Thèse n° 10 310

On design strategies for binary dielectric metasurfaces based on the Fourier modal method

Présentée le 9 octobre 2020

à la Faculté des sciences et techniques de l'ingénieur
Laboratoire de nanophotonique et métrologie
Programme doctoral en photonique

pour l'obtention du grade de Docteur ès Sciences

par

Kévin MÜLLER

Acceptée sur proposition du jury

Prof. C. Moser, président du jury
Dr T. Scharf, Dr W. Noell, directeurs de thèse
Prof. C. Rockstuhl, rapporteur
Prof. T. Pertsch, rapporteur
Prof. H. Altug, rapporteuse

École
polytechnique
fédérale
de Lausanne

2020

Acknowledgements

First of all, I would like to thank my thesis supervisor, Toralf Scharf, who supported me during those four years of my PhD. I have always felt that he was interested in what I was doing, and I'm very grateful to him for that. In addition, his support over the past few months has been priceless.

A thesis is hard work, but it is a unique opportunity to gain in-depth knowledge on a subject and to get to know each other as well. Ken Weible, Reinhard Völkel, Toralf Scharf and Wilfried Noell gave me the opportunity to do a thesis and I am very grateful to them.

I spent the first seven months of my PhD at the Karlsruhe Institute of Technology in the group of Carsten Rockstuhl. I learned a lot and really enjoyed my stay there, thanks to him and his great group.

It was also in Karlsruhe that I started working on the Poynting operation. The idea came from a discussion that I had with Fernando Negredo and, with the help of Carsten Rockstuhl and Ivan Fernandez-Corbaton, I developed this operation. The resulting paper went through multiple rounds of review, and I would like to thank the reviewers, along with Carsten, Fernando and Ivan, who gave me constructive comments and helped me to improve my work.

For the next 18 months, I spent my PhD at SUSS MicroOptics SA and worked in collaboration with Yan Chen, Raoul Kirner and Wilfried Noell. Yan Chen asked to me to give my opinion on several interesting papers and to solve problems, and it was always a pleasure to help. It made me feel involved in something that I can make it better, giving me extra motivation to pursue my work.

For the last part of my PhD, I worked in the group of Olivier Martin at EPFL. I would like to thank him and his group for all the meaningful discussions and supports.

I was involved in the fabrication of two metasurfaces, an anti-reflective metasurface and a resonant metasurface, but I was not able to get satisfactory results in time. Many people helped me in those projects and I would like to express my gratitude to: Andreas Vetter for making a sample of the anti-reflective metasurface, Toralf for mounting the optical setup for the measurements, Philippe Wyss for the deposition of hydrogenated amorphous silicon and helping me in the measurements of the layer thickness and optical properties, Karim Achouri

for the e-beam exposure and SEM images, and for the many discussions on the fabrication process, Zdenek Benes for helping me to improve the e-beam writing strategy, Debdatta Ray for helping me in the clean room and the CMI staff for training and advice.

This thesis is part of the NOLOSS project, which includes 15 PhD students. We have had several workshops during the NOLOSS project, and I would like to thank my fellow PhD students from NOLOSS for the constructive discussions and the good time spent during those workshops. I would also like to thank Toralf again, who is the coordinator, for managing this project, and the European Union for the funding.

I would like to thank all person who have made my life outside of the PhD enjoyable and interesting. It includes my flatmates, my friends and the people from the Kung-Fu Zentrum in Karlsruhe, the Wong Shaolin Kung-Fu club in Neuchâtel and the Discofox course from K'danse in Lausanne. I would also like to express my gratitude to my parents and my brother, now sister, for their support.

Funding:
This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 675745.

Lausanne, August 21, 2020 Kevin Müller

Abstract

Binary dielectric metasurfaces are arrays of sub-wavelength structures that act as a thin layer of artificial material. They are generally lossless and relatively simple to fabricate since only a single structuring step is required. By carefully designing the metasurface, the phase, amplitude and the polarization of the incident light can be controlled at will. In practice, fabrication constraints and the limited choice of materials reduce what can be done with metasurfaces. But a wide range of functionalities can be implemented with the proper design techniques and knowledge. This thesis contributes to both.

The modes are key to understand the phenomena occurring inside a metasurface. To facilitate the analysis of the modes, the Poynting operation, which is related to the Poynting vector, is introduced. We describe how this operation can be used to reformulate the boundary condition in order to estimate the reflection and transmission coefficients with reduced knowledge on the modes involved, and to orthonormalized the modes.

The Fourier modal method, which is the method used for the rigorous simulation of metasurfaces in this work, is improved in order to facilitate the access to valuable information that can be used to better understand the phenomena occurring inside a metasurface, and to speed up the design and optimization process. This method computes the eigen-modes present in the metasurface. To better analyze them, the eigen-modes are orthonormalized using the Poynting operation. We show that most of the modes can be filter out in order to simulate a metasurface with different thicknesses in milliseconds.

From the analysis of the modes propagating in metasurfaces, two types of metasurface are identified: single-mode metasurfaces and multi-mode metasurfaces. For single-mode metasurfaces, we provide design techniques that translates the desired response into internal properties of the metasurfaces. For multi-mode metasurfaces, the concept of self-coupling mode is developed. We show that, based on this concept, the angular and spectral response of a metasurface can be interpolated safely with a few simulations even if high-Q resonances are present. For both types of metasurfaces, examples of design are provided.

Gradient-based optimization methods allow to obtain the optimal metasurface in a few iterations, but the derivative of the merit function is necessary. The adjoint method computes the functional derivative of the merit function with respect to the permettivity and permeability. We provide the equations of the adjoint method and apply them to diffractive optical elements such that they can be used in conjunction with the Fourier modal method.

This thesis contributes to design challenges for complex electromagnetic problems,and it does not only provides concepts, but also the tools to put them into operation.

Keywords: Dielectric metasurface, modal analysis, Poynting vector, Fourier modal method, metasurface design, hologram, adjoint method, diffractive optical element, resonance analysis, resonant metasurface.

Résumé

Les métasurfaces diélectriques binaires sont des réseaux de structures plus petites que la longueur d'onde qui agissent comme une fine couche de matériau artificiel. Elles sont généralement sans perte et relativement simples à fabriquer puisqu'une seule étape de structuration est nécessaire. En concevant soigneusement la métasurface, la phase, l'amplitude et la polarisation de la lumière incidente peuvent être contrôlées à volonté. En pratique, les contraintes de fabrication et le choix limité de matériaux réduisent ce qui peut être fait avec les métasurfaces. Mais un large éventail de fonctionnalités peut être mis en oeuvre avec les techniques de conception et les connaissances appropriées. Cette thèse contribue à ces deux aspects.

Les modes sont essentiels pour comprendre les phénomènes qui se produisent à l'intérieur d'une métasurface. Pour faciliter l'analyse des modes, l'opération de Poynting, qui est liée au vecteur de Poynting, est introduite. Nous décrivons comment cette opération peut être utilisée pour reformuler la condition aux interfaces afin d'estimer les coefficients de réflexion et de transmission avec une connaissance réduite des modes impliqués, et pour orthonormaliser les modes.

La méthode modale de Fourier, qui est la méthode utilisée pour la simulation rigoureuse des métasurfaces dans cette thèse, est améliorée afin de faciliter l'accès à des informations précieuses qui peuvent être utilisées pour mieux comprendre les phénomènes se produisant à l'intérieur d'une métasurface, et pour accélérer le processus de conception et d'optimisation. Cette méthode permet de calculer les modes propres présents dans la métasurface. Pour mieux les analyser, les modes propres sont orthonormalisés à l'aide de l'opération de Poynting. Nous montrons que la plupart des modes peuvent être filtrés afin de simuler une métasurface avec des épaisseurs différentes en quelques millisecondes.

A partir de l'analyse des modes se propageant dans les métasurfaces, deux types de métasurfaces sont identifiés : les métasurfaces monomodes et les métasurfaces multimodes. Pour les métasurfaces monomodes, nous fournissons des techniques de conception qui traduisent la réponse souhaitée en propriétés internes des métasurfaces. Pour les métasurfaces multimodes, le concept de mode d'auto-couplage est développé. Nous montrons que, sur la base de ce concept, la réponse angulaire et spectrale d'une métasurface peut être interpolée en toute sécurité avec quelques simulations, même si des résonances à haut facteur de qualité sont présentes. Pour les deux types de métasurfaces, des exemples de conception sont fournis.

Les méthodes d'optimisation basées sur le gradient permettent d'obtenir la métasurface optimale en quelques itérations, mais la dérivée de la fonction de mérite est nécessaire. La méthode adjointe calcule la dérivée fonctionnelle de la fonction de mérite en fonction de la

perméabilité et de la perméabilité. Nous fournissons les équations de la méthode adjointe et les appliquons aux éléments optiques diffractifs de manière à ce qu'ils puissent être utilisés en conjonction avec la méthode modale de Fourier.
Cette thèse contribue aux défis de conception pour des problèmes électromagnétiques complexes, et elle ne fournit pas seulement des concepts, mais aussi les outils pour les mettre en oeuvre.

Mots-clés : Métasurface diélectrique, analyse modale, vecteur de Poynting, méthode modale de Fourier, conception de métasurface, hologramme, méthode adjointe, élément optique diffractif, analyse de résonance, métasurface résonante.

Contents

Acknowledgements **i**

Abstract (English/Français) **iii**

1 Introduction **1**

- 1.1 Introduction to the design process for diffractive optical elements 5
- 1.2 Structure of the thesis 10
- 1.3 Notation and convention 11

2 Poynting operation **13**

- 2.1 Introduction 13
- 2.2 Discussion on operations 15
- 2.3 Poynting operation 16
- 2.4 Orthogonality relation 19
- 2.5 Reformulation of the boundary condition 21
- 2.6 Operations on the Gram matrix 23
- 2.7 Reformulation of the boundary condition: Applications 25
  - 2.7.1 Estimation of the coupling efficiencies at the surface of a metamaterial . 25
  - 2.7.2 Fresnel coefficients for uniaxial media 30
- 2.8 Operations on the Gram matrix: Applications 33
  - 2.8.1 Orthogonality of plane waves in lossy homogenous isotropic media . . . 33
  - 2.8.2 Orthonormalization of propagating and evanescent mode 35
  - 2.8.3 Orthonormalization of complex modes 38
  - 2.8.4 Rotation of propagating, evanescent and complex modes 42
- 2.9 Conclusion 45
- 2.10 Proofs 45
  - 2.10.1 Proof of the orthogonality relation 45
  - 2.10.2 Proof of the reformulation of the boundary condition 48
  - 2.10.3 Proof of the operations on the Gram matrix 51

3 Improved Fourier modal method **57**

- 3.1 Introduction 57
  - 3.1.1 Overview of the improved Fourier modal method 58
  - 3.1.2 Representation of a multi-layer structure and design strategy 62

- 3.2 Computation of the eigen-modes 64
  - 3.2.1 Convention for plane waves 69
- 3.3 Computation of the coupling matrices 71
  - 3.3.1 Coupling matrices at an interface between two homogeneous media . . 74
- 3.4 Layer reduction 74
- 3.5 Mode filtering and contribution of the eigen-modes to the power flow 76
- 3.6 Conclusion 82
- 3.7 Proofs 82
  - 3.7.1 Proof of the coupling matrices at an interface between two heterogeneous media 82
  - 3.7.2 Proof of the layer reduction 83

4 Single-mode metasurface 87

- 4.1 Introduction 87
- 4.2 Design of ideal single-mode metasurfaces 89
  - 4.2.1 Comparison between ideal and real single-mode metasurfaces 93
  - 4.2.2 Design of metasurface-based hologram 98
- 4.3 Design of a pair of ideal single-mode metasurfaces 105
  - 4.3.1 Design of a polarization rotator and a pseudo-quarter-wave plate 107
- 4.4 Design of anti-reflective metasurfaces 109
  - 4.4.1 Design of glass and silicon anti-reflective metasurfaces 114
- 4.5 Design of a half-wave plate 120
- 4.6 Conclusion 122
- 4.7 Proofs 123
  - 4.7.1 Proof of the symmetric Jones matrix from a transformation of polarization state 123
  - 4.7.2 Proof of the parameters of an ideal single-mode metasurfaces from its Jones matrix 128
  - 4.7.3 Proof of the design of a pair of ideal single-mode metasurfaces 129

