matrix generated by

the Fourier $\dot{\boldsymbol{a}}_n$ of the profile derivative. The numerical solution of system (8.11) requires a truncation order M. Then, the covariant components $F^{(m)}(x,u)$ and $G^{(m)}(x,u)$ are described by only $2M+1$ expansion coefficients $f_n^{(m)}(u)$ and $g_n^{(m)}(u)$, and the Cartesian components $F_c^{(m)}(x,y)$ and $G_c^{(m)}(x,y)$ within the region $y \geq y_1$ or $y \leq y_2$ by the sum of $2M+1$ outgoing plane waves (amplitude $c_n^{(1+)}, c_n^{(2-)}$) and $2M+1$ incoming waves (amplitude $c_n^{(1-)}, c_n^{(2+)}$), see figure 3.2.

8.3 Numerical Implementations

The differential system (8.11) has non-constant coefficients and represents an initial value problem. We propose a procedure for obtaining the $N-$dimensional $\boldsymbol{S}-$matrix ($N = 4M+2$). First, we define N independent vectors satisfying the boundary conditions on the grating interface $u = 0$. In the $E_{//}$ polarization, the continuity relations on the electric and magnetic components are given by:

$$\begin{cases} F^{(1)}(x,u=0) = F^{(2)}(x,u=0) \\ \nu^{(1)} G^{(1)}(x,u=0) = \nu^{(2)} G^{(2)}(x,u=0) \end{cases} \tag{8.13}$$

In $H_{//}$ polarization, we have:

$$\begin{cases} \nu^{(1)} F^{(1)}(x,u=0) = \nu^{(2)} F^{(2)}(x,u=0) \\ G^{(1)}(x,u=0) = G^{(2)}(x,u=0) \end{cases} \tag{8.14}$$

Substituting (8.10) into (8.13) and projecting on function $exp(-j\alpha_n x)$ give in $E_{//}$ polarization:

$$\begin{cases} f_n^{(1)}(u=0) = f_n^{(2)}(u=0) \\ \nu^{(1)} g_n^{(1)}(u=0) = \nu^{(2)} g_n^{(2)}(u=0) \end{cases} \tag{8.15}$$

Similarly, in $H_{//}$ polarization, we get:

$$\begin{cases} \nu^{(1)} f_n^{(1)}(u=0) = \nu^{(2)} f_n^{(2)}(u=0) \\ g_n^{(1)}(u=0) = g_n^{(2)}(u=0) \end{cases} \tag{8.16}$$

$f_n^{(m)}(u=0)$ and $g_n^{(m)}(u=0)$ are the initial values of field components $F^{(m)}(x,u)$ and $G^{(m)}(x,u)$. We define N independent vectors from different initial values $f_n^{(m)}(u=0)$ and $g_n^{(m)}(u=0)$. In the $E_{//}$ polarization, they are contained in the following matrix:

$$\begin{pmatrix} \boldsymbol{F}^{(m)}(u=0) \\ \boldsymbol{G}^{(m)}(u=0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I}/\nu^{(m)} & \boldsymbol{I}/\nu^{(m)} \end{pmatrix} \tag{8.17}$$

I is the $(2M+1)-$dimensional identity matrix. In the $H_{//}$ polarization, we use:

$$\begin{pmatrix} \boldsymbol{F}^{(m)}(u=0) \\ \boldsymbol{G}^{(m)}(u=0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}/\nu^{(m)} & \boldsymbol{I}/\nu^{(m)} \\ \boldsymbol{I} & -\boldsymbol{I} \end{pmatrix} \tag{8.18}$$

For a perfectly conducting grating, the fields inside the conductor vanish, see equation (2.87) - (2.90). The tangential component of the electric field is zero on the grating surface. The tangential component of the magnetic field is different to zero and gives the surface current density. For a perfect conduction, we consider $2M+1$ independent vectors and in $E_{//}$ polarization,

$$\begin{pmatrix} \boldsymbol{F}^{(m)}(u=0) \\ \boldsymbol{G}^{(m)}(u=0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{pmatrix} \tag{8.19}$$

and in $H_{//}$ polarization,

$$\begin{pmatrix} \boldsymbol{F}^{(m)}(u=0) \\ \boldsymbol{G}^{(m)}(u=0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix} \tag{8.20}$$

The coupled differential equation system (8.11) is solved for each column vector of initial condition matrices. For each medium, this step requires numerical integrations with iterative algorithm from $u=0$ to $u=y_m$ and gives final values of N vectors (contained in matrices $\boldsymbol{F}^{(m)}(u=y_m)$ and $\boldsymbol{G}^{(m)}(u=y_m)$). In the horizontal plane $y=y_m$, the covariant components $F^{(m)}(x,u)$ and $G^{(m)}(x,u)$ become identified with the Cartesian components $F_c^{(m)}(x,y)$ and $G_c^{(m)}(x,y)$. By projecting on basis functions $exp(-j\alpha_n x)$ of the connection relationship, we find:

$$\begin{cases} f_n^{(m)}(u=y_m) = c_n^{(m+)} exp(-j\beta_n^{(m)} y_m) + c_n^{(m-)} exp(+j\beta_n^{(m)} y_m) \\ g_n^{(m)}(u=y_m) = c_n^{(m+)} \frac{\beta_n^{(m)}}{k^{(m)}} exp(-j\beta_n^{(m)} y_m) - c_n^{(m-)} \frac{\beta_n^{(m)}}{k^{(m)}} exp(+j\beta_n^{(m)} y_m) \end{cases} \tag{8.21}$$

For each medium, we deduce from (8.21) the amplitudes $c_n^{(m\pm)}$. For a dielectric grating, the $N-$dimensional $S-$matrix is obtained as follows:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{C}^{(1+)} \\ \boldsymbol{C}^{(2-)} \end{pmatrix} \begin{pmatrix} \boldsymbol{C}^{(1-)} \\ \boldsymbol{C}^{(2+)} \end{pmatrix}^{-1} \tag{8.22}$$

For a perfectly conducting grating, we have:

$$\boldsymbol{S} = \boldsymbol{C}^{(1+)}(\boldsymbol{C}^{(1-)})^{-1} \tag{8.23}$$

8.4 On the computational time

The used iterative algorithm is a variable order Adams-Bashforth-Moulton PECE solver (Prediction/Evaluation/Correction/Evaluation). It is a multistep solver and needs the solutions at several preceding spatial points to compute the current solution. Results presented in the next section are provided using MATLAB and the solver ODE113.

The dominant computational cost of the proposed method is due to numerical integrations and depends on the relative and absolute tolerances used by the algorithm. This relative tolerance controls the number of correct digits in all solution components, except those smaller than the absolute tolerance thresholds. The absolute tolerance is a threshold below which the value of the i^{th} solution component is unimportant. The absolute error tolerances determine the accuracy when the solution approaches zero. For a given grating with given tolerances, the computational cost is $\mathcal{O}(N^3)$.

For a multilayer grating ($n+1$ layer labeled with $1, 2, ..., n, n+1$ in sequence), we will show that it is possible to form the local scattering matrix $\boldsymbol{S}_{i,i+1}$ and then glue them to form the global matrix $\boldsymbol{S}_{1,n+1}$. We can also explore the parallelism in this gluing operation. For example, we can in the first setp, obtain $\boldsymbol{S}_{1,3}$, $\boldsymbol{S}_{3,5}$, $\boldsymbol{S}_{5,7}$... in parallel, and in the second step, obtain $\boldsymbol{S}_{1,5}$, $\boldsymbol{S}_{5,9}$,... in parallel, and so on. This needs only $\mathcal{O}(\log(n))$ steps to get the global scattering matrix. Moreover, for the calculation of the local scattering matrix $\boldsymbol{S}_{i,i+1}$, the proposed method leads to systems of first-order linear differential equations, the solution of which requires the choice of an iterative algorithm. The proposed method is based on numerical integrations with N independent initial vectors. The kernel of this computing process is the numerical integration.

This approach is also particularly well adapted to large-scale parallel and distributed architectures. Indeed, in the context of a distributed system comprising a network of machines, each of the problems could be solved on a machine whose architecture can be single or multiple processors. The proposed new method has a significant degree of coarse grain parallelism and requires little communication. These features offer the possibility of reducing dramatically the computation time. It's an advantage compared with the conventional C-method which leads to eigenvalue problems.

8.5 Numerical results

We consider a perfectly conducting sinusoidal grating defined by $a(x) = h_0 \cos(2\pi x/D)$ with $D = \lambda$. Under the incidence angle $\theta_0 = 30°$, the grating is in first order Littrow mounting and $\theta_{-1} = -\theta_0$, see paragraph 3.1.2. Figure 8.2 shows the efficiency curves under the polarization $H_{//}$ with h_0 varying from 0 to λ. The truncation order M is equal to 9 and $y_1 = -y_2 = 1.01max(a(x))$. The relative and absolute tolerances are equal to 10^{-6} and 10^{-9}, respectively. For a perfectly conducting grating illuminated under first-order Littrow mounting, the efficiency curves when the groove depth increases oscillate between 0 and 1.