5 Multi-mode metasurface, resonant metasurface and self-coupling mode 135

- 5.1 Introduction 135
- 5.2 Self-coupling mode 139
  - 5.2.1 Definition and computation of the self-coupling modes 139
  - 5.2.2 Contribution of the self-coupling modes on the response of a metasurface 141
- 5.3 Application of the self-coupling mode on multi-mode metasurfaces 143
  - 5.3.1 Magnetic and electric dipole resonances in Huygens' metasurface 143
  - 5.3.2 Narrowband metasurface and interpolation 147
  - 5.3.3 GaAlAs metasurface for laser application and computation of the angular spectrum of a resonance for a symmetric metasurface 151
  - 5.3.4 Design of high-Q metasurface for sensing 159
- 5.4 Conclusion 168

6 Adjoint method **171**
6.1 Introduction 171
6.2 Plane waves generated by a plane source 172
6.3 Variation of the fields due to a change in material parameters and reciprocity . 175
6.4 Adjoint method for periodic diffractive optical elements 177
6.5 Application of the adjoint method for the design of a beam splitter 180
6.6 Conclusion 185
6.7 Proofs 186
  6.7.1 Proof of the plane waves generated by a plane source 186
  6.7.2 Derivation of the adjoint method for periodic diffractive optical elements 188

7 Conclusion **195**

Bibliography **199**

Curriculum Vitae **215**

1 Introduction

The ability to control light has always been associated with technological advances. The first optical elements that were designed to manipulate light are mirrors and lenses, and they allowed us to see and understand what is too small or too far away for our naked eyes. Those elements are based on the concept of rays. The first great textbooks on this concept date back to the Ancient Greece with the work of Euclid [1] and, a thousand years later, to the Islamic Golden Age with the work of Ibn al-Haytham [2]. However, it is during the Renaissance, at the beginning of the 17th century, that Galileo Galilei demonstrated the first high-performance telescope and microscope. Since then, lens systems have been continuously improved, leading to the current cameras, microscopes, telescopes up to the lithography lens systems [3,4].

Also in the 17th century, the model of light as waves emerges, mainly from C. Huygens with the Huygens' principle according to which each point of a wavefront is a spherical source. It was not until the very beginning of the 19th century that the wave theory of light became popular in the scientific community thanks to the double slits experiment of T. Young. From the work of many scientists over the following decades, J. C. Maxwell were able to propose an unifying theory of electromagnetism through the Maxwell equation [5], which is the core of the rigorous simulation methods used today.

Since light is an electromagnetic wave, it is possible to bend light by varying its phase and amplitude on a plane, which can be done using a diffractive optical element. The first diffractive optical element was made at the end of the 18th century by D. Rittenhouse [6,7], but it is 150 years later that D. Gabor was able to fabricate a hologram [8] by recording an interference pattern on a photographic plate. It is with the invention of the laser and the advance in micro-fabrication and materials that the first computer generated hologram was made [9]. Computer generated hologram allows to control the phase and the amplitude of the incident light while being extremely thin in comparison to a lens. The first computer generated hologram was amplitude-only, but, nowadays, they are usually phase-only because of the high transmission efficiency. Computer generated hologram are a type of diffractive optical elements and they are typically used for the external cavity of lasers [7], structured light profilometry [10] and security features [11].

The diffractive optical elements mentioned earlier are composed of features larger the wavelength and they mainly affect the amplitude and phase of an incident beam. In order to control the polarization, a possibility is to use uniaxial material whose extraordinary axis can be oriented differently at different points in space, but such optical component is not practical because, at our knowledge, it is not possible to fabricate a polycrystalline material where the orientation of the crystals can be controlled independently. A more interesting solution is to properly designed structures called meta-atoms, leading to an artificial material also known as metamaterial. The condition is that the dimensions of the structures and the distance between the structures are much smaller than the wavelength in order to avoid scattering [12]. Due to this constraint, the first metamaterial was done for microwaves application in 1948 by W. Kock [13]. In the microwave range, metals are excellent candidate because they are lossless and interact strongly with light. Metamaterials became more popular in the scientific community 40 years later with the work of J. B. Pendry and D. R. Smith [14–16]. The notion of metasurface has been introduced in the microwaves community a few years later [17].

The wavelength of microwaves is between 1 mm and 1 m, meaning that structures in the range of 100 μm can be used as the building blocks for metamaterials and metasurfaces since they are deep subwavelength. Due to the interesting properties of metasurfaces, the idea of metasurface has been brought to the visible and near-infrared domain. The first metallic metasurfaces are perforated metallic layers, such as the metallic-dielectric-metallic structures in [18] that behaves as material with a negative refractive index, the metalens composed of a hole array [19] and another metalens composed of v-shaped structures [20], which was inspired by [21]. The disadvantages that metallic metasurface in the microwave regime does not have is that the structures have to be simple due to the small size of the structures and the resulting difficulties during fabrication, and metal absorbs light at those wavelength. Moreover, metallic usually works on resonance, meaning that the absorption is amplified. An interesting metallic metasurface which is not based on resonances is given in [22], where the transmission efficiency is above 85% on average.

To avoid absorption, dielectric metasurfaces have been proposed. However, since dielectrics have a weaker effect on the propagation of light, dielectric metasurfaces need to be thicker, making the distinction between dielectric metasurface, metamaterial and gratings unsharp. As with metallic metasurfaces, the structure that compose the dielectric metasurfaces have to be as simple as possible in order to be able to fabricate them properly. Therefore, most dielectric metasurfaces are binary, requiring a single etching step for their fabrication.

One of the early dielectric metasurfaces, named zero-order grating at that time, has been designed for the mid-infrared regime [23] and was composed of lines and spaces. A few years later, the first dielectric metasurfaces composed of an array of cylinders have been fabricated [24], followed by a metasurface based on the Pancharatnam-Berry phase [25]. Metasurfaces composed of cylinders [26–30] (fig. 1.1a) behaves as usual diffractive optical elements except that the accumulated phase is continuous even if the metasurface is binary. Metasurfaces based on the Pancharatnam-Berry phase [31–35] (fig. 1.1b) are composed of ellipses or rectangles and the accumulated phase depends on their orientation. They are polarization dependent and have the particularity to change the direction of rotation of circular polarized light. Typical applications for those kind of metasurfaces are beam shaping [30], beam deflection [28], phase-only holograms [32] and light focusing [26, 27, 29, 33, 34].

Metasurfaces affect both the phase and the polarization [38, 39]. Therefore, it is possible to design a metasurface such that it generates two different holograms depending on the polarization of the illumination [36, 40] (fig. 1.1c), or spatially separates an incident beam into two polarized beams [41, 42].

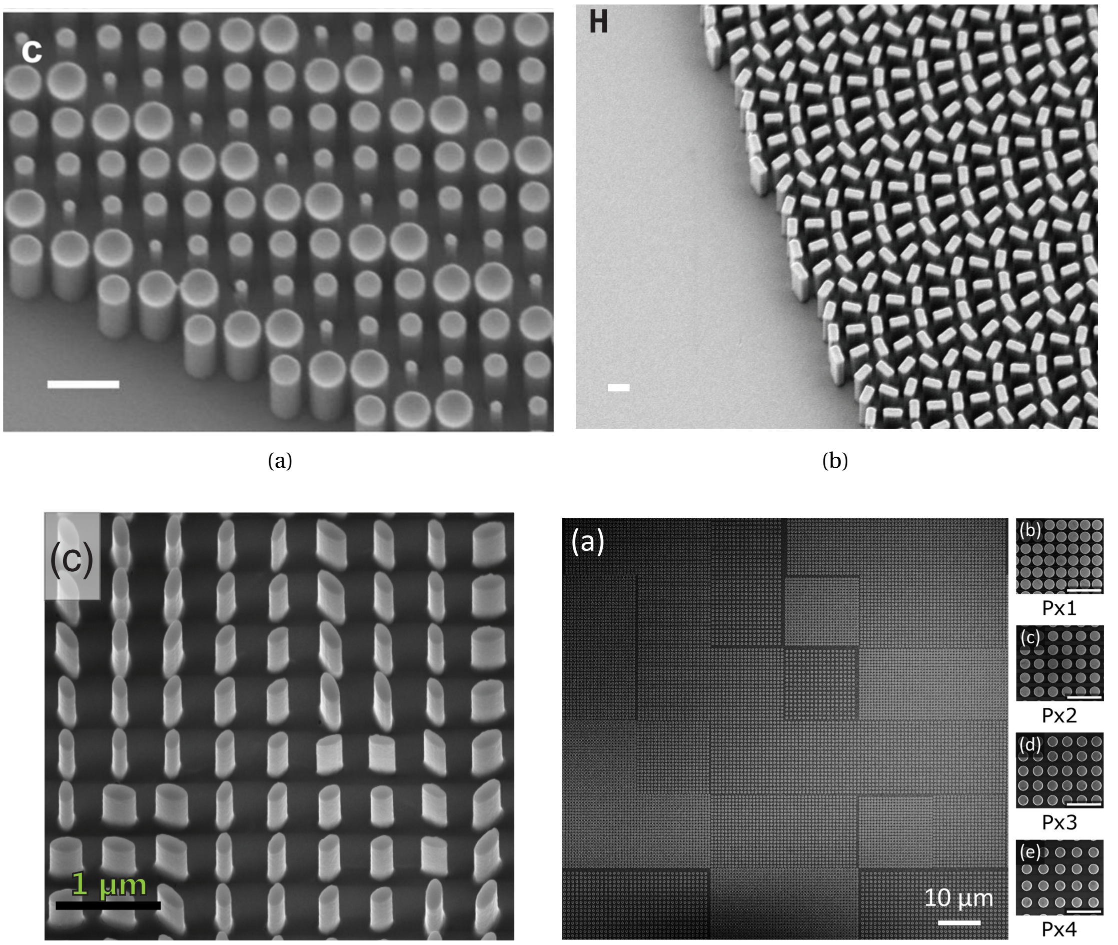
The dielectric metasurfaces mentioned earlier can be seen as an array of waveguides and, by changing the dimensions of those waveguides, different phase accumulations can be obtained. The phase can also be controlled using Huygens' metasurfaces, which are dielectric metasurfaces with two overlapping resonances [43]. Hence, Huygens metasurfaces are used to generate holograms [37, 44, 45] (fig. 1.1d). Resonant dielectric metasurfaces have many other applications. They can be used as color filter [46–50], polarization filter [51], generalized Hartmann-Shack array [52], molecule sensing [53] and light emitter [54]. Moreover, due to the large fields inside resonant metasurfaces, otherwise negligible effects such as second harmonic generation [55–57], Kerr nonlinearities [58] and Faraday rotation [59] can be greatly enhanced, and, by using graphene, tunable resonant metasurfaces can be made [60].

Metasurfaces are complex structures with a high diversity of responses, and they requires proper design techniques and strategies in order to unlock their full potential. The work presented here goes into that direction by providing a new set of techniques that facilitates the design of binary dielectric metasurfaces and allows a more systematic analysis of the metasurface response.

The simulation method used in this work is the Fourier modal method [61] and it has been improved in order to greatly facilitate the design of metasurfaces (chapter 3). This method is particularly well suited for the simulation of binary metasurfaces and, at the same time, provides valuable information that can be used in the design process. The most important information that cannot be obtained through other methods such as the Finite Difference Time Domain method (FDTD) and the Finite Element Method (FEM) is the number of propagating modes inside the metasurface.