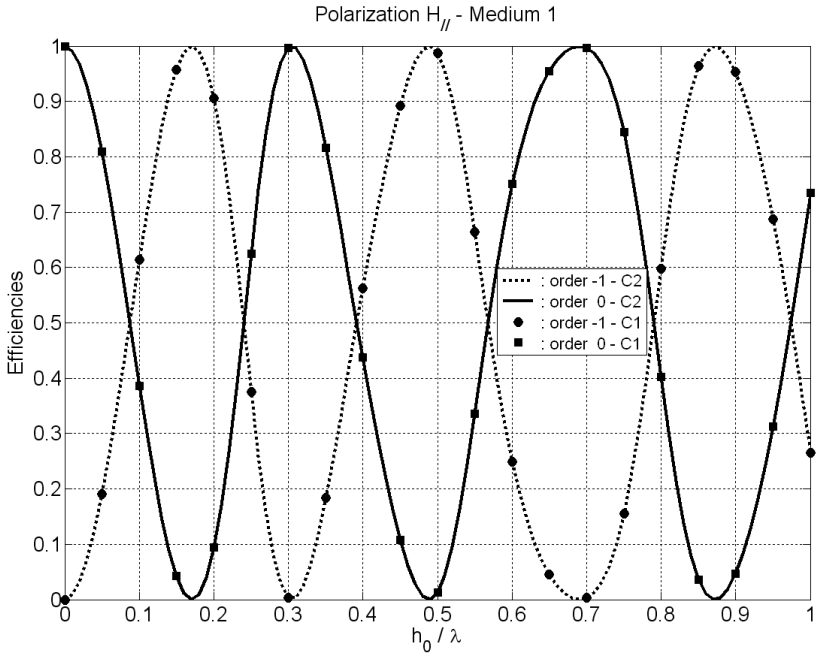

FIGURE 8.2: Reflected efficiencies versus sinusoidal grating amplitude Perfectly conducting grating in $H_{//}$ polarization.

As a result, the grating can present a perfect blazing in the minus-first-order. Figure 8.2 gives the efficiency curves obtained with the reference C-method (C1) and based on Fourier series factorization rules and the curves derived from the new version of the C-method (C2). For the reference method, the truncation order M is 18. This value provides a very good accuracy on the efficiencies. Figure 8.2 shows superimposed curves. Comparisons are conclusive and validate the proposed method. As shown in Figure 8.3, the new version of the C-method used with $M = 9$ checks the power balance with an error smaller than 10^{-3} whatever the groove depth. We obtain similar results in polarization $E_{//}$. The new approach is well-adapted to analyze this sinusoidal grating with the peak-peak amplitude smaller or equal to two wavelengths. The proposed method only uses Rayleigh expansions outside the grooves and does not use the Rayleigh hypothesis stipulating that the scattered field away from the surface can be extended down onto the grating even though it is formed by solely up-going waves. The theoretical validity of the Rayleigh hypothesis has given rise to some works for diffraction gratings. A classical result can be mentioned: for a perfectly conducting grating defined by $a(x) = h\cos(2\pi x/D)$ with the Dirichlet condition,the assumption does not hold if $h/D > 7\%$ [7]. The proposed method gives the efficiencies with a good accuracy with $h/D = 1$. Figure 8.3 also shows the error on the power balance for the conventional C-method used with the same value of truncation order. The error given by C1 is lowest when $h_0 < 3\lambda/4$. When $3\lambda/4 < h_0 < \lambda$, C1 and C2 methods are equivalent in terms of accuracy on the power balance.

Figure 8.4 shows the transmitted efficiencies in $H_{//}$ polarization for a lossless dielectric grating with a sinusoidal profile. The simulation parameters are: $\theta_0 = 15°, \nu^{(m)} = 3/2, D = 3\lambda/2$ and $0 \leq h_0 \leq \lambda$. The truncation order is equal to 9 and $y_1 = y_2 = 1.01max(a(x))$. The relative and absolute tolerances are equal to 10^{-6} and 10^{-9}. The grating presents four diffraction orders: $\theta_2 = 45.8°, \theta_1 = 15.8°, \theta_0 = 9.94°$ and $\theta_1 = 38.1°$. For $h_0/\lambda = 0$, the zeroth-order transmitted efficiency is equal to 0.98 and all other transmitted efficiencies are null. The incident energy is distributed into different diffraction orders when the groove depth increases. For $h_0/\lambda > 0.42$, the zeroth-order transmitted efficiency is smaller than 50%. Comparison between the reference C-method (C1) used with $M = 18$ and the new version (C2) used with $M = 9$ is conclusive and curves of efficiencies are superimposed.

Figure 8.5 gives the error on the power balance. The error increases with the groove

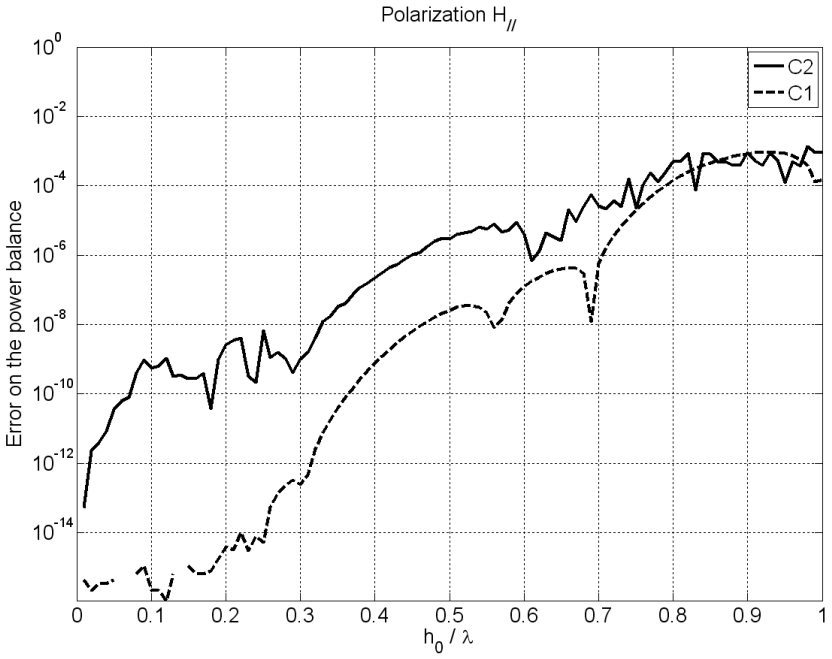

FIGURE 8.3: Error on the power balance versus sinusoidal grating amplitude. Perfectly conducting grating in $H_{//}$ polarization.

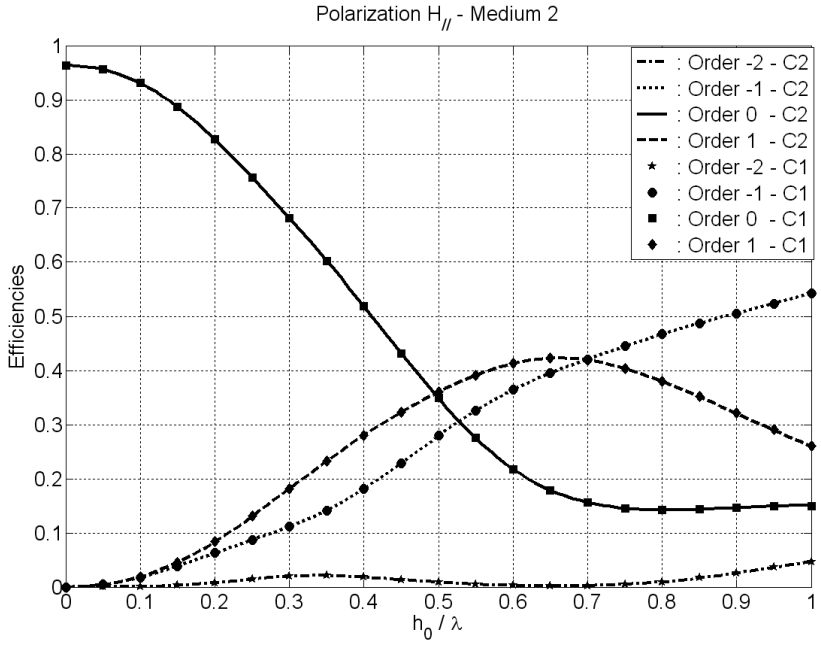

FIGURE 8.4: Transmitted efficiencies versus sinusoidal grating amplitude. Lossless dielectric grating in $H_{//}$ polarization.

depth and it is smaller than 10^{-3} whatever the profile amplitude. We obtain similar results in polarization $E_{//}$ .The new approach of the C-method is well-adapted to analyze this lossless dielectric grating when the peak-peak amplitude is smaller or equal to two wavelengths. Figure 8.5 also shows the error on the power balance for the conventional

C-method used with $M = 9$. The error given by C1 is lowest when $h_0 < 3\lambda/10$. On the range $3\lambda/10 < h_0 < 3\lambda/5$, C1 and C2 methods are equivalent in terms of accuracy on the power balance. When $3\lambda/5 < h_0 < \lambda$, the proposed method leads to the lowest errors.

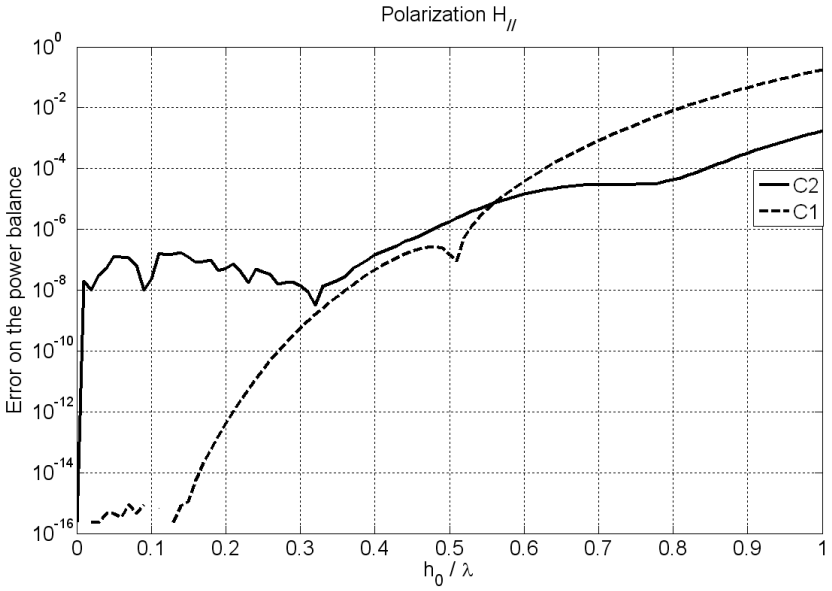

FIGURE 8.5: Error on the power balance versus sinusoidal grating amplitude. Lossless dielectric grating in $H_{//}$ polarization.

Figure 8.6 gives the reflected efficiencies versus the depth parameter p for a H-polarized metallic grating whose the profile is defined by,

$$a(x) = 0.48\lambda p \cos(2\pi x/D - 11\pi/90) + 0.12\lambda p \cos(4\pi x/D + 11\pi/90) \tag{8.24}$$

with $\lambda/D = 1$ and $0 \leq p \leq 1$. The optical index is equal to $0.07624 - 1.431j$ (the optical index of gold at $3200\mathring{A}$) [95]. The grating illuminated under 20° presents two diffraction orders ($\theta_0 = 20°, \theta_1 = 46.2°$). The relative and absolute tolerances are equal to 10^{-6} and 10^{-9}, respectively. Comparison between the reference C-method (C1) used with $M = 18$ and the new version (C2) used with $M = 9$ and $y_1 = y_2 = 1.01max(a(x))$ is conclusive. Curves of efficiencies are superimposed. For $p = 0.54$, both the efficiencies are close to zero.

Figure 8.7 gives the sum of efficiencies for the two polarizations. In the $E_{//}$ polarization case, the conduction losses are weak. In $H_{//}$, for the configuration defined by $p = 0.54$, surface plasmons are excited and cause a quasi-total absorption of incident energy [95].

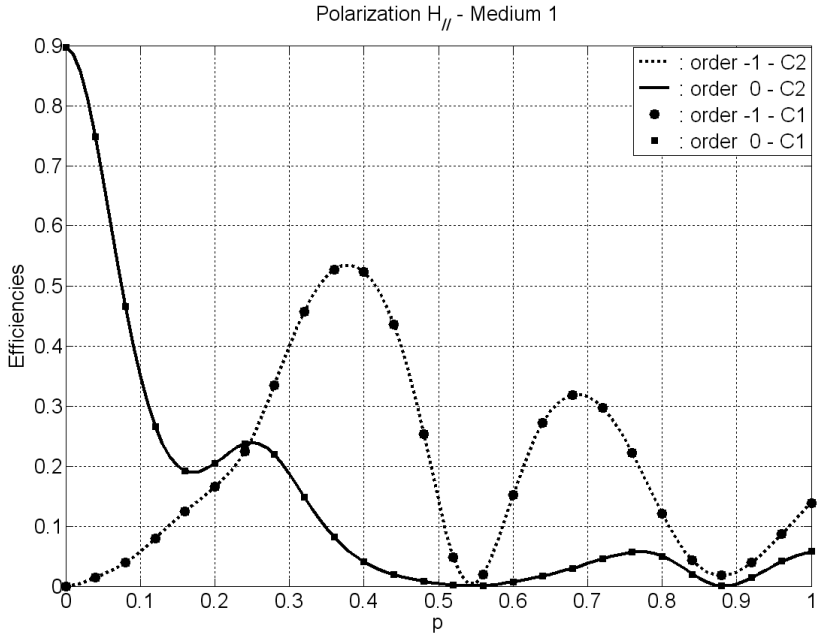

FIGURE 8.6: Reflected efficiencies versus grooves depth. Metallic grating in $H_{//}$ polarization.

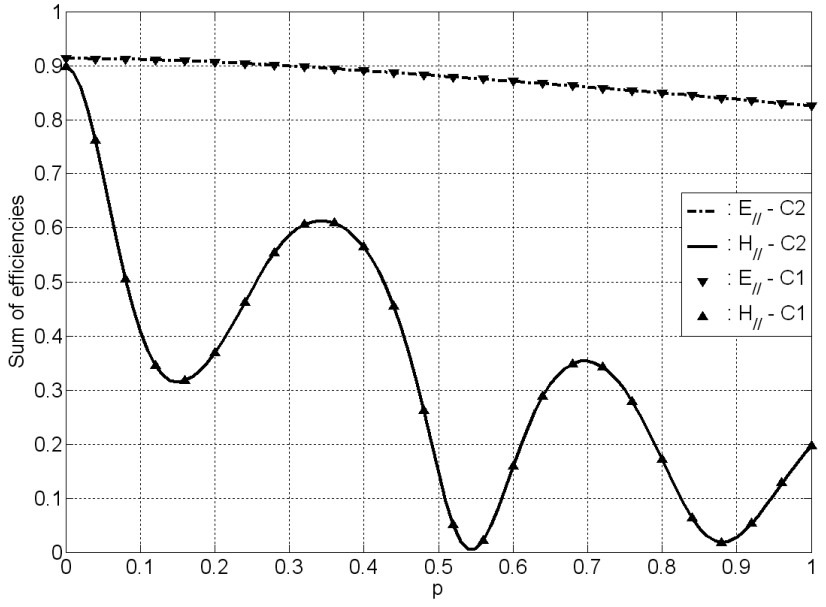

FIGURE 8.7: Sum of reflected efficiencies versus groove depth. Metallic grating in $E_{//}$ and $H_{//}$ polarizations.

The proposed method allows analyzing the metallic grating under consideration and the phenomenon of quasi-total absorption by surface plasmons in $H_{//}$ polarization.

8.6 Application to multilayer gratings with homogeneous medium

In this section, we extend the C-method as an initial value problem to multilayer gratings with homogeneous medium. We consider a $n+1$ layer diffraction grating, thus there are n interfaces separating the layers. From the uppermost to downmost, these layers are composed of medium 1 to medium $n+1$. Each medium has a constant optical index. Thus we can calculate the scattering matrix $\boldsymbol{S}_{i,i+1}$ which associate the incoming and outcoming waves from medium i and $i+1$. Then we collect all the scattering matrix of adjacent medium, $\boldsymbol{S}_{i,i+1}, i = 1, ..., n$ and obtain the global matrix $\boldsymbol{S}_{1,n+1}$ by combination of elementary matrices $\boldsymbol{S}_{i,i+1}, i = 1, ..., n$.

We consider a $n+1$ layer diffraction grating. The interface is represented by a periodic cylindrical surface $y = a_i(x), 1 \leq i \leq n$. This surface separates the medium i from the medium $i+1$. In Figure 8.8, two representative adjacent interfaces are shown. We consider separable layered grating, meaning that there exists a horizontal line $y = y_{i+1}$ separates the interface $y = a_i(x)$ and the interface $y = a_{i+1}(x)$ for each i. The interfaces in general have different functional forms and amplitudes, but there exists a value D such that D is the period of the function or a multiple of the period. The thickness h_i is measured between the middle lines of the two boundaries. The medium between the interface $y = a_i(x)$ and the interface $y = a_{i+1}(x)$ is homogeneous with optical index $\nu^{(i)}$, impedance $Z^{(i)}$ and wave number $k^{(i)}$.

As the interfaces are separable, we can consider each interface separately and then combine to form the global matrix. For each interface $y = a_i(x)$, we consider the problem as in the previous section. Then, we reduced the problem to the previous one.

We define the local scattering matrix($\boldsymbol{S}$-matrix) to relate the amplitudes of outgoing plane waves $(\boldsymbol{c}_n^{(i+)}, \boldsymbol{c}_n^{((i+1)-)})$ to those of incoming waves $(c_n^{(i-)}, c_n^{((i+1)+)})$ such that:

$$\begin{pmatrix} \boldsymbol{c}^{(i+)} \\ \boldsymbol{c}^{((i+1)-)} \end{pmatrix} = \boldsymbol{S}_{i,i+1} \begin{pmatrix} \boldsymbol{c}^{(i-)} \\ \boldsymbol{c}^{((i+1)+)} \end{pmatrix} \tag{8.25}$$

here we use $\boldsymbol{c}^{(m\pm)}$ to represent a vector containing the scattering amplitudes $c_n^{(m\pm)}$, $m = i, i+1$.