Metasurfaces with a single propagating mode per polarization are called here single-mode metasurfaces. They are broad-band, have a high transmission efficiency and can be used to control both phase and polarization. Moreover, they can be described using simple models, allowing to predict the functionalities they can perform and their limitations. Those models are given in chapter 4 and they are used to design different type of holograms, anti-reflective metasurfaces and metasurfaces that act as waveplates.

When more propagating modes are present, the metasurface is called a multi-mode metasurface. Due to intereference between the modes [62], multi-mode metasurfaces have more



(a) (b)

(c) (d)

Figure 1.1 – a) SEM image of the metalens proposed in [27], which is composed of TiO_2 cylinders on a glass substrate. The length of the scale bar is 600 nm. This figure is from [27]. b) SEM image of the metasurface based on the Pancharatnam-Berry proposed in [33], which is composed of TiO_2 nanofins on a glass substrate. The length of the scale bar is 300 nm. This figure is from [33]. c) SEM image of a metasurface that generates two holograms depending of the polarization of the illumination. The metasurface has been proposed in [36] and is composed of TiO_2 cylinders with varying dimensions and orientation. This figure is from [36]. d) SEM image of the Huygens' metasurface proposed in [37], which is composed of silicon cylinders embedded in glass. This figure is from [37].

complex behaviors and high-Q resonances can occur. The standard technique to detect a resonance is to look at sharp features in the metasurface spectral response, but it leads to several issues. First, the metasurface has to be simulated at multiple wavelengths in order to know if there is a resonance. Second, it is possible to miss a high-Q resonance if the sampling is too coarse. The concept of self-coupling mode developed in chapter 5 solves those issues.

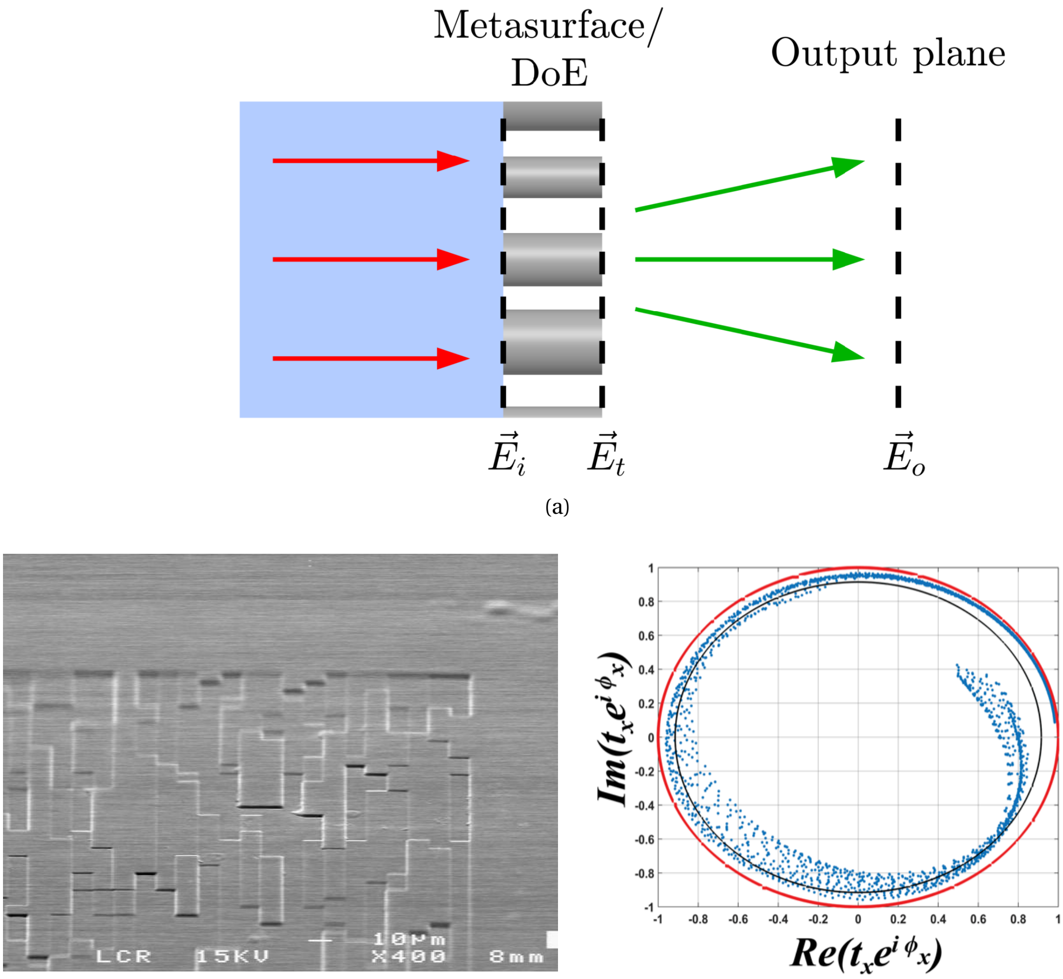
The different design techniques proposed in this work use assumptions and constraints in order to obtain a metasurface close to the optimal one with minimal computational effort. Gradient-based optimization methods are ideal to get the optimal metasurface in a reasonable amount of time because only a few iterations are required to reach a local optimum. The condition is obviously that the gradient of the merit function needs to be computed, and the adjoint method can do this with only two simulations [63]. Using the Fourier modal method, the number of simulations reduces to one at normal incidence. The implementation of the adjoint method depends of the simulation method and the adjoint method is provided for the Fourier modal method in chapter 6.

1.1 Introduction to the design process for diffractive optical elements

In this section, several design processes for diffractive optical elements (DoE) are presented based on what has been done the literature. The term DoE includes here any diffractive optical element made of dielectric material, but the focus is on dielectric binary metasurfaces. When working in transmission, the system to optimize can be represented as shown in fig. 1.2a, where the incident field just before the DoE, denoted $\vec{E}_i$, is known and the objective is to design the DoE such that the field at the output plane, denoted $\vec{E}_o$, fulfills some specifications. For a given DoE, the simulation method to get the transmitted field just after the DoE, denoted $\vec{E}_t$, from the field $\vec{E}_i$ is different than the simulation method to get the field $\vec{E}_o$ from the field $\vec{E}_t$ since it is a propagation in free-space. The fields $\vec{E}_i$, $\vec{E}_t$ and $\vec{E}_o$ depend on position and wavelength, but, for many applications, the wavelength is fixed.

Before starting the optimization, several choices have to be made, namely the merit function, the type of DoE and the simulation methods. The choice of the merit function is critical because it determines the DoE that will be obtained after the optimization. If a gradient-based optimization method is used in the design process, the merit function has to be differentiable, or, at least, piece-wise differentiable.

The simplest and most common merit function is the mean square of the difference between the obtained values, such as the intensity at the output plane, and the desired ones. An example is shown in section 6.5 (equation (6.23)). However, there are usually other constraints, typically fabrication constraints, and multiple quantities to optimize that need to be included in the merit function. It can be done by using a min-max multi-objective formulation [66] or by constructing a merit function which is a linear composition of the multiple merit functions related to the different quantities that need to be optimized. An example of such merit function is in section 5.3.4 (equation (5.25)). The constraints can also be incorporated into the



(a)

(b)

(c)

Figure 1.2 – a) Schema representing a Diffractive optical Element (DoE), which can be a metasurface, with three different planes where the field is computed during the design process. The red arrows represents the incident light and the green arrows represents the light propagating from the DoE to the output plane. $\vec{E}_i$ is the incident field just before the DoE, $\vec{E}_t$ is the transmitted field just after the DoE and $\vec{E}_o$ is the field at the output plane, which is used to compute the merit function. b) SEM image of the multi-level DoE made of polycarbonate proposed in [64]. The maximum etch depth is 810 nm and the wavelength of operation is 633 nm. c) Real and imaginary part of the responses in transmission of a set of cuboids made of TiO_2 on a glass substrate. The simulations have been done by J. B. Mueller and the figure is from [65] (fig. 8.12.2). Each blue dot is the response of a cuboid for different lateral dimensions. The height of the cuboid is fixed to 600 nm. The red circle corresponds to full transmission and the black circle is the average of the transmitted amplitudes. The incident field is polarized along one side of the cuboid.

optimization process with the use of constrained optimization methods [67].

The choice for the type of DoE depends of the desired functionality. If the DoE have to reflect significantly the light, the possible candidates are DoEs in front of a mirror [68, 69], structures composed of metals and multi-mode metasurfaces. As mentioned in the introduction, resonant metasurfaces are multi-mode metasurfaces, but it is also possible to have metasurfaces that act as broad-band mirrors [70].

If the DoE affects mainly the phase of the incident light, meaning that the transmission efficiency is high, multi-mode metasurfaces should be avoided with some exceptions such as the Huygens' metasurfaces [43]. In the case where the diffraction angle is small, meaning that the phase of the field after the DoE is smooth, conventional DoEs should be considered. Conventional DoEs have features larger than the wavelength, meaning that diffraction effects inside the DoE can be neglected. Hence, the phase delay is proportional to the height profile of the DoE. This approximation is called the thin element approximation [71]. An example of a conventional DoE is shown in fig. 1.2b. However, for large diffraction angle, the features size is in the order of the wavelength leading to diffraction effects inside the DoE that need to be taken into account. Therefore, rigorous simulation methods are required.

Another type of DoE that can be used to affects the phase of the incident light are single-mode metasurfaces, also called waveguide-type metasurfaces [72]. Single-mode metasurfaces are composed of cylinders with various cross-sections, and each of those cylinders acts as a waveguide, meaning that the light is confined mostly inside the cylinders as discussed in [72]. By using cylinders with elliptical or rectangular cross-section (figs. 1.1b and 1.1c), single-mode metasurfaces can control both the phase and the polarization. There is also a continuum from single-mode metasurface to conventional DoE, but those two extremes are more simple to simulate and to design as explained later.

Before starting to optimize the DoE, the simulation methods has to be chosen in order to get the field $\vec{E}_o$ from the field $\vec{E}_i$. As shown in fig. 1.2a, the system is divided into two regions: a heterogeneous medium, which corresponds to the DoE, and a homogeneous medium, where free space propagation takes place. Because those regions are fundamentally different, the simulation methods used are not the same. For the homogeneous medium, common methods are the Rayleight-Sommerfeld diffraction integral (section 2.1 of [73]) and the angular spectrum of plane wave (section 3.10 of [74]). In many applications, the output plane is in the far-field. In this case, the field $\vec{E}_o$ is simply the Fourier transform of the field $\vec{E}_t$. Finally, the propagation of light can be approximated by rays optics if the field $\vec{E}_t$ is a smooth function, meaning that diffraction effects during propagation are negligible. In order to get the propagation direction of the rays from the field $\vec{E}_t$, the generalized law of refraction is used [21].

For the heterogeneous medium, the thin element approximation should be used when this approximation is valid since getting the field $\vec{E}_t$ from the field $\vec{E}_i$ becomes then trivial. If diffraction effects inside the DoE are significant, a rigorous method such as the Finite-Difference Time-Domain method (FDTD) [75], the Finite Element Method (FEM) [76–78] and the Fourier Modal Method (FMM) [61, 79–81], also known as the Rigorous Coupled Wave Analysis (RCWA) [82], is used. However, rigorous methods are too computationally expensive for the simulation of large area, meaning that strategies are needed in order to simulate aperiodic DoE such as holograms and metalenses. One strategy given in [83, 84] is to divide the DoE into sub-area, simulate this sub-area assuming that it is repeated periodically [84] or by adding a perfectly matched layer [83], and get the field just after this sub-area. The field $\vec{E}_t$ is given by the combination of the results of all those partial simulations.

For single-mode metasurfaces, a similar strategy can be taken. Since the light is confined mostly inside the cylinders, the field after a single-mode metasurface at the location of a cylinder is weakly affected by the dimensions of the neighboring cylinders. Therefore, the field at that location is assumed to be the same as the field of the transmitted plane wave after a zeroth-order grating composed of a periodic arrangement of this cylinder. This approximation is called the locally periodic approximation [66]. By simulating multiple zeroth-order gratings, each of which is composed of a periodic arrangement of a cylinder present in the single-mode metasurface, the field $\vec{E}_t$ is obtained. As shown in [66], interpolation can be used to reduce the number of zeroth-order gratings being simulated. Most single-mode metasurfaces are simulated using this strategy [26–36, 40, 66].