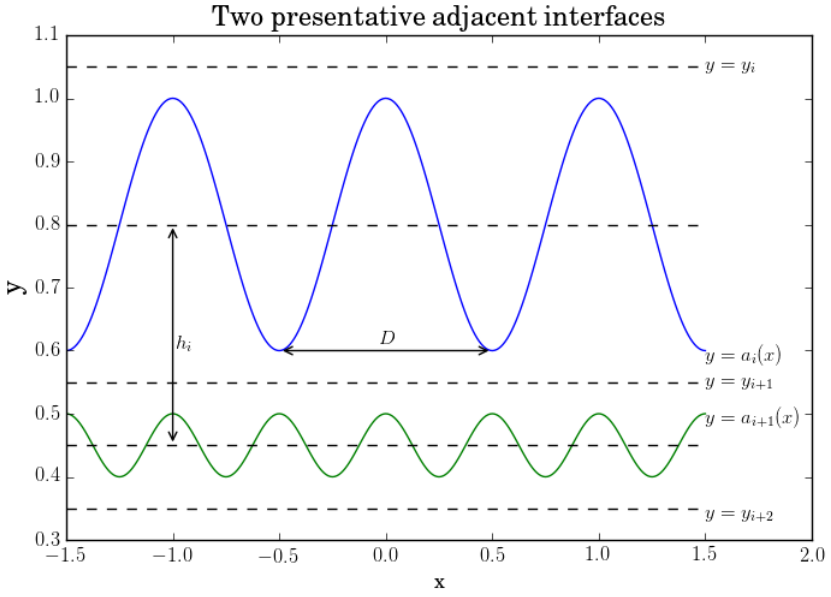

FIGURE 8.8: Notation for the description of a layered grating

Within the regions A_i and A_{i+1} defined by $a_i(x) \leq y \leq y_i$ and $y_{i+1} \leq y \leq a_i(x)$, we consider the non-orthogonal coordinate system defined as in the equation (8.1).

We then combine the local scattering matrices to form the global scattering matrix $\boldsymbol{S}_{1,n+1}$. In fact, if we have local scattering matrices $\boldsymbol{S}_{p,q}$ and $\boldsymbol{S}_{q,r}$ such that $p < q < r$, then we can obtain the scattering matrix $\boldsymbol{S}_{p,r}$.

$$\begin{cases} \begin{pmatrix} \boldsymbol{c}^{(p+)} \\ \boldsymbol{c}^{(q-)} \end{pmatrix} = \boldsymbol{S}_{p,q} \begin{pmatrix} \boldsymbol{c}^{(p-)} \\ \boldsymbol{c}^{(q+)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{S}_{p,q}^{(+-)} & \boldsymbol{S}_{p,q}^{(++)} \\ \boldsymbol{S}_{p,q}^{(--)} & \boldsymbol{S}_{p,q}^{(-+)} \end{pmatrix} \begin{pmatrix} \boldsymbol{c}^{(p-)} \\ \boldsymbol{c}^{(q+)} \end{pmatrix} \\ \begin{pmatrix} \boldsymbol{c}^{(q+)} \\ \boldsymbol{c}^{(r-)} \end{pmatrix} = \boldsymbol{S}_{q,r} \begin{pmatrix} \boldsymbol{c}^{(q-)} \\ \boldsymbol{c}^{(r+)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{S}_{q,r}^{(+-)} & \boldsymbol{S}_{q,r}^{(++)} \\ \boldsymbol{S}_{q,r}^{(--)} & \boldsymbol{S}_{q,r}^{(-+)} \end{pmatrix} \begin{pmatrix} \boldsymbol{c}^{(q-)} \\ \boldsymbol{c}^{(r+)} \end{pmatrix} \end{cases} \tag{8.26}$$

Eliminating the vectors $\boldsymbol{c}^{(q+)}$ and $\boldsymbol{c}^{(q-)}$, one glue the two scattering matrices to be one $\boldsymbol{S}_{p,r}$:

$$\boldsymbol{S}_{p,r} = \begin{pmatrix} \boldsymbol{S}_{p,q}^{(+-)} + \boldsymbol{S}_{p,q}^{(++)}(I - \boldsymbol{S}_{q,r}^{(+-)}\boldsymbol{S}_{p,q}^{(-+)})^{-1}\boldsymbol{S}_{q,r}^{(++)}\boldsymbol{S}_{p,q}^{(--)} & \boldsymbol{S}_{p,q}^{(++)}(I - \boldsymbol{S}_{q,r}^{(+-)}\boldsymbol{S}_{p,q}^{(-+)})^{-1}\boldsymbol{S}_{q,r}^{(++)} \\ \boldsymbol{S}_{q,r}^{(--)}(I - \boldsymbol{S}_{p,q}^{(-+)}\boldsymbol{S}_{q,r}^{(+-)})^{-1}\boldsymbol{S}_{p,q}^{(--)} & \boldsymbol{S}_{q,r}^{(-+)} + \boldsymbol{S}_{q,r}^{(--)}(I - \boldsymbol{S}_{p,q}^{(-+)}\boldsymbol{S}_{q,r}^{(+-)})^{-1}\boldsymbol{S}_{p,q}^{(-+)}\boldsymbol{S}_{q,r}^{(++)} \end{pmatrix} \tag{8.27}$$

So in this way, we glue $\boldsymbol{S}_{1,2}$ and $\boldsymbol{S}_{2,3}$ to get $\boldsymbol{S}_{1,3}$, and then glue $\boldsymbol{S}_{1,3}$ and $\boldsymbol{S}_{3,4}$ to get $\boldsymbol{S}_{1,4}$ and so on until we get the global matrix $\boldsymbol{S}_{1,n+1}$.

We perform an experiment in the article [96]. The grating considered is sinusoidal and defined by $a_1(x) = a_1 \cos(\frac{2\pi}{D}x), a_2(x) = a_2 \cos(\frac{2\pi}{D}x)$. Figure 8.9 shows the efficiency curves under the polarization $E_{//}$, with the value of $\sin\theta_0$ varying from 0.24 to 0.38. With the same parameters as in that paper, figure 8.9 is exactly the same as in the paper [96]. We can also analyze structure with non parallel interfaces. We perform more experiments to see how this figure changes when we change the amplitude of function $y = a_1(x)$. With other parameters fixed, figure 8.10 shows the curve when $a_1 = 0.01$, figure 8.11 shows the curve when $a_1 = 0.03$.

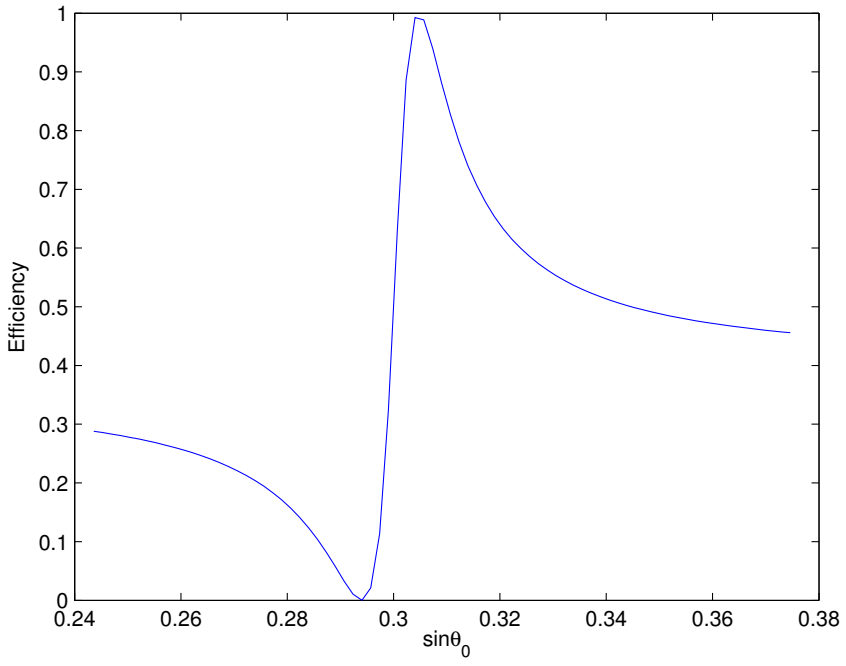

FIGURE 8.9: Diffraction efficiency of the zeroth reflected order of the sinusoidal grating. Parameters of the system are: $\nu^{(1)} = \nu^{(3)} = 1, \nu^{(3)} = 2.3, h_1 = 0.19\mu m, D = 0.37\mu m, \lambda = 0.6328\mu m, a_1 = a_2 = 0.02\mu m$, for $E_{//}$ polarization

We see from Figure 8.9 that the zeroth-order efficiency changes from 0 to 1. In Figure 8.10, the efficiency can not reach 0, it changes from a small positive value to 1. In figure 8.11, the efficiency can neither reach 0 nor 1. It can be observed that when a_1 varies from $0.03\mu m$ to $0.01\mu m$, the jump becomes steeper. One can also observe that the place of the jump moves towards the left direction. When we vary a_2 from $0.03\mu m$ to $0.01\mu m$, no similar phenomena can be observed. In fact, the curves seem almost stay the same. Figure 8.12 and figure 8.13 shows the curve when $a_2 = 0.03$ and $a_2 = 0.01$, respectively.

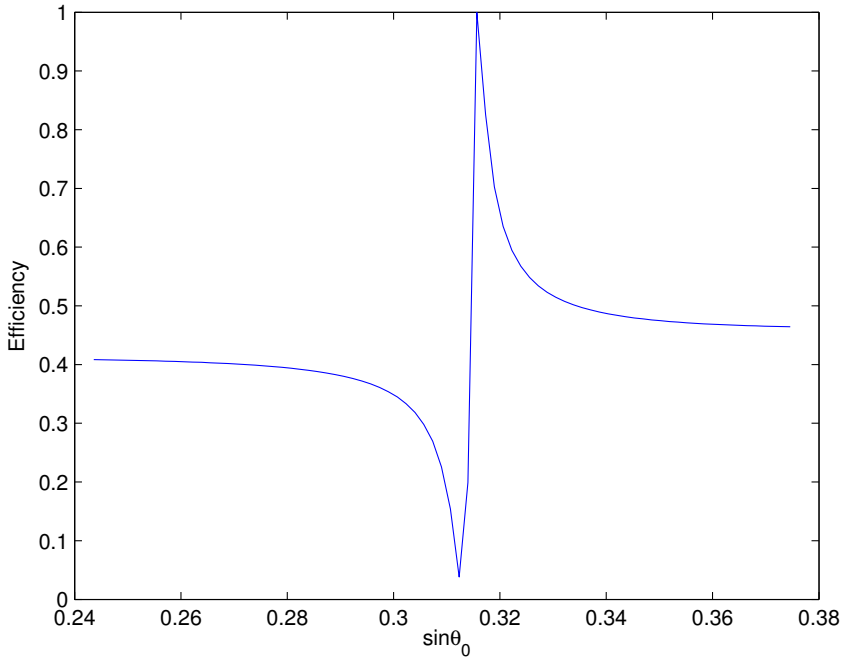

FIGURE 8.10: Diffraction efficiency of the zeroth reflected order of the sinusoidal grating. Parameters of the system are: $\nu^{(1)} = \nu^{(3)} = 1, \nu^{(2)} = 2.3, h_1 = 0.19\mu m, D = 0.37\mu m, \lambda = 0.6328\mu m, a_1 = 0.01\mu m, a_2 = 0.02\mu m$, for $E_{//}$ polarization

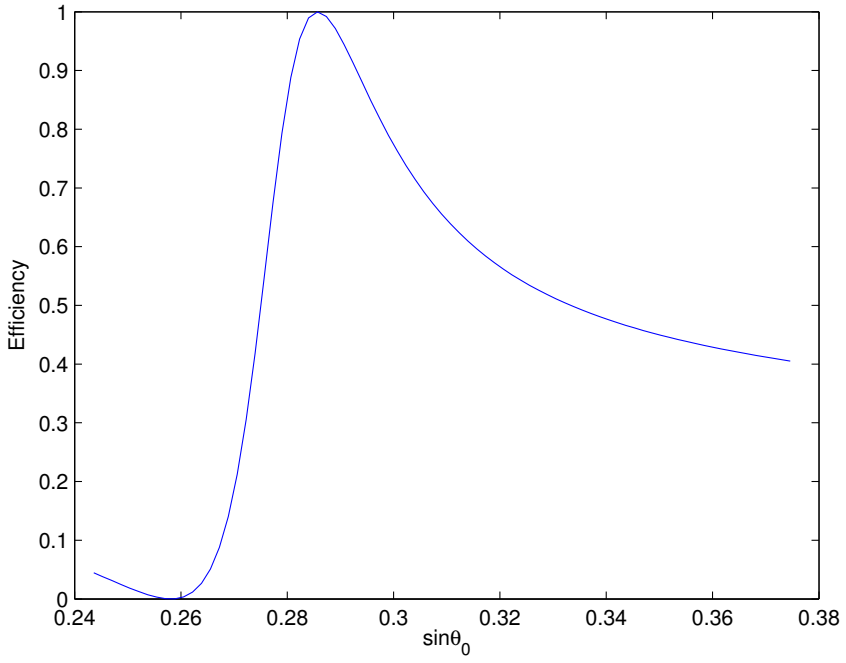

FIGURE 8.11: Diffraction efficiency of the zeroth reflected order of the sinusoidal grating. Parameters of the system are: $\nu^{(1)} = \nu^{(3)} = 1, \nu^{(2)} = 2.3, h_1 = 0.19\mu m, D = 0.37\mu m, \lambda = 0.6328\mu m, a_1 = 0.03\mu m, a_2 = 0.02\mu m$, for $E_{//}$ polarization

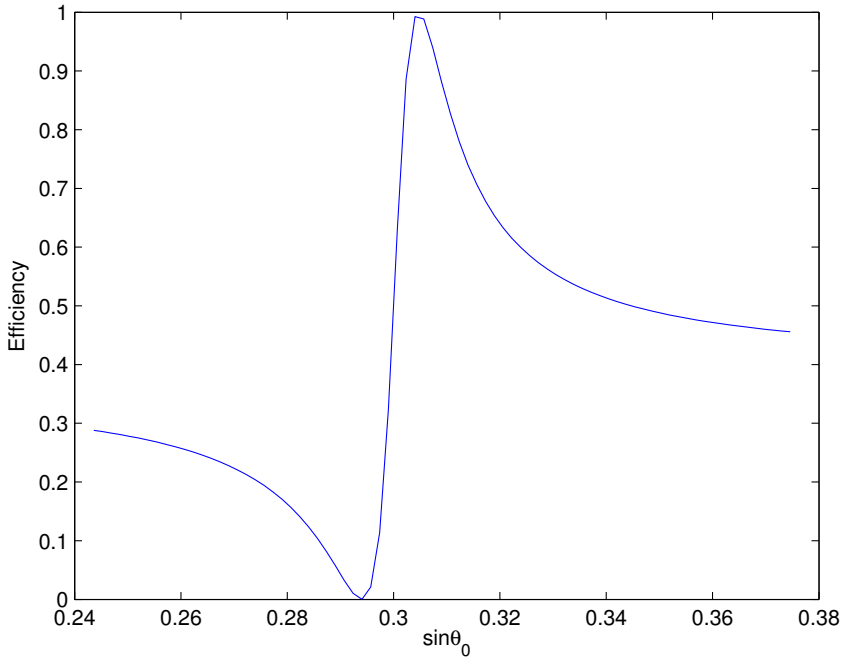

FIGURE 8.12: Diffraction efficiency of the zeroth reflected order of the sinusoidal grating. Parameters of the system are: $\nu^{(1)} = \nu^{(3)} = 1, \nu^{(2)} = 2.3, h_1 = 0.19\mu m, D = 0.37\mu m, \lambda = 0.6328\mu m, a_2 = 0.03\mu m, a_1 = 0.02\mu m$, for $E_{//}$ polarization

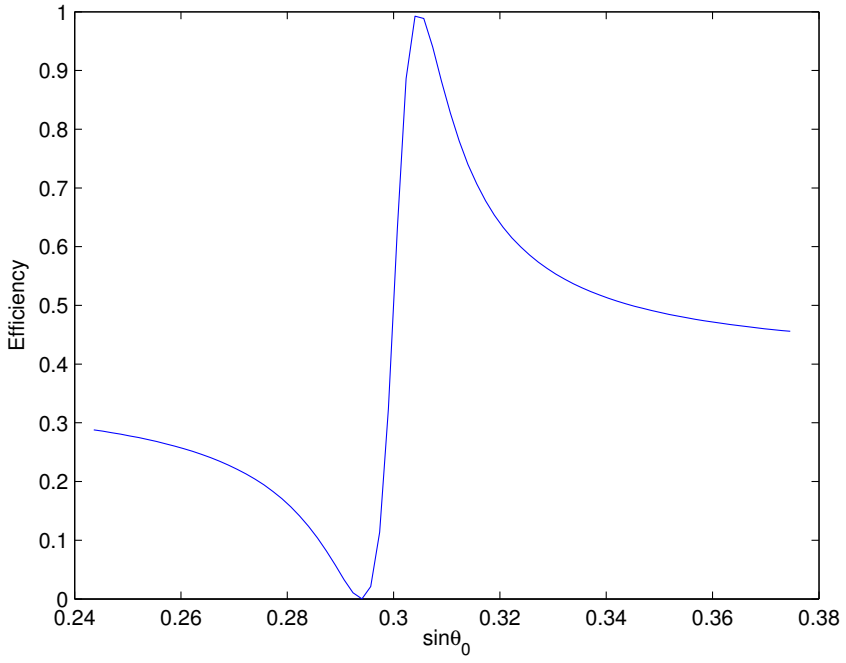

FIGURE 8.13: Diffraction efficiency of the zeroth reflected order of the sinusoidal grating. Parameters of the system are: $\nu^{(1)} = \nu^{(3)} = 1, \nu^{(2)} = 2.3, h_1 = 0.19\mu m, D = 0.37\mu m, \lambda = 0.6328\mu m, a_2 = 0.01\mu m, a_1 = 0.02\mu m$, for $E_{//}$ polarization

8.7 Conclusion

In this chapter, we studied our approach of the C-method as an initial value problem for the efficient calculation of the N-dimensional scattering matrix of a grating. We have shown that this formulation is an interesting tool for analyzing perfectly conducting or dielectric gratings with deep grooves. The proposed method allows analyzing the complex phenomenon of incident energy absorption. We then extend this method to multilayer with homogeneous medium. We applied this method to multilayer gratings with an arbitrary number of interfaces. We have shown how to combine the local scattering matrix to obtain the global one. We have validate our method by comparison our results with that from published paper. The proposed method has very good accuracy as well as a natural of two level parallel property. This new version of C-method is an attractive alternative to analyze multilayered grating having parallel or non-parallel interfaces. We are currently working on the extension of this method to apply it to multilayer with inhomogeneous medium and non-parallel interfaces.

Chapter 9

Conclusion

In this thesis, we study the electromagnetic diffraction by gratings and random rough surfaces. The C-method is an exact method based on Maxwells equations under covariant form written in a nonorthogonal coordinate system. The C-method leads to the eigenvalue problem of the high dimension, dense and non-symmetric scattering matrix. All the eigenvalues and eigenvectors of the scattering matrix are needed. The scattered field is expanded as a linear combination of eigensolutions satisfying the outgoing wave condition. The boundary conditions allow the diffraction amplitudes to be determined.