Once the merit function, the DoE type and the simulation methods have been chosen, the optimized DoE can be found through different design techniques, three of which are described here. The first one is parameter sweep. It is typically used when there are a few parameters that are optimized, meaning that it is mainly used to design zeroth-order gratings. This technique also allows to explore the possible responses a DoE can have. For example, to design a resonant metasurface composed of cylinders, prior information can be used in order to choose roughly the dimensions of the cylinder such that a dipole resonance, or a higher order resonance [85], providing the desired functionality, occurs. Then, the response of a set of metasurfaces which is close to the resonance is computed and the optimized metasurface is found by interpolating the obtained responses. This design technique is also used for single-mode metasurface as explained later. In this work, the Fourier modal method proposed in chapter 3 facilitates such parameter sweep and, for resonant metasurfaces, the concept of self-coupling mode presented in chapter 5 allows to safely interpolate the response of the metasurface even if high-Q resonances are present.

The second design technique is known as gradient-based topology optimization. Since the merit function is computed with the DoE geometry as the input, it is possible to get the effect of a variation of this geometry on the merit function. Using this derivative, gradient-based optimization methods, such as gradient descent and quasi-Newton method [67], can be used leading to a very fast convergence to the optimal DoE even for a large number of parameters. Two methods can be used to compute this derivative: the algorithmic differentiation [86, 87] and the adjoint method [63, 88, 89]. The algorithmic differentiation is based on the fact that a sequence of operations applied on the inputs has been made in order to compute the merit function. Hence, the chain rule is applied to get the derivative of the merit function with respect to those inputs. The adjoint method allows to get the functional derivative of the merit function with respect to the permettivity and permeability by computing the field generated by the primary source and an adjoint source. A review on topology optimization applied on optical systems is given in [90]. In chapter 6, the equations required to apply the adjoint method when using the Fourier modal method, are provided.

The last design technique presented here is based on the design technique commonly used for conventional DoE, but it can also be used in the design of single-mode metasurfaces (sections 8.11 to 8.13 in [65]). For conventional DoE, the main assumption is that the DoE affects only the phase, meaning that the amplitude of the fields $\vec{E}_i$ and $\vec{E}_t$ are the same. Hence, the optimization of the conventional DoE is reduced to finding the phase of the field $\vec{E}_t$ that minimizes the merit function. This step is usually done with the Gerchberg-Saxton algorithm also known as the Iterative Fourier Transform Algorithm (IFTA) [91]. The height profile of the conventional DoE is directly obtained from the phase of the field $\vec{E}_t$ due to the thin element approximation.

To adapt this design technique to single-mode metasurfaces, two major differences from conventional DoEs have to be taken into account. First, there is no more a direct relationship between the phase delay and the geometry of the metasurface. Therefore, the relationship between the field $\vec{E}_i$ and $\vec{E}_t$ at the location of the cylinders that compose the metasurface has to be computed using a rigorous method. Since it is usually not feasible to simulate rigorously the whole DoE, the locally periodic approximation is used. Second, the cylinders which are not circularly symmetrical in cross-section affect the polarization of the incident light as a birefringent material (section 8.11 in [65]). Hence, the transfer function describing the relationship between the field $\vec{E}_i$ and $\vec{E}_t$ is a Jones matrix (section 6.1 in [92]), which is wavelength and position-dependant.

In order to find the Jones matrices that transform a given polarization state into another one with the desired phase delays, typically from 0 to 2π, different techniques can be used. Those techniques have to take into account that the Jones matrix describes the response of a cylinder behaving as birefringent materials, leading to constraints on the Jones matrix. The technique used in the literature is based on the Poincaré sphere [25,65], where the incident and transmitted polarization states are points located on the Poincaré sphere and the polarization state changes continuously while propagating inside the cylinder, making a trajectory on the Poincaré sphere. Due to the constraints on the Jones matrix, the trajectory has to follow a set of rules, which are given in section 8.11 in [65], and, from this trajectory, the Jones matrix can be found. We propose a different technique in section 4.2, where we provide a set of equations that gives directly all the possible Jones matrix that transform the incident polarization state into the desired transmitted polarization state. We also discuss in section 4.2.2 the best type of single-mode metasurface , which is related to a set of Jones matrices, for a given functionality.

Once the set of Jones matrices is obtained, the next step is to find the dimensions of the cylinders, such that their response is as close as possible to the desired Jones matrices. Two approaches to obtain those dimensions are proposed in the literature. The first one is to do a parameter sweep (section 8.12 in [65]), whose obtained responses can be represented as in fig. 1.2c. The second approach is to perform a gradient-based topology optimization [66, 93]. Then, the optimized single-metasurface can be found with a modified version of the Gerchberg-Saxton algorithm that takes into account the available Jones matrices (section 8.13 in [65]).

1.2 Structure of the thesis

Each chapter of this work introduces or develops techniques to design or analyze metasurfaces. Chapters are divided into three parts: theory, examples of application and proofs.

Chapter 2 introduces the Poynting operation, which is based on the Poynting vector. It is used to reformulate the boundary condition and greatly simplify the orthonormalization of the eigen-modes. Both applications are used in the following chapters, but an important advantage of the Poynting operation comes from the formalism itself. Because the properties of the Poynting operation are clearly stated, the power flow due to non-orthogonal modes, evanescent modes and complex modes can be obtained in an automatic and simple way.

Chapter 3 presents the Fourier modal method used in this thesis. It has been improved in order to facilitate the design of metasurfaces. Its main differences with current Fourier modal methods is that it is presented and implemented such that the simulation can be easily optimized depending on the information the designer is looking for. Moreover, the layers thickness can be fixed later in the simulation and, after the mode filtering described in section 3.5, the metasurface response for different layers thicknesses is computed in a few milliseconds. For example, if the designer wants to analyse in depth the central layer in a multi-layer structure or change its thickness, the layers before and after this central layer can be reduced into two interfaces, which are represented by a S-matrix, and change the thickness of the central layer afterward, leading to reduced memory usage and faster computation time. The equations provided in the literature [80, 81, 94, 95] does not allow such manipulation.

Chapter 4 provides design techniques for single-mode metasurfaces, which is typically composed of cylinders with circular, elliptical or rectangular cross-section. The first part of this chapter gives for different types of hologram the required orientation of the cylinder and the relationship between the cylinders' height and the propagation constant of the eigen-modes, assuming that no reflection occurs. Equations are also provided for two metasurfaces in series, giving all the possible solutions. Those equations can be used for wave plates. The second part presents a simple technique for the design of anti-reflective metasurfaces followed by the detailed design of a half-wave plate.

Chapter 5 develops the concept of self-coupling mode, which can be used for the analysis and design of resonant multi-mode metasurface. It also greatly facilitates the interpolation of the response of resonant metasurfaces. Four examples are provided: the Huygens' metasurface, a narrowband metasurface, a metasurface-based laser and a very high-Q metasurface for sensing application. Each of those examples presents a different aspect in the use of the self-coupling modes.

Chapter 6 adapts the adjoint method used in [63,89] for the Fourier modal method. The adjoint method gives the functional derivative of a merit function in function of the permittivity and permeability and it is used in conjunction with a gradient-based optimization method, such as the gradient-descent or the quasi-Newton method. As a proof of concept, the adjoint method is applied to the optimization of a 5x7 beam-splitter.

1.3 Notation and convention

The notation used in this paper is the following: $\vec{x}$ is a vector, $\hat{x}$ is a matrix, x_{mn} is the element at the m-th line and n-th column of the matrix $\hat{x}$, x^* and $\bar{x}$ denotes the complex conjugate of x, x^T and x^H denote respectively the transpose and the conjugate transpose of x, $\vec{x}_{/\!/}$ is the tangential components of the vector $\vec{x}$ relative to a surface and $x_\perp$ is its normal component. Moreover, the magnetic field $\vec{H}$ and the magnetization density $\vec{M}$ are normalized in the following way:

$$\vec{H} = \sqrt{\frac{\mu_0}{\epsilon_0}}\vec{H}' \qquad \vec{M} = \sqrt{\frac{\mu_0}{\epsilon_0}}\vec{M}', \tag{1.1}$$

where $\vec{H}'$ is the standard magnetic field, $\vec{M}'$ is the standard magnetization density, ϵ_0 is the permittivity in vacuum and μ_0 is the permeability in vacuum. The normalized magnetic field has the same unit as the electric field. The implicit time dependence is $e^{-i\omega t}$, where ω is the angular frequency.

Using the normalized magnetic field and magnetization density, the Maxwell equation for isotropic media with source becomes:

$$\begin{aligned} \nabla \times \vec{E} &= ik_0(\mu\vec{H} + \vec{M}) \\ \nabla \times \vec{H} &= -ik_0(\epsilon\vec{E} + \vec{P}), \end{aligned} \tag{1.2}$$

where $\vec{E}$ is the electric field, $\vec{P}$ is the polarization density, k_0 is the wavenumber in vacuum, ϵ is the relative permittivity and μ is the relative permeability. $\vec{P}$ and $\vec{M}$ represent a source. Except in chapter 6, the source terms $\vec{P}$ and $\vec{M}$ are zero.

In chapter 2, a bianisotropic media without a source is considered and the Maxwell equation is

$$\begin{aligned}\nabla\times\vec{E} &= ik_0(\hat{\zeta}\vec{E}+\hat{\mu}\vec{H}),\\ \nabla\times\vec{H} &= -ik_0(\hat{\epsilon}\vec{E}+\hat{\xi}\vec{H}),\end{aligned}\tag{1.3}$$

where $\hat{\zeta}$ and $\hat{\xi}$ are the bianisotropy parameters.

Only the Maxwell-Faraday equation and the Ampère's circuital law, which are given by equations eq. (1.2), are considered in this work since the Gauss's laws are redundant in the time-harmonic regime when the frequency is different than zero.

The structures considered in this work are a stack of layers composed of z-invariant z-symmetry invariant (ZSI) media and the interfaces between the layers are perpendicular to the z-axis. The ZSI property is introduced in section 2.1 and a formal definition is given in section 2.4. In chapters 4 and 5, the structures are composed of three layers: the substrate, a metasurface and a superstrate. If not stated otherwise, the illumination is a plane wave that propagates in the z direction.

Modes are solutions of the Maxwell equation in a region of space and they are denoted by ψ. Two types of modes are mentioned in this work: the eigen-modes and the self-coupling modes. Eigen-modes are modes whose z-dependency is $e^{i\gamma z}$, where γ is the propagation constant. In homogeneous media, the eigen-modes are plane waves and they are described in section 3.2.1. The self-coupling modes are present only in chapter 5 and they are defined in section 5.2.1.

2 Poynting operation

2.1 Introduction

We consider here optical materials that are invariant in their geometry along a principal propagation direction for which we take the z-axis. In such z-invariant medium, the electromagnetic field can be decomposed into modes, which are eigen-functions of the Maxwell equations. This mode decomposition is commonly used for the analysis and the simulation of optical fibers, such as photonic crystal fibers [96], but it can also be used for z-invariant metasurfaces, like the ones shown in the review of Genevet et *al.* [97]. Modes constitute the core of the Fourier modal method [61, 79] and similar methods [82, 98].

Modes inside a z-invariant heterogenous medium have several properties depending on its geometrical cross-section and the materials that compose the medium, as described in [99]. The study of such properties can lead to a better insight such as the classification of modes propagating inside a lossless medium [99, Chapter 11]. In some media, each mode has its backward-propagating counterpart. Those media are called bidirectional and the conditions for bidirectionality have been stated in the literature [100, 101]. In a bidirectional medium, the number of modes that need to be computed can usually be reduced, speeding up the simulation of the light propagating in such medium.