We propose the specifically designed parallel QR algorithm for the C-method. We present why we propose the "early shift" and how it can be used to accelerate the convergence. We also present the techniques of parallel QR with tightly coupled bulge chasing and parallel AED. These techniques are used to reduce the computational time of the C-method. We apply this specifically designed parallel QR algorithm to the scattering matrix. We also compare the computation time with that of the sequential code. The results show a significant speed up to approximately 40 for 64 cores with our new QR algorithm. This combination of early shift and other shifts can also be used in the problems such as linear-quadratic optimal control problem where a large number of eigenvalues and eigenvectors are needed and background of the original problem can provides very good initial approximations. This parallel QR algorithm can be used for analyzing crossed gratings or random rough surfaces. Comparisons with experimental

data for moderate roughness and isotropic or anisotropic very rough surfaces are conclusive in both co-polarized and cross-polarized components. Comparisons allow the validity of our approach.

As a prospect, we propose a spectral projection method to solve the eigenvalue problem efficiently. We propose a global eigensolver by a combination of the SS method and MIRAMns. This proposed global eigensolver allows us to calculate a large number of (or all) the eigenvalues of a generalized matrix. Compared to QR algorithm, this method has the advantage of having very good scalability. This promising method can be continued in future work.

This is the first attempt to combine MIRAMns and SS method to form a global eigensolver. Numerical experiments show this combination allows us to get all the eigenvalues and their corresponding eigenvectors. MIRAMns converges with less iterations than IRAM, and the SS method is very suitable for parallelization. The scalability of the global eigensolver is very good, we get almost linear speed up. The complexity of computation can be varying with the precision that is required. The precision can be increased with smaller sub-domain. .

For gratings, we propose a new version of C-method which leads to a differential system with initial conditions. We studied the new version of C-method as an initial value problem for the efficient calculation of the N-dimensional scattering matrix of a grating. We have shown that this formulation is an interesting tool for analyzing perfectly conducting or dielectric gratings with deep grooves. The proposed method allows analyzing the complex phenomenon of incident energy absorption. We then extend this method to multilayer with homogeneous medium. We applied this method to multilayer gratings with an arbitrary number of interfaces. We have shown how to combine the local scattering matrix to obtain the global one. We have validate our method by comparison our results with that from published paper. The proposed method has very good accuracy as well as a natural of two level parallel property. This new version of C-method is an attractive alternative to analyze multilayered grating having parallel or non-parallel interfaces.

For the future work, we plan to extend the spectral projection method to more general case. We also plan to extend our new version of C-method to multilayer with inhomogeneous medium.

Appendix A

Résumé

Titre:

Diffraction électromagnétique par des réseaux et des surfaces rugueuses aléatoires; Mise en œuvre de méthodes hautement efficaces pour la résolution de systèmes aux valeurs propres et de problèmes aux conditions initiales

Résumé:

Dans cette thèse, nous étudions la diffraction électromagnétique par des réseaux et des surfaces rugueuses aléatoires. Le méthode C est une méthode exacte développée pour ce but. Elle est basé sur des équations de Maxwell sous forme covariante écrites dans un système de coordonnées non orthogonales [27–29]. La discrétisation des équations de Maxwell dans ce système de coordonnées et la méthode de séparation des variables conduisent à une matrice de diffusion pleine et non symétrique dont il faut déterminer les valeurs propres. Toutes les valeurs et vecteurs propres sont nécessaires. Le champ diffusé est représenté par une combinaison linéaire des solutions propres satisfaisant à la condition d'onde sortante. Les conditions aux limites permettent de déterminer les amplitudes de diffusion associées à chaque solution propre. Cette méthode a été utilisée pour l'analyse des réseaux de diffraction utilisés en optique [30–50], des guides d'ondes et des surfaces rugueuses pour des problèmes de télédétection [55–65].

Nous nous concentrons sur l'aspect numérique de la méthode C, en développant une mise en œuvre efficace de cette méthode exacte. Des méthodes itératives de recherche de valeurs propres telles que les méthodes de sous-espace de Krylov ou les méthodes de Jacobi-Davidson [66] ont été développées pour traiter de problèmes de valeur propre

de très grande taille. Cependant, elles ne fournissent pas systématiquement toutes les valeurs propres et leurs vecteurs propres correspondants. Ainsi, ces méthodes itératives sont inefficaces pour la méthode C car toutes les valeurs propres et leurs vecteurs propres associés sont nécessaires. La méthode QR qui est basée sur les transformations semblables, calcule tous les éléments propres d'une matrice dense sans danger de manquer des solutions propres particulières. Nous proposons un algorithme QR parallèle conçu spécifiquement pour la méthode C pour résoudre le problème de valeur propre.

Cet algorithme QR parallèle est une variante de l'algorithme QR basée sur trois techniques: "early shift"[1], "bulge chasing"[2] [67, 68] parallèle et "aggressive early deflation (AED)" [3] parallèle [68, 69]. Nous proposons la technique "early shift" pour la matrice de diffusion en fonction des propriétés que nous avons observées. En effet, la méthode C et l'interprétation physique derrière la méthode C fournissent une très bonne approximation de certaines valeurs propres avant les calculs. L'utilisation de ces approximations comme "early shift" offre la possibilité de déflation rapide. Nous avons combiné le "early shift", le "shift" de Wilkinsons ansi que "le shift exceptionnel" afin d'accélérer la convergence de la méthode QR. Plus particulièrement, nous utilisons le "early shift" afin de déflater les valeurs propres approchées de la matrice de diffusion et accélérer ainsi la convergence de la méthode. L'algorithme double shift QR, pour des raisons d'économie et d'accélération de convergence, combine deux itérations avec shift en une seule itération avec double shift. A chaque itération, il engendre un bulbe d'éléments non-nuls à chasser par la suite (bulge chasing). Nous utilisons la version multishift de l'algorithme QR. Ainsi, pour le "bulge chasing", au lieu de chasser un seul bulbe, contenant deux shifts, une chaîne de plusieurs bulbes étroitement couplés, contenant chacun deux shifts, est poursuivi au cours d'une itération de QR multishift. Cette idée et la technique de "retard et accumulation" [67, 68] permet d'effectuer la majeur partie des calculs en termes d'opérations BLAS de niveaux 3 (essentiellement produits matrice-matrice) et augmenter ainsi l'efficacité de l'algorithme en termes de performances. L'AED est une stratégie de la déflation qui profite des perturbations de la matrice en dehors des éléments sous-diagonaux de la matrice de Hessenberg. Elle identifie et déflate les valeurs propres convergées longtemps avant la stratégie classique de déflation et peut améliorer considérablement la convergence de l'algorithme QR. Les résultats présentés dans cette

[1] La transformation de la matrice avec un décalage "prématuré" des éléments diagonaux permettant d'accélérer la convergence du processus itératif.

[2] La chasse aux élément non-nuls introduits au cours du calcul

[3] La dimiuntion de la taille de la matrice en fonction des valeurs propres connues et/ou déjà calculées

thèse mettent en évidence cette amélioration de performances pour le problème considéré.

En perspective, nous proposons une méthode de projection spectrale pour résoudre le problème de valeurs propres efficacement. Cette méthode proposeé afin de palier au problème de "scalability" de la méthode QR. Elle est baseé sur une combinaison de la méthode de Sakuria et Sugiura (SSM) [71] et "multiple implicitly restarted Arnolid method" avec des sous-espace imbriqué (MIRAMns) proposé par S. A. Shahzadeh Fazeli et al [72]. La méthode proposée nous permet de calculer un grand nombre de (ou toutes) les valeurs propres de la matrice généralisée. Comparé à l'algorithme QR, cette méthode a l'avantage d'avoir une très bonne "scalability". Les premiers resultats experimentaux sont très encourageants. Cette méthode prometteuse peut être poursuivie dans les travaux futurs.

La méthode C originale impose la résolution d'un système aux valeurs propres. En définissant un nouveau système de coordonnées non orthogonales, nous établissons une formulation qui évite la résolution d'un système aux valeurs propres. En particulier, nous voulons explorer la parallélisation potentielle de cette nouvelle méthode pour étudier des réseaux multicouches. Pour les réseaux de diffration à une interface, nous montrons que cette nouvelle version de la méthode C conduit à un système différentiel avec les conditions initiales. Nous montrons que cette nouvelle version de la méthode C peut être utilisée pour l'étude des réseaux comme un empilement d'interfaces délimitant des couches homogènes. Nous montrons que cette formulation est un outil efficace pour analyser des réseaux parfaitement conducteurs ou des réseaux diélectriques aux sillons très profonds. La méthode proposée permet d'analyser le phénomène complexe de l'absorption totale d'énergie incidente (par des modes plasmons). Nous avons appliqué cette méthode à des réseaux multicouches avec un nombre arbitraire d'interfaces. Nous avons montré comment combiner la matrice de diffusion locale pour obtenir une matrice globale caractéristique de la structure multicouches. Nous montrons que cette nouvelle version de la méthode C permet d'analyser les réseaux multicouches ayant des interfaces parallèles ou non parall'eles. Nous avons validé notre méthode en comparant les résultats fournis par cette méthode avec des résultats publiés. Nous montrons que la nouvelle version de la méthode C a une très bonne précision et permet une parallélisation à deux niveaux de la propriété parallèle à deux niveaux [73, 74].

Appendix B

Publications

Published:

C.Pan, R.Dusséaux and N.Emad, "Parallel QR algorithm for the C-method: Application to the diffraction by gratings and rough surfaces". *Apr. 2015*
23^{rd} High Performance Computing Symposium (HPC 2015), pp. 166-173, Alexandria, VA, USA.