An important property is the mode orthogonality. In order to determine whether two modes are orthogonal, an operation has to be defined that maps two modes into a complex number. Two modes are orthogonal if the defined operation applied on those two modes gives zero. The equation that states the condition for mode orthogonality is called an orthogonality relation. Several orthogonality relations have been found [99, 102–104], usually based on the Lorentz reciprocity ([103], [104, Chapter 31]). In [105–107], the boundary condition is expressed using an operation related to an orthogonality relation. Due to mode orthogonality, the boundary condition is then greatly simplified.

The use of the appropriate operation for a given problem gives significant advantages such as a general expression of the boundary condition that has been simplified using the orthogonality of the modes [107]. An operation is also required for the normalization of modes. In order to facilitate the choice between the different operations, one has to get a comprehensive picture of the advantages and disadvantages related to their use. This work addresses this issue by giving a deep insight of the use of a set of operations that have the same properties as the following operation:

$$\frac{1}{2}\int_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n \times \vec{H}_m^*) \cdot \vec{n} ds, \tag{2.1}$$

where S is a surface, $\vec{n}$ is the surface normal and $(\vec{E}_m, \vec{H}_m)$ is the electric and magnetic fields of the mode m. For z-invariant heterogenous media, S is a plane perpendicular to the z-axis and the modes are solutions of the source-free Maxwell equations in the time harmonic regime.

We call the operation (2.1) the Poynting operation because the operation (2.1) is related to the complex Poynting vector, whose real part usually represents the power flow [108]. By extension, the operations that have the same properties are also called Poynting operation.

The operation (2.1) is well suited for lossless and z-symmetry invariant (ZSI) media, also known as strictly bidirectional media [99], due to the orthogonality relation presented in section 2.4. In ZSI media, which are a special case of bidirectional media, the fields of a mode propagating in one direction can be directly deduced from the fields of the related mode propagating in the opposite direction. Any medium composed of isotropic materials is a ZSI medium. If the materials that compose the medium are anisotropic or gyrotropic, the medium is still ZSI if the optical axis or the axis of gyration is parallel to the z-axis. Bianisotropic materials such as the Tellegen metacrystals presented in [109] are also ZSI. A formal definition of a ZSI medium is given in section 2.4.

This chapter is structured as follow. In section 2.2, different operations are presented followed by a discussion on their advantages and disadvantages. The presented operations are chosen such that they can be used to express the boundary condition in the same way, meaning that only the tangential components of the fields are required. In section 2.3, the Poynting operation is defined in an abstract way based on the properties of the operation (2.1). Then, additional properties, which have a physical meaning or are used latter in this chapter, are derived. A sesquilinear form associated to the Poynting operation is introduced. In section 2.4, an orthogonality relation for lossless and ZSI media in the bianisotropic case is presented. In the same section, the definition and properties of a ZSI medium are given. The derivation of the orthogonality relation is in section 2.10.1. In section 2.5, the fields on both sides of an interface are decomposed into modes and systems of equations involving the Poynting operation are proposed for the computation of the coupling coefficients between the modes. Their derivation is in section 2.10.2. In section 2.7.1, the coupling coefficients at the interface between air and a lossless z-invariant metamaterial are estimated, considering only the main mode in both media. We also compare our method presented in this work with other methods proposed in the literature [105, 106, 110]. In section 2.7.2, the Fresnel conditions generalized for uniaxial media are derived from the same equations used in section 2.7.1. In section 2.6, a set of operations applied on the Gram matrix of the sesquilinear form introduced in section 2.3, is proposed. Those matrix operations are similar to the elementary operations in the Gaussian elimination and they can be used to orthonormalize a set of modes. Their derivation is in section 2.10.3. In sections 2.8.2 to 2.8.4, the orthonormalization and rotation of propagating, evanescent and complex modes in a lossless uniaxial ZSI medium which is invariant to a 90°-rotation around the z-axis, are given using an algorithm based on the matrix operations presented in section 2.6. A rotation is defined here as a transformation from a set of orthonormal modes to another set of orthornormal modes.

2.2 Discussion on operations

An important criterion when choosing an operation is its usefulness, meaning that its use leads to important simplifications in derivations. For instance, the operation used in [105–107] has been selected because it is related to an orthogonality relation which is valid for reciprocal media. Based on the orthogonality relations in the literature [104, 111–113], the following operations can be distinguished:

$$\frac{1}{2}\int_S (\vec{E}_m \times \vec{H}_n - \vec{E}_n \times \vec{H}_m)\cdot \vec{n}ds. \tag{2.2}$$

$$\frac{1}{2}\int_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n^* \times \vec{H}_m)\cdot \vec{n}ds, \tag{2.3}$$

In a z-invariant reciprocal medium, modes are, in most cases, orthogonal if the operation (2.2) is used, meaning that an orthogonality relation exists. If the operation (2.3) is considered instead, modes are mostly orthogonal in a z-invariant lossless medium. The orthogonality relation for bianisotropic lossless media, which is related to the operation (2.3), is proved in section 2.10.1 and the orthogonality relation for anisotropic reciprocal media, which is related to the operation (2.2), is proved in the Chapter 31 of [104]. It can be easily generalized for bianisotropic media. If the medium has, in addition, the ZSI property, the operations (2.2) and (2.3) can be modified in the following way while keeping the orthogonality of the modes:

$$\frac{1}{2}\int_S (\vec{E}_m \times \vec{H}_n + \vec{E}_n \times \vec{H}_m)\cdot \vec{n}ds. \tag{2.4}$$

$$\frac{1}{2}\int_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n \times \vec{H}_m^*)\cdot \vec{n}ds, \tag{2.5}$$

In order to compare the different operations, the physical meaning, the occurrence of self-orthogonal modes, and the validity of the orthogonality relation are considered. An operation related to a physical quantity, such as the operations (2.3) and (2.5), leads to a meaningful normalization. Hence, the weight of the different modes has a physical meaning. The operation (2.3) is related to the real Poynting vector. The operation (2.5) is related to the complex Poynting vector. To our knowledge, the operations (2.2) and (2.4) do not have a physical meaning.

Since the operations (2.2) to (2.5) are not definite, self-orthogonal modes can be present, complicating the normalization of the modes. Hence, the operation (2.4) is a better option in that respect than the operation (2.2), since all the modes are self-orthogonal when the operation (2.2) is used. Since the operation (2.5) is related to the complex Poynting vector, the active and reactive power of a mode is obtained, avoiding the self-orthogonality of the evanescent modes.

The operations (2.2) to (2.5) are related to an orthogonality relation [99, 103, 104] but are valid only under certain conditions. For the operations (2.2) and (2.4), one condition is that the medium has to be reciprocal. For the operations (2.3) and (2.5), the medium has to be lossless. In that sense, the operations (2.2) and (2.4) have an advantage since reciprocal media are more common than lossless media. However, in the case of a periodic structure, the orthogonality relation related to the operations (2.2) and (2.4) is no more valid when the Bloch phase is not null, which is the reason behind the use of self-adjoint modes in [106]. For the operations (2.3) and (2.5), the orthogonality relation is still valid. Compared to the operations (2.2) and (2.3), the additional condition in order to have a valid orthogonality relation for the operations (2.4) and (2.5) is the ZSI property of the medium.

The choice to focus on the operation (2.5) instead of the operation (2.3) is a matter of taste because both are based on the same sesquilinear form as shown in section 2.3, meaning that formalisms based on those operations are similar.

2.3 Poynting operation

In ZSI z-invariant medium, the Poynting operation can be defined as:

$$[\psi_m|\psi_n] := \frac{1}{2}\int_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n \times \vec{H}_m^*) \cdot \vec{n}\, ds, \tag{2.6}$$

where $[\cdot|\cdot]$ represents the Poynting operation, ψ are modes, and S is a plane perpendicular to the z-axis. S is typically an infinite plane for an aperiodic medium and a unit cell for a periodic medium. A formal and more generalized definition is given later in this section. A complete set of modes can be computed by finding the solutions of Maxwell's equations (1.3) of the form:

$$(\vec{E}(x,y,z), \vec{H}(x,y,z)) = (\vec{E}_0(x,y), \vec{H}_0(x,y))e^{i\gamma z}, \tag{2.7}$$

where γ is called the propagation constant. However, any linear combination of those modes also gives another mode. In the definition of the Poynting operation (2.6), only the tangential components of the electric and magnetic fields are required. Therefore, the mode ψ is defined as:

$$\psi := (\vec{E}_{\parallel}, \vec{H}_{\parallel}). \tag{2.8}$$

In ZSI z-invariant medium, if the mode ψ with a propagation constant γ exists, a mode with the propagation constant $-\gamma$, called ψ^-, is also a solution to Maxwell's equations. The fields of the modes ψ and ψ^- are related:

$$\psi^- = (\vec{E}_{\parallel}, -\vec{H}_{\parallel}). \tag{2.9}$$

We call the operator $(\cdot)^-$ the minus operator and it changes a forward-propagating mode into the corresponding backward-propagating mode. In passive media, the amplitude of a forward-propagating mode decreases along z and the z-component of its power flux is positive.

In an isotropic non-dispersive medium, the power flux of a given mode m through the surface S is the real part of $[\psi_m|\psi_m]$ multiplied by a constant [108]. The imaginary part of $[\psi_m|\psi_m]$ is known as the reactive power. In general, the power flux Φ_S carried by a set of modes through the surface S is proportional to:

$$\Phi_S \propto \operatorname{Re}\left\{ \left[\sum_{m=1}^{M} a_m \psi_m + b_m \psi_m^- \middle| \sum_{m=1}^{M} a_m \psi_m + b_m \psi_m^- \right] \right\}. \tag{2.10}$$

The choice of the definition of the Poynting operation only affects the validity of the orthogonality relation presented in section 2.4. For the reformulation of the boundary condition and the operations on the Gram matrix presented in, respectively, sections 2.5 and 2.6, only the properties of the Poynting operation are needed. Therefore, the definition of the Poynting operation is generalized such that any operation that has the same properties as the operation (2.6) is also the Poynting operation.

The Poynting operation is an operation with the map:

$$[\cdot|\cdot] : \mathbb{V} \times \mathbb{V} \to \mathbb{C}, \tag{2.11}$$

and the properties:

$$
\begin{aligned}
[\psi_m|\psi_n] &= [\psi_n|\psi_m], && (2.12a)\\
[\psi_p|\psi_m+\psi_n] &= [\psi_p|\psi_m]+[\psi_p|\psi_n], && (2.12b)\\
[\psi_m|k\psi_n] &= \mathrm{Re}\{k\}[\psi_m|\psi_n]+i\,\mathrm{Im}\{k\}[\psi_m|\psi_n^-], && (2.12c)\\
[\psi_m^-|\psi_n^-] &= -[\psi_m|\psi_n]. && (2.12d)
\end{aligned}
$$

$\mathbb{V}$ is a vector space over the field $\mathbb{C}$ and ψ is an element of $\mathbb{V}$. k is a complex number. Due to property (2.12c), the Poynting operation is neither a bilinear map nor a sesquilinear map. The minus operator associated to the Poynting operation is:

$$
(\cdot)^- : \mathbb{V} \to \mathbb{V} \tag{2.13}
$$

with the following properties:

$$
\begin{aligned}
(\psi_m+\psi_n)^- &= \psi_m^-+\psi_n^-, && (2.14a)\\
(k\psi)^- &= k\psi^-, && (2.14b)\\
\psi^{--} &= \psi. && (2.14c)
\end{aligned}
$$

From the properties (2.12), a set of additional properties can be derived:

$$
\begin{aligned}
[\psi_m|\psi_n^-] &= -[\psi_m^-|\psi_n], && (2.15a)\\
[\psi_m|\psi_m^-] &= 0, && (2.15b)\\
[\psi_m|i\psi_m^-] &= i[\psi_m|\psi_m], && (2.15c)\\
[s\psi_m|t\psi_n] &= \mathrm{Re}\{\bar{s}t\}[\psi_m|\psi_n]+i\,\mathrm{Im}\{\bar{s}t\}[\psi_m|\psi_n^-], && (2.15d)\\
[s\psi_m|t(\psi_n+\psi_n^-)] &= \bar{s}t[\psi_m|\psi_n+\psi_n^-], && (2.15e)
\end{aligned}
$$

where s and t are complex numbers. The properties (2.12d) and (2.15a) express that a minus sign appears when the propagation direction of both modes is flipped. From the property (2.15b), a forward-propagating mode is orthogonal to its backward-propagating counterpart. However, under the property (2.15c), an evanescent mode, suggesting that $[\psi|\psi]$ is purely imaginary, carries power if it interacts with its backward-propagating counterpart dephased by ±90°. The property (2.15d) is the general formula when both modes are weighted. The use of the properties (2.12c) and (2.15d) complicates the derivations for the different proofs. Therefore, we introduce the function σ defined by

$$\sigma(\psi_m, \psi_n) := [\psi_m | \psi_n + \psi_n^-]. \tag{2.16}$$

Due to properties (2.12b) and (2.15e), σ is a sesquilinear form. It has additional properties which are related to the minus operator:

$$\sigma(\psi_m^-, \psi_n) = -\sigma(\psi_m, \psi_n), \tag{2.17a}$$

$$\sigma(\psi_m, \psi_n^-) = \sigma(\psi_m, \psi_n). \tag{2.17b}$$

The sesquilinear form σ is mostly used for the proof of the reformulation of the boundary condition and the operation on the Gram matrix (sections 2.10.2 and 2.10.3). Using the definition (2.6):

$$\sigma(\psi_m, \psi_n) = \int_S (\vec{E}_n \times \vec{H}_m^*) \cdot \vec{n} ds, \tag{2.18a}$$

$$\sigma(\psi_n, \psi_m) + \sigma(\psi_m, \psi_n)^* = \int_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n^* \times \vec{H}_m) \cdot \vec{n} ds. \tag{2.18b}$$

The right-hand term of equation (2.18b) is the operation (2.3). Hence, an important part of this work can be used when the operation (2.3) is considered instead of the Poynting operation.

2.4 Orthogonality relation

In this section, the condition for mode orthogonality is derived in a bianisotropic lossless non-dispersive ZSI z-invariant medium. The constitutive relation for a bianisotropic medium is given by

$$\begin{pmatrix} \frac{\vec{D}}{\epsilon_0} \\ c_0 \vec{B} \end{pmatrix} = \begin{pmatrix} \hat{\epsilon} & \hat{\xi} \\ \hat{\zeta} & \hat{\mu} \end{pmatrix} \begin{pmatrix} \vec{E} \\ \vec{H} \end{pmatrix}, \tag{2.19}$$

where c_0 is the speed of light in vacuum, $\vec{D}$ is the displacement field, and $\vec{B}$ is the magnetic flux density. The constitutive relation (2.19) is similar to the one found in [114] up to some constants and it leads to Maxwell's equations (1.3). Due to the use of the normalized magnetic field, $\hat{\epsilon}$, $\hat{\mu}$, $\hat{\xi}$, and $\hat{\zeta}$ are unitless. To be able to derive the condition for mode orthogonality when the surface of integration is a plane at $z = constant$, several assumptions have to be done. First, the medium is z-invariant meaning that $\hat{\epsilon}$, $\hat{\mu}$, $\hat{\xi}$, and $\hat{\zeta}$ depend only on x and y. Hence, the electric and magnetic field of a mode has the following form:

$$\vec{X}_m(x,y,z) = \vec{X}_{m0}(x,y)e^{i\gamma_m z}, \tag{2.20}$$

where $\vec{X}$ can be the electric or the magnetic field and γ_m is the propagation constant. Second, the medium has to be lossless, meaning that [114]

$$\hat{\epsilon} = \hat{\epsilon}^H, \qquad \hat{\mu} = \hat{\mu}^H, \qquad \hat{\zeta} = \hat{\xi}^H. \tag{2.21}$$

Third, the following assumption has to be fulfilled:

$$\oint_{\partial S} \vec{n} \times (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n^* \times \vec{H}_m) \cdot d\vec{l} = 0. \tag{2.22}$$

It means that the field has to vanish at the boundary of S or, for a periodic medium, the surface S corresponds to a unit cell. Using the assumptions (2.21) and (2.22), the following orthogonality relation is valid:

$$(\gamma_m - \bar{\gamma}_n) \iint_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n^* \times \vec{H}_m) \cdot \vec{n} ds = 0. \tag{2.23}$$

Finally, the medium has to be ZSI. As stated by [101], in a ZSI medium, if the mode described by $(\vec{E}_\parallel, E_\perp, \vec{H}_\parallel, H_\perp, \gamma)$ fulfils the Maxwell equations, the mode described by $(\vec{E}_\parallel, -E_\perp, -\vec{H}_\parallel, H_\perp, -\gamma)$ is still a solution of the Maxwell equations. The property of ZSI media is stated in [115] and it is

$$\begin{gathered}
\hat{\epsilon} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & 0 \\ \epsilon_{21} & \epsilon_{22} & 0 \\ 0 & 0 & \epsilon_{33} \end{pmatrix}, \qquad \hat{\mu} = \begin{pmatrix} \mu_{11} & \mu_{12} & 0 \\ \mu_{21} & \mu_{22} & 0 \\ 0 & 0 & \mu_{33} \end{pmatrix}, \\
\hat{\zeta} = \begin{pmatrix} 0 & 0 & \zeta_{13} \\ 0 & 0 & \zeta_{23} \\ \zeta_{31} & \zeta_{32} & 0 \end{pmatrix}, \qquad \hat{\xi} = \begin{pmatrix} 0 & 0 & \xi_{13} \\ 0 & 0 & \xi_{23} \\ \xi_{31} & \xi_{32} & 0 \end{pmatrix}.
\end{gathered} \tag{2.24}$$

If the assumptions (2.21), (2.22) and (2.24) are fulfilled and the definition (2.6) is used, the following orthogonality relations hold:

$$(\gamma_m^2 - \bar{\gamma}_n^2)[\psi_m|\psi_n] = 0, \tag{2.25}$$

$$(\gamma_m^2 - \bar{\gamma}_n^2)[\psi_m|\psi_n + \psi_n^-] = 0. \tag{2.26}$$

As a curiosity, a mode is self-orthogonal if its propagation constant has a real and an imaginary part. In lossless media, those modes are called complex waves or complex modes and it has been studied in [116] and in the Chapter 11-12 of [99]. The proof of the orthogonality relations (2.23), (2.25) and (2.26) is in section 2.10.1.

2.5 Reformulation of the boundary condition

Let us consider a surface S that acts as an interface between two media. The boundary condition states that the components of the electric and magnetic fields that are tangential to the surface is equal on both sides of the interface. By using the definition of the Poynting operation given in (2.6) where the surface of integration is the interface, the modes are described only by the tangential components of their fields. Hence, if L are the modes in the left medium, R are the modes in the right medium, and the forward direction is from left to right, the boundary condition can be written as:

$$\sum_{m=1}^{M} p_m L_m + r_m L_m^- = E + \sum_{n=1}^{N} q_n R_n^- + t_n R_n, \tag{2.27}$$

where p_m, r_m, q_n, and t_n are the weights of the different modes and E is the error term. The weights which are unknown depend on the problem at hand. Usually, p_m and q_n, which are the weights of the modes propagating towards the interface, are known and r_m and t_n are the unknowns. Please note that the Poynting operation (2.6) can only be used if the interface is perpendicular to the z-axis and both media are ZSI z-invariant.

If all the modes on both media are considered, the unknowns have to be found such that the error term E, which is the fields mismatch at the interface, is null. In the case where only a subset of modes is considered, equation (2.27) may not admit a solution, but the unknowns can be estimated based on some assumptions. A naive way is to find the unknowns such that the error term E is minimized in the least-mean-square sense, but it has been shown that it gives inaccurate results [105]. The different equations proposed in this section allow to find the unknowns based on the result of the Poynting operation applied on the error term E and the considered modes L_m and R_n. Let us introduce the following expressions:

$$\begin{aligned}
S_{Lu} &:= [L_u|E+E^-] &&= \sum_{m=1}^{M} (p_m + r_m)[L_u|L_m + L_m^-] &&- \sum_{n=1}^{N} (t_n + q_n)[L_u|R_n + R_n^-],\\
S_{Rv} &:= [R_v|E+E^-] &&= \sum_{m=1}^{M} (p_m + r_m)[R_v|L_m + L_m^-] &&- \sum_{n=1}^{N} (t_n + q_n)[R_v|R_n + R_n^-],\\
T_{Lu} &:= [E|L_u + L_u^-]^* &&= \sum_{m=1}^{M} (p_m - r_m)[L_m|L_u + L_u^-]^* &&- \sum_{n=1}^{N} (t_n - q_n)[R_n|L_u + L_u^-]^*,\\
T_{Rv} &:= [E|R_v + R_v^-]^* &&= \sum_{m=1}^{M} (p_m - r_m)[L_m|R_v + R_v^-]^* &&- \sum_{n=1}^{N} (t_n - q_n)[R_n|R_v + R_v^-]^*,
\end{aligned} \tag{2.28}$$

where $v \in [1, M]$ and $u \in [1, N]$. The expressions (2.28) can be written in a compact form:

$$\begin{aligned}
\vec{S}_L &= \hat{G}_{LL}(\vec{p}+\vec{r}) - \hat{G}_{LR}(\vec{t}+\vec{q}),\\
\vec{S}_R &= \hat{G}_{RL}(\vec{p}+\vec{r}) - \hat{G}_{RR}(\vec{t}+\vec{q}),\\
\vec{T}_L &= \hat{G}^H_{LL}(\vec{p}-\vec{r}) - \hat{G}^H_{RL}(\vec{t}-\vec{q}),\\
\vec{T}_R &= \hat{G}^H_{LR}(\vec{p}-\vec{r}) - \hat{G}^H_{RR}(\vec{t}-\vec{q}),
\end{aligned} \tag{2.29}$$

where the m-th row and the n-th column of $\hat{G}_{XY}$ is given by

$$\hat{G}_{XYmn} = [X_m | Y_n + Y_n^-]. \tag{2.30}$$

Due to the orthogonality relation presented in section 2.4, $\hat{G}_{LL}$ and $\hat{G}_{RR}$ are sparse matrices when lossless ZSI media are considered. When E is null, the expressions (2.28) are equal to zero. Hence, unknowns in equation (2.27) can be found as the solution of a system of equations composed of the expressions S and T. The expressions S and T are related to each others by the equations

$$\begin{aligned}
\sum_{m=1}^{M} (\bar{p}_m - \bar{r}_m) S_{Lm} &= \sum_{n=1}^{N} (\bar{t}_n - \bar{q}_n) S_{Rn},\\
\sum_{m=1}^{M} (\bar{p}_m + \bar{r}_m) T_{Lm} &= \sum_{n=1}^{N} (\bar{t}_n + \bar{q}_n) T_{Rn}.
\end{aligned} \tag{2.31}$$

Hence, the expressions S and T are not independent when E is null. As an example, if r and t are the unknowns and p and q are known, the system of equations composed of all the expressions S admits an infinity of solutions.

The expressions S and T are related by the partial derivative of the Poynting operation applied on E. Therefore, they can be used for optimization purposes:

$$\begin{aligned}
\frac{\partial}{\partial \bar{p}_u}[E|E] &= S_{Lu}, & \frac{\partial}{\partial \bar{p}_u}[E|E]^* &= T_{Lu},\\
\frac{\partial}{\partial \bar{r}_u}[E|E] &= -S_{Lu}, & \frac{\partial}{\partial \bar{r}_u}[E|E]^* &= T_{Lu},\\
\frac{\partial}{\partial \bar{t}_v}[E|E] &= -S_{Rv}, & \frac{\partial}{\partial \bar{t}_v}[E|E]^* &= -T_{Rv},\\
\frac{\partial}{\partial \bar{q}_v}[E|E] &= S_{Rv}, & \frac{\partial}{\partial \bar{q}_v}[E|E]^* &= -T_{Rv},
\end{aligned} \tag{2.32}$$

where $\frac{\partial}{\partial \bar{z}}$ is one of the Wirtinger derivatives, which is defined as

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right) \tag{2.33}$$

with $z = x + iy$ and x and y are real.