C.Pan, R.Dusséaux, M.Fall and N.Emad, "The curvilinear coordinate method as an initial value problem: Application to gratings". *Jan. 2015*
JOSA A, Vol. 32, Issue 1, pp. 143-149 (2015).

C.Pan, R.Dusséaux and N.Emad, "The C-method as an initial value problem: application to multilayer gratings". *July 2015*
Progress In Electromagnetics Research Symposium (PIERS 2015), pp.1297-1301, Prague, Czech Republic.

In preparation:

C.Pan, R.Dusséaux and N.Emad, "Numerical Study of scattering of electromagnetic waves from two-dimensional random rough surface: parallel QR algorithm for the C-method ".

C.Pan, R.Dusséaux and N.Emad, "An extension of the curvilinear coordinate method - Application to gratings of inhomogeneous and multilayered media ".

C.Pan, N.Emad and R.Dusséaux, "Spectral projection method as a global eigensolver ".

Bibliography

[1] R. Gómez Martín. *Electromagnetic field theory for physicists and engineers: Fundamentals and Applications*. Universidad de Granada, 2009.

[2] J.D. Jackson. *Classical Electrodynamics (3^{rd} Ed.)*. Wiley, August 1998.

[3] J.A. Stratton and S.J. Adams. *Electromagnetic Theory*. International series in physics. McGraw-Hill, 2007.

[4] A. Ishimaru. *Electromagnetic wave propagation, radiation and scattering*. Prentice Hall, 1991.

[5] C. Palmer and E. Loewen. *Diffraction grating handbook*. Newport Corporation, 2005.

[6] R. Petit. *Electromagnetic Theory of Gratings*. Springer-Verlag. Berlin, Heidelberg, New-York, 1980.

[7] R. Petit and M. Cadilhac M. Sur la diffraction d'une onde plane par un rseau infiniment conducteur. *C.R. Acad. Sci. B*, pages 468–471, 1966.

[8] R.F. Millar. On the Rayleigh assumption in scattering by a periodic surface. *Proc. Camb. Phil. Soc.*, 65:773–791, 1969.

[9] R.F. Millar. On the Rayleigh assumption in scattering by a periodic surface - II. *Proc. Camb. Phil. Soc.*, 69:217–225, 1971.

[10] P.M. Van den Berg and J.T. Fokkema. The Rayleigh hypothesis in the theory of reflection by a grating. *J. Opt. Soc. Am.*, 69:27–31, 1979.

[11] J.B. Keller. Singularities and Rayleigh's hypothesis for diffraction gratings. *J.Opt.Soc.Am.A*, 17:456–457, 2000.

[12] A.I. Kleev and A.B. Manenkov. The convergence of point-matching techniques. *IEEE Trans. Antennas Propagat.*, 37:50–54, 1989.

[13] P. Beckmann and A. Spizzichino. *The scattering of electromagnetic waves from rough surfaces.* Pergamon, Oxford, 1963.

[14] J. A. Ogilvy. *Theory of wave scattering from random rough surfaces.* Adam Hilger, 1991.

[15] C. Bourlier. Azimuthal harmonic coefficients of the microwave backscattering from a non-Gaussian ocean surface with the first-order SSA model. *IEEE Trans. Geosc. Remote Sens*, 42:2600–2611, 2004.

[16] C. Bourlier, G. Berginc, and J. Saillard. Theoretical study of the Kirchhoff integral from a two-dimensional randomly rough surface with shadowing effect: application to the backscattering coefficients for a perfectly-conducting surface. *Waves in Random Media*, pages 91–118, 2001.

[17] G. Voronovich. *Wave scattering from rough surfaces.* Springer, Berlin, 1994.

[18] K.F. Warnick and W.C. Chew. Numerical simulation methods for rough surface scattering. *Wave Random Media*, 11(1):R1–30, 2001.

[19] M. Saillard and A. Sentenac. Rigorous solutions for electromagnetic scattering from rough surfaces. *Wave Random Media*, 11(3):R103–137, 2001.

[20] K. Pak, L. Tsang, and J.T. Johnson. Numerical simulations and backscattering enhancement of electromagentic waves from two-dimensional dielectric random rough surfaces with the sparse-matrix canonical grid method. *J.Opt.Soc.Am.A*, 18:1515–1529, 1997.

[21] R.L. Wagner, J. Song, and W.C. Chew. Monte Carlo simulations of electromagentic scattering from two-dimensional random rough surfaces. *IEEE Trans. Antennas Propagat.*, 45:235–245, 1997.

[22] V. Jandhyala, B. Shanker, E. Michielssen, and W.C. Chew. Fast algorithm for the analysis of scattering by dielectric rough surfaces. *J.Opt.Soc.Am.A*, 15:1877–1885, 1998.

[23] D. Torrungrueng and J.T. Johnson. Numerical studies of backscattering enhancement of electromagnetic waves from two-dimensional random rough surfaces with the forward-backward novel spectral acceleration method. *J.Opt.Soc.Am.A*, 18:2518–2526, 2001.

[24] G. Soriano and M. Saillard. Scattering of electromagentic waves from two-dimensional rough surfaces with an impedance approximation. *J.Opt.Soc.Am.A*, 18:124–133, 2001.

[25] L. Tsang, Q. Li, D. Chen, P. Xu, and V. Jandhyala. Wave scattering with the UV multilevel partitioning method: 2. three-dimensional problem of nonpenetrable surface scattering. *Radio Sci.*, 39:RS5011, 1–11, 2004.

[26] M. Born and E. Wolf. *Principles of optics Electromagnetic theory of propagation, Interference and diffraction of light.* Pergamon Press, 1959.

[27] E.J. Post. *Formal structure of electromagnetic.* North-Holland, Amsterdam, 1962.

[28] J. Chandezon, G. Raoult, and D. Maystre. A new theoretical method for diffraction gratings and its numerical application. *Journal of Optics*, 11(4):235–241, 1980.

[29] J. Chandezon and G. Raoult. Application d'une nouvelle méthode de résolution des équations de Maxwell à l'étude de la propagation des ondes électromagnétiques dans les guides périodiques. *Ann Telecom*, 36:305–314, 1981.

[30] E. Popov, L. Mashev, and D. Maystre. Conical diffraction mounting. generalization of a rigourous differential method. *J.Optics (Paris)*, 17:175–180, 1986.

[31] S.J. Elston, G.P. Bryan-Brown, and J.R. Sambles. Polarization conversion from diffraction gratings. *Phys.Rev.B*, 44:6393–6400, 1991.

[32] J. Chandezon, M.T. Dupuis, G. Cornet, and D. Maystre. Multicoated gratings, a differential formalism applicable in the entire optical region. *J.Opt.Soc.Am.A*, 72:839–846, 1982.

[33] L. Li. Multilayer-coated diffraction gratings: differential method of Chandezon et al revisited. *J. Opt. So. Am. A*, 11:2816–2828, 1994.

[34] N.P.K. Cotter, T.W. Preist, and J.R. Sambles. Scattering-matrix approach to multilayer diffraction. *J.Opt.Soc.Am.A*, 72:1097–1103, 1995.

[35] G. Granet, J.P. Plumey, and J. Chandezon. Scattering by a periodically corrugated dielectric layer with non-identical faces. *Pure Appl. Opt.*, 4:1–5, 1995.

[36] T.W. Preist, N.P.K Cotter, and J.R. Sambles. Periodic multilayer gratings of arbitrary shape. *J.Opt.Soc.Am.A*, 12:1740–1748, 1995.

[37] L. Li, G. Granet, J.P. Plumey, and J. Chandezon. Some topics in extending the C-method to multilayer-coated gratings of different profiles. *Pure Appl.Opt.*, 5:141–156, 1996.

[38] G. Granet. Diffraction par des surfaces bipériodiques: résolution en coordonnées non-orthogonales. *Pure Appl. Opt.*, 4:777–793, 1995.

[39] G. Granet. Analysis of diffraction by surface-relief crossed gratings with use of the Chandezon method: Application to multilayer crossed gratings. *J.Opt.Soc.Am.A*, 15:1121–1131, 1998.

[40] R. Dusséaux, C. Faure, J. Chandezon, and F. Molinet. New pertubation theory of diffraction gratings and its application to the study of ghosts. *J.Opt.Soc.Am.A*, 12:1271–1282, 1995.

[41] R. Dusséaux. Model with two roughness levels for diffraction gratings: the generalized Rayleigh expansion. *J.Opt.Soc.Am.A*, 15:2684–2697, 1998.

[42] E. Popov and M. Nevière. Surface-enhanced second harmonic generation in nonlinear corrugated dielectrics: new theorical approaches. *J.Opt.Soc.Am.B*, 11:1555–1564, 1994.

[43] J.B. Harris, T.W. Preist, and J.R. Sambles. Differential formalism for mulitlayer diffraction gratings made with uniaxial meterials. *J.Opt.Am.A*, 12:1965–1973, 1995.

[44] J.B. Harris, T.W. Preist, and J.R. Sambles. Conical diffraction from multicoated gratings containing uniaxial materials. *J.Opt. So. Am.A*, 12:1965–1973, 1995.

[45] M.E. Inchaussandague and R.A. Depine. Polarization conversion from diffraction gratings made of uniaxial crystals. *Phys.Rev.E*, 24:2899–2911, 1996.

[46] M.E. Inchaussandague and R.A. Depine. Rigorous vector theory for diffraction from gratings made of biaxial crystals. *J.Mod.Opt*, 44:1–27, 1997.