The proof of the statements of this section is in section 2.10.2 and an example of application is given in section 2.7.1, where an estimation of the coupling coefficients at an interface between air and a lossless ZSI z-independent metamaterial is given. The accuracy of the estimation of the method based on the Poynting operation is compared to other methods proposed in the literature.

2.6 Operations on the Gram matrix

Two modes ψ_m and ψ_n are orthogonal with respect to the Poynting operation if $[\psi_m|\psi_n] = 0$ and $[\psi_m|\psi_n^-] = 0$ for $m \neq n$. Hence, a set of modes is orthogonal if the matrix $\hat{G}$ is diagonal with

$$G_{mn} = [\psi_m|\psi_n + \psi_n^-], \; m, n \in [1, M] \tag{2.34}$$

since the expressions $[\psi_m|\psi_n]$ and $[\psi_m|\psi_n^-]$ are given by

$$\begin{aligned} [\psi_m|\psi_n] &= \frac{1}{2}(G_{mn} + G_{nm}), \\ [\psi_m|\psi_n^-] &= \frac{1}{2}(G_{mn} - G_{nm}). \end{aligned} \tag{2.35}$$

The matrix $\hat{G}$ is a square matrix of size $M \times M$ and it is the Gram matrix of the sesquilinear form σ. To diagonalize $\hat{G}$, an algorithm similar to the Gaussian elimination can be used where the usual operations are replaced by the operations shown in table 2.1. For the listing of operations on $\hat{G}$, L_{Gm} and C_{Gm} refer to, respectively, the m-th line and column of the matrix $\hat{G}$, $B \longleftarrow A$ means that the object B is replaced by the object A, and $A \longleftrightarrow B$ means that the objects A and B are swapped. In practice, it can be preferable to not change the modes at each iteration and to have the modes at the r-th iteration in this form:

$$\vec{\psi}_r = \frac{1}{2}\hat{A}_r(\vec{\psi}_0 + \vec{\psi}_0^-) + \frac{1}{2}\hat{B}_r(\vec{\psi}_0 - \vec{\psi}_0^-), \tag{2.36}$$

where $\vec{\psi}_r$, $\hat{A}_r$, and $\hat{B}_r$ are the mode ψ, the matrix $\hat{A}$ and the matrix $\hat{B}$ after the r-th iteration, and $\vec{\psi}_0$ are the initial modes. $\hat{A}_0$ and $\hat{B}_0$ are the identity matrix.

Using the operations in table 2.1, any Gram matrix can be transformed into the identity matrix

Operation name	Mode operation
Mode swapping	$\psi_m \longleftrightarrow \psi_n$
Mode scaling	$\psi_m \longleftarrow k\psi_m$
Mode reversal	$\psi_m \longleftarrow \psi_m^-$
Mode composition 1	$\psi_m \longleftarrow \psi_m + \frac{k}{2}(\psi_n + \psi_n^-)$
Mode composition 2	$\psi_m \longleftarrow \psi_m + \frac{k}{2}(\psi_n - \psi_n^-)$

Operation name	Operation on $\hat{A}$, $\hat{B}$	Operation on $\hat{G}$
Mode swapping	$L_{Am} \longleftrightarrow L_{An}$ $L_{Bm} \longleftrightarrow L_{Bn}$	$C_{Gm} \longleftrightarrow C_{Gn}$ $L_{Gm} \longleftrightarrow L_{Gn}$
Mode scaling	$L_{Am} \longleftarrow kL_{Am}$ $L_{Bm} \longleftarrow kL_{Bm}$	$C_{Gm} \longleftarrow kC_{Gm}$ $L_{Gm} \longleftarrow \bar{k}L_{Gm}$
Mode reversal	$L_{Bm} \longleftarrow -L_{Bm}$	$L_{Gm} \longleftarrow -L_{Gm}$
Mode composition 1	$L_{Am} \longleftarrow L_{Am} + kL_{An}$	$C_{Gm} \longleftarrow C_{Gm} + kC_{Gn}$
Mode composition 2	$L_{Bm} \longleftarrow L_{Bm} + kL_{Bn}$	$L_{Gm} \longleftarrow L_{Gm} + \bar{k}L_{Gn}$

Table 2.1 – A set of operations that can be done on $\hat{G}$ and its consequences on different quantities.

if modes with different propagation constants and different propagation direction can be combined. However, if the combination of modes with different propagation directions is forbidden, the Gram matrix can only be diagonalized since the operations that keep the Gram matrix diagonal, namely "Mode scaling" and "Mode reversal", cannot transform a diagonal matrix with complex values into the identity matrix. Moreover, the diagonalization of the Gram matrix when a lossy medium is considered becomes challenging because the operation "Mode composition 1" has to be applied with "Mode composition 2", transforming both a line and a column of the Gram matrix.

The operations "Mode composition 1" commutes with "Mode composition 2" and they can be combined into the following operation:

$$\psi_m \longleftarrow \psi_m + \frac{k_1}{2}(\psi_n + \psi_n^-) + \frac{k_2}{2}(\psi_n - \psi_n^-), \tag{2.37}$$

where k_1 and k_2 are the constants related to the operation "Mode composition 1" and "Mode composition 2" respectively. The Gram matrix at the r-th iteration $\hat{G}_r$ can directly be computed using $\hat{A}_r$, $\hat{B}_r$, and the initial Gram matrix $\hat{G}_0$:

$$\hat{G}_r = \hat{B}_r^* \hat{G}_0 \hat{A}_r^T. \tag{2.38}$$

The different operations on the Gram matrix presented in this section are proved in section 2.10.3 and, as an example of application, a set of complex modes are orthonormalized in section 2.8.3.

2.7 Reformulation of the boundary condition: Applications

2.7.1 Estimation of the coupling efficiencies at the surface of a metamaterial

The reformulation of the boundary condition presented in section 2.5 can be used in a similar way as in [106], where a small set of modes is considered on both sides of an interface and the reflection and transmission coefficients are estimated based only on the electric and magnetic fields of the considered modes. In this Appendix, we compare different methods to estimate the coupling efficiency of the system shown in fig. 2.1a. The considered methods are those presented in section 2.5 and in [105, 106], and the overlap integral often used for the estimation of the coupling into a fiber [117].

As shown in fig. 2.1a, the considered system is composed of two lossless media. The left medium is air. The right medium is a 2D-periodic z-invariant metamaterial composed of cylinders in air. The material that composes the cylinder is either glass, with a refractive index of $n = 1.44$, or silicon, with a refractive index of $n = 3.48$. The lattice dimension of the right medium is chosen such that a single mode propagates in both media. The incident mode L is an x-polarized plane wave propagating in the left medium at normal incidence at a wavelength of 1550 nm. The mode R propagating in the right medium is also known and it has been computed using the Fourier modal method. The amplitude of the x-component of the electric field of the mode R is shown in figs. 2.1b to 2.1d for different lattice dimensions and for both silicon and glass cylinders. The system in fig. 2.1a has been chosen because the mode R is significantly different from the mode L, making the estimation of the coupling coefficients a challenge.

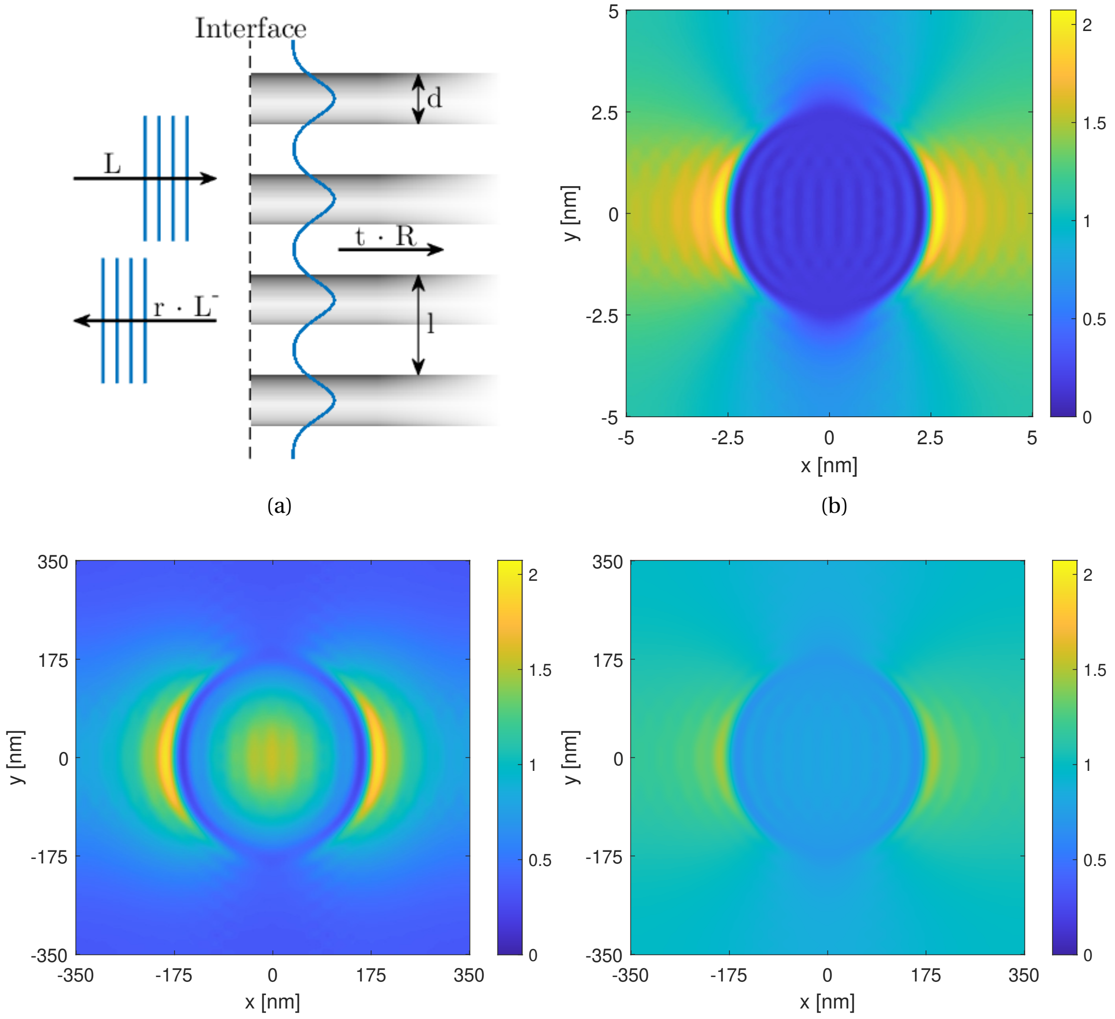
For the estimation of the coupling coefficients based on the Poynting operation, the following definition, similar to (2.6), is used:

$$[\psi_m|\psi_n] := \frac{1}{2|\Lambda|}\int_\Lambda (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n \times \vec{H}_m^*)\cdot \vec{n}\, ds, \tag{2.39}$$

where Λ is the lattice and $|\Lambda|$ is the lattice area. The operation presented in [103, 105, 106] is defined as

$$\langle\psi_m|\psi_n\rangle := \frac{1}{2|\Lambda|}\int_\Lambda (\vec{E}_m \times \vec{H}_n - \vec{E}_n \times \vec{H}_m)\cdot \vec{n}\, ds. \tag{2.40}$$

Since the considered modes are propagating modes in lossless media, the modes can be scaled



(a)

(b)

(c)

(d)

Figure 2.1 – a) Schematic of the system, which is composed of two media. The left medium is air. The right medium is a 2D periodic arrangement of cylinders in air. The lattice dimension varies from $l = 10\,nm$ to $l = 1000\,nm$ and the diameter d is half the lattice dimension. The cylinders are made of glass ($n = 1.44$) or silicon ($n = 3.48$). For a normal incident plane wave L, only one mode per medium are propagating, namely the mode L^- in the left medium and R in the right medium. (b-c) Amplitude of the x-component of the electric field of the mode R when the cylinders are made of silicon and the lattice dimension is $l = 10\,nm$ and $l = 700\,nm$ respectively. (d) Amplitude of the x-component of the electric field of the mode R when the cylinders are made of glass and the lattice dimension is $l = 700\,nm$. The color bars at the right hand-side of figs. 2.1b to 2.1d are the same for comparison purposes.

by a complex number such that the tangential components of the electric and magnetic fields of the modes are purely real, meaning that

$$[\psi_m|\psi_n] = \langle \psi_m^-|\psi_n\rangle \tag{2.41}$$

and those two quantities are purely real. In the definition of the operator in [106], the adjoint field of the modes, which are solution of Maxwell's equations for the reversed Bloch phase, is used. Since normal incidence is considered in this Appendix, the Bloch phase is zero and the operator presented in [106] is the same as in [105].