[47] G. Granet, J.Chandezon, and O. Coudert. Extension of the C-method to nonhomogeneous media: applications to nonhomogeneous layers with parallel modulated faces and to inclined lamellar gratings. *J.Opt.Soc.Am.A*, 14:1576–1582, 1997.

[48] L. Li and J. Chandezon. Improvement of the coordinate transformation method for surface-relief gratings with sharp edges. *J. Opt. So. Am. A*, 13(11):2247–2255, Nov 1996.

[49] J.P. Plumey, B. Guizal, and J.Chandezon. Coordinate transformation method as applied to asymmetric gratings with vertical facets. *J.Opt.Soc.Am.A*, 14:610–617, 1997.

[50] G. Granet and J.Chandezon. La méthode des coordonnées curvilignes appliquée à la diffraction par des réseaux dont le profil est donné par des équations paramétriques: application à la diffraction par un réseau cycloidal. *Pure App.Opt*, 6:727–740, 1997.

[51] R. Dusséaux, P.Cornet, and P.Chambelin. Etude de transformateur plan-E dans un système de coordonnées non orthogonales. *Ann. Telecommun*, 5-6:311–323, 1999.

[52] R. Dusséaux, P. Chambelin, and C. Faure. Analysis of rectangular waveguide H-plane junctions in nonorthogonal coordinate system. *Progress In Electromagnetics Research*, 28:205–229, 2000.

[53] R. Dusséaux and C. Faure. Analyse de composants plan-E symétriques en guides d'onde à section rectangulaire. *Ann. Telecommun*, 9-10:834–855, 1999.

[54] R. Dusséaux and C. Faure. Telegraphist's equations for rectangular waveguides and analysis in nonorthogonal coordinates. *Prog. Electromagn. Res. PIER*, 88:53–71, 2008.

[55] A. Benali, J. Chandezon, and J. Fontaine. A new theory for scattering of electromagentic waves from conducting or dielectric rough surfaces. *IEEE Trans. Antennas Propagat.*, 40:141–148, 1992.

[56] R. Dusséaux and C. Baudier. Scattering of a plane wave by one-dimensional dielectric rough surfaces-study of the field in a nonorthogonal coordinate system. *PIER*, 37:289–317, 2002.

[57] C Baudier, R Dusséaux, K S Edee, and G Granet. Scattering of a plane wave by one-dimensional dielectric random rough surfaces - study with the curvilinear coordinate method. *Waves Random Media*, 14:61–74, 2004.

[58] K.A. Braham, R.Dusseaux, and G. Granet. Scattering of electromagnetic waves from two-dimensional perfectly conducting random rough surfaces - study with the curvilinear coordinate method. *Wave Random Complex Media*, 18(2):255–274, 2008.

[59] R. Dusséaux, K.A. Braham, and G.Granet. Implementation and validation of the curvilinear coordinate method for the scattering of electromagnetic waves from two-dimensional dielectric random rough surfaces. *Wave Random Complex Media*, 18(4):551–570, 2008.

[60] K. A. Braham and R. Dusséaux. The curvilinear coordinate method associated with the short-coupling-range approximation for the study of scattering from one-dimensional random rough surfaces. *Optcs Comm.*, 281:5504–5510, 2008.

[61] R. Dusséaux, E.Vannier, O.Taconet, and G.Granet. Study of backscatter signature for seedbed surface evolution under rainfall - influence of radar precision. *Progress in Electromagnetics Research*, pages 415–437, 2012.

[62] K.Edee, B. Guizal, G. Granet, and A. Moreau. Beam implementation in a nonorthogonal coordinate system: application to the scattering from random rough surfaces. *J.Opt.Soc.Am.A*, 25:796–804, 2008.

[63] R. Dusséaux, K.A.Braham, and N.Emad. Eigenvalue system for the scattering from rough surfaces- saving in computation time by a physical approach. *Optics Comm*, 282:3820–3826, 2009.

[64] D. Prémel. Computation of a quasi-static field induced by two long straight parallel wires in a conductor with a rough surface. *J.Phys.D:Appl.Phys.*, 41:1–12, 2008.

[65] D. Prémel. Generalization of the second order vector potential formulation for arbitrary non-orthogonal curvilinear coordinates systems from the covariant form of Maxwell's equations. *Journal of Electromagnetic Analysis and Applications*, pages 400–409, 2012.

[66] Z.Bai, J.W.Demmel, J.J.Dongarra, A.Ruhe, and H. van der Vorst. Templates for the solution of algebraic eigenvalue problems. *Software, Environments, and Tools. SIAM*, 2000.

[67] Karen Braman, Ralph Byers, and Roy Mathias. The multishift QR algorithm. part I: Maintaining well-focused shifts and level 3 performance. *SIAM J. Matrix Anal. Appl*, 23:929–947, 2002.

[68] Robert Granat, Bo Kågström, and Daniel Kressner. A novel parallel QR algorithm for hybrid distributed memory HPC systems. *SIAM J. Sci. Comput.*, 32(4):2345–2378, Aug 2010.

[69] Karen Braman, Ralph Byers, and Roy Mathias. The multishift QR algorithm. part II: Aggressive early deflation, 2002.

[70] http://www.netlib.org/blas/.

[71] T. Sakuria and H.Sugiura. A projection method for generalized eigenvalue problem using numerical integration. *J.Comp.Appl.Math.*, 159:119–128, 2003.

[72] S. A. Shahzadeh Fazeli, N. Emad, and Z. Liu. A key to choose subspace size in implicitly restarted arnoldi method. *Numerical Algorithms, Springer US*, pages 1–20, 2015.

[73] C. Pan, R. Dusséaux, M. Fall, and N. Emad. Curvilinear coordinate method as an initial value problem: application to gratings. *Journal of the Optical Society of America A*, 32:143–149, 2015.

[74] C. Pan, R. Dusséaux, M. Fall, and N. Emad. The c-method as an initial value problem: application to multilayer gratings. *PIERS*, pages 1297–1301, 2015.

[75] F. Caire, D. Prémel, and G. Granet. Semi-analytical computation of a quasi-static field induced by a 3D eddy current probe scanning a 2D layered conductor with parallel rough interfaces. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, pages 600–613, 2014.

[76] M. Charbit. Eléments de théorie du signal: les signaux aléatoires. *Ellipses: Paris*, 1990.

[77] T.M. Elfouhaily and C.A. Guérin. A critical survey of approximate scattering wave theories from random rough surfaces. *Waves Random Media*, 14:R1–10, 2004.

[78] Z.Bai and J.W.Demmel. On a block implementation of hessenberg multishift qr iteration. *International Journal of High Speed Computing*, 1:97 – 112, 1989.

[79] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3^{rd} Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[80] http://www.netlib.org/scalapack/.

[81] Mark R. Fahey. A parallel eigenvalue routine for complex Hessenberg matrices. *ACM Trans. Math. Softw.*, 29:326–336, 2003.

[82] http://netlib.org/scalapack/slug/node75.html.

[83] G. Berginc. Small-slope approximation method: a further study of vector wave scattering from two-dimensional surfaces and comparison with experimental data. *Progress in electromagnetic research*, 37:251–287, 2002.

[84] J.T. Johnson, Leung Tsang, R.T. Shin, K.Pak, C.H. Chan, A. Ishimaru, and Y. Kuga. Backscattering enhancement of electromagnetic waves from two-dimensional perfectly conducting random rough surfaces: a comparison of Monte Carlo simulations with experimental data. *Antennas and Propagation, IEEE Transactions on*, 44(5):748–, May 1996.

[85] P. Phu, A. Ishimaru, and Y. Kuga. Copolarized and cross-polarized enhanced backscattering from two-dimensional very rough surfaces at millimeter wave frequencies. *Radio Sci.*, 29(5):1275–1291, 1994.

[86] James W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1997.

[87] G. Henry and R. Geijn. Parallelizing the qr algorithm for the unsymmetric algebraic eigenvalue problem: Myths and reality. *SIAM J. Sci. Comput*, 17:870883, 1996.

[88] Grey Ballard, James Demmel, and Ioana Dumitriu. Minimizing communication for eigenproblems and the singular value decomposition. 2010.

[89] James Demmel, Ioana Dumitriu, and Olga Holtz. Fast linear algebra is stable. *Numerische Mathematik*, 108(1):59–91, 2007.

[90] D. C. Sorensen. Implicitly restarted Arnoldi/Lanczos methods for large scale eigenvalue calculations. *Institute for Computer Applications in Science and Engineering (ICASE)*, 1996.

[91] Y. Saad. Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices. *Linear Algebra Applications*, 34:269–295, 1980.

[92] T.Sakurai and H.Tadano. CIRR: a Rayleigh-Ritz type method with contour integral generalized eigenvalue problems. *Hokkaido Math.J.*, 36:745–757, 2007.

[93] K.Senzaki, H. Tadano, and T. Sakurai. An estimation method of eigenvalues distribution with substructuring. *Proc. Annual meeting of Japan SIAM*, pages 132–133, 2006.

[94] S. Afifi and R. Dusséaux. Statistical study of radiation loss from planar optical waveguides: The curvilinear coordinate method and the small pertubation method. *J. Opt. Soc. Am. A*, 27:1171–1184, 2010.

[95] H. Raether. Surface plasmons on smooth and rough surfaces and on gratings. *Springer-Verlag, Berlin*, 1988.

[96] E. Popov, L. Mashev, and D. Maystre. Theoretical study of the anomalies of coated dielectric gratings. *Optica Acta*, 33:607–619, 1986.