When a single mode is considered in both media, the expressions (2.28) become:

$$\begin{aligned} S_L &= (1+r)[L|L] && - t[L|R+R^-], \\ S_R &= (1+r)[R|L+L^-] && - t[R|R], \\ T_L &= (1-r)[L|L]^* && - t[R|L+L^-]^*, \\ T_R &= (1-r)[L|R+R^-]^* && - t[R|R]^*. \end{aligned} \tag{2.42}$$

If the modes L and R do not couple with the modes that are not considered in the system, the system of equations composed of $S_L = 0$ and $S_R = 0$, or $T_L = 0$ and $T_R = 0$, does not admit an unique solution due to equation (2.31). Therefore, four different systems of equations are proposed for the estimation of the coupling coefficients r and t:

$$\begin{cases} S_L = (1+r_1)[L|L] - t_1[L|R+R^-] = 0 \\ T_L = (1-r_1)[L|L]^* - t_1[R|L+L^-]^* = 0 \end{cases} \Rightarrow \begin{aligned} r_1 &= \frac{[L|L]^*[L|R+R^-]-[L|L][R|L+L^-]^*}{[L|L]^*[L|R+R^-]+[L|L][R|L+L^-]^*} \\ t_1 &= \frac{2|[L|L]|^2}{[L|L]^*[L|R+R^-]+[L|L][R|L+L^-]^*} \end{aligned} \tag{2.43a}$$

$$\begin{cases} S_L = (1+r_2)[L|L] - t_2[L|R+R^-] = 0 \\ T_R = (1-r_2)[L|R+R^-]^* - t_2[R|R]^* = 0 \end{cases} \Rightarrow \begin{aligned} r_2 &= \frac{|[L|R+R^-]|^2-[L|L][R|R]^*}{|[L|R+R^-]|^2+[L|L][R|R]^*} \\ t_2 &= \frac{2[L|L][L|R+R^-]^*}{|[L|R+R^-]|^2+[L|L][R|R]^*} \end{aligned} \tag{2.43b}$$

$$\begin{cases} S_R = (1+r_3)[R|L+L^-] - t_3[R|R] = 0 \\ T_L = (1-r_3)[L|L]^* - t_3[R|L+L^-]^* = 0 \end{cases} \Rightarrow \begin{aligned} r_3 &= \frac{[R|R][L|L]^*-|[R|L+L^-]|^2}{[R|R][L|L]^*+|[R|L+L^-]|^2} \\ t_3 &= \frac{2[L|L]^*[R|L+L^-]}{[R|R][L|L]^*+|[R|L+L^-]|^2} \end{aligned} \tag{2.43c}$$

$$\begin{cases} S_R = (1+r_4)[R|L+L^-] - t_4[R|R] = 0 \\ T_R = (1-r_4)[L|R+R^-]^* - t_4[R|R]^* = 0 \end{cases} \Rightarrow \begin{aligned} r_4 &= \frac{[R|R][L|R+R^-]^*-[R|R]^*[R|L+L^-]}{[R|R][L|R+R^-]^*+[R|R]^*[R|L+L^-]} \\ t_4 &= \frac{2[R|L+L^-][L|R+R^-]^*}{[R|R][L|R+R^-]^*+[R|R]^*[R|L+L^-]} \end{aligned} \tag{2.43d}$$

An interesting property of the solutions (r_2, t_2) and (r_3, t_3) is that they satisfy the following equality:

$$[L + r_m L^-|L + r_m L^-] = [t_m R|t_m R], \; m \in \{2,3\}. \tag{2.44}$$

Since the integration of the Poynting vector is the same on both sides of the interface, it is guaranteed that no absorption or power generation can occur at the interface, but it also means that no power is carried by other modes. In this example, all the propagating modes that can be excited by the mode L are considered. Therefore, such solutions are consistent with the system at hand since no other mode than L and R can carry power away from the interface.

When the tangential components of the electric and magnetic fields are purely real, which is the case in this example, the solutions (r_1, t_1) and (r_4, t_4) are related to the solutions proposed in [105, 106]. In [106], the estimations of the coupling coefficients are:

$$r_P = -\frac{\langle R|L\rangle}{\langle R|L^-\rangle}, \qquad t_P = -\frac{\langle L^-|L\rangle}{\langle L^-|R\rangle}. \tag{2.45}$$

By applying equation (2.41), r_P and t_P are equal to r_1 and t_1 respectively. In [105], the estimations of the coupling coefficients are:

$$r_S = -\frac{\langle L|R\rangle}{\langle L^-|R\rangle} \qquad t_S = \langle R^-|L\rangle - \frac{\langle R^-|L^-\rangle\,\langle R|L\rangle}{\langle R|L^-\rangle} \tag{2.46}$$

In this case, r_S and t_S are equal to r_4 and t_4 respectively. Since R and L are propagating modes, r_1 is also equal to r_4.

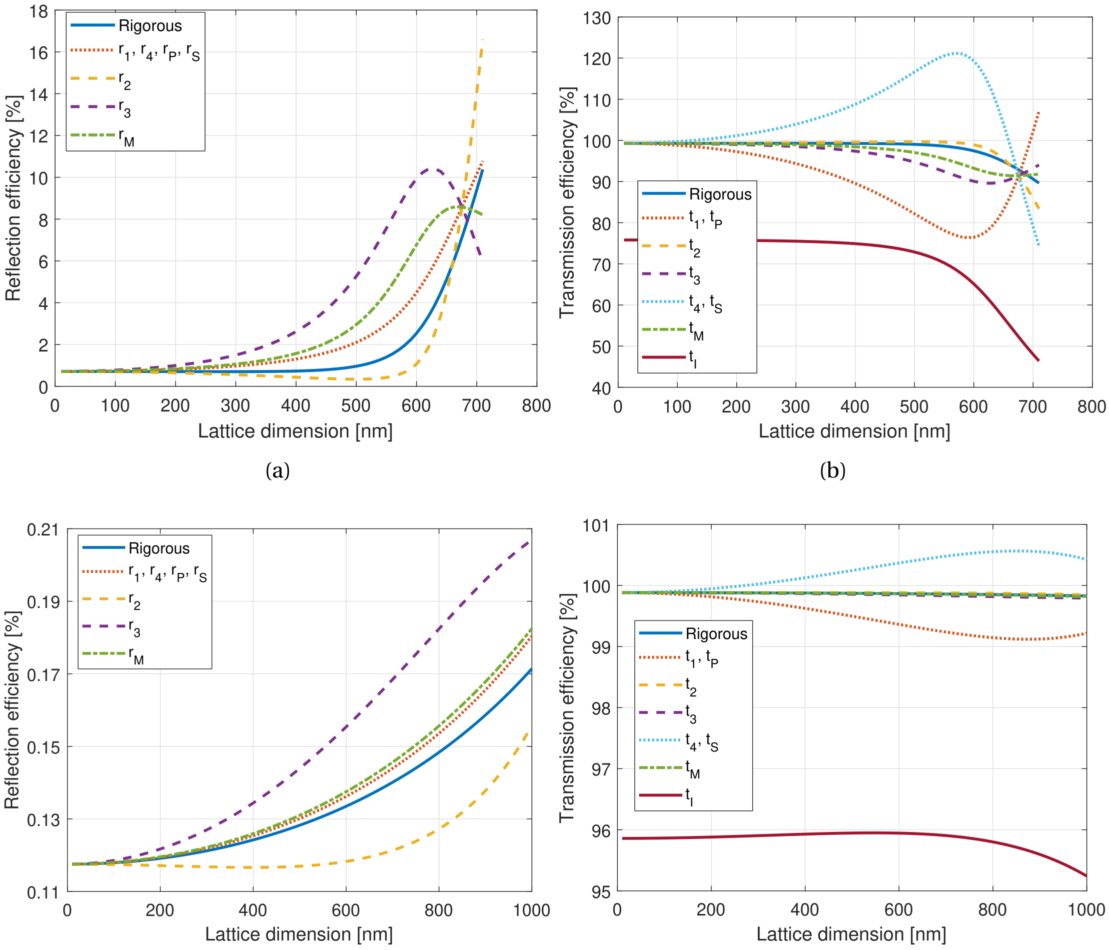
Two other methods are proposed. For the first method, in order to avoid making an arbitrary choice between the solutions (r_2, t_2) and (r_3, t_3), the reflection and transmission coefficients, called r_M and t_M, are chosen such that equation (2.44) is satisfied and the following quantity is minimized:

$$|S_L|^2 + |S_R|^2 + |T_L|^2 + |T_R|^2. \tag{2.47}$$

The second method is the overlap integral for the estimation of the transmission efficiency, which is given by:

$$t_I = \frac{\int_\Lambda \vec{E}_L \cdot \vec{E}_R^* ds}{\sqrt{\int_\Lambda |\vec{E}_L|^2 ds \int_\Lambda |\vec{E}_R|^2 ds}}, \tag{2.48}$$

where E_L and E_R are the electric field of the mode L and R respectively. In the literature, the overlap integral is usually computed from the scalar field [110, 117] but it is used when the z-component of the electric field is negligible compared to the tangential field. We choose to use the vectorial field in the definition of the overlap integral. Using only the x-component of the electric field in the definition of the overlap integral doesn't improve its ability to estimate the reflection and transmission coefficients.



(a)

(b)

(c)

(d)

Figure 2.2 – (a-b) Reflection and transmission efficiencies computed with different methods when the right medium is composed of silicon cylinders. (c-d) Reflection and transmission efficiencies computed with different methods when the right medium is composed of glass cylinders. For the transmission efficiency plot, the curves obtained from rigorous simulation and from the transmission coefficients t_2, t_3 and t_M are visually superimposed.

In fig. 2.2, the coupling efficiencies are plotted for different lattice dimensions and for the system composed with silicon cylinders and the system composed with glass cylinders. The reflection efficiency η_r and transmission efficiency η_t are defined as:

$$\eta_r = |r|^2 \qquad\qquad \eta_t = \frac{[R|R]}{[L|L]}|t|^2 \tag{2.49}$$

In order to compare the different methods, the coupling coefficients are computed rigorously using the Fourier modal method and are used as the reference. The estimation obtained using the overlap integral is clearly inaccurate and it should not be used for such systems. For the other methods, they all converge to the reference for small lattice dimensions and start to significantly diverge when the lattice dimension is larger than one tenth of the wavelength. When the estimated coupling coefficients starts diverging, the different methods give also different results. It is an advantage of having multiple formula for the estimation of the coupling coefficients since it gives an indication of the validity of the estimation. However, exceptions can occur as shown in figs. 2.2a and 2.2b for the lattice dimension $l \approx 675\,\mathrm{nm}$.

The performance of each method for the estimation of the coupling coefficients depends on the considered system. For example, r_2 and t_2 estimate better the coupling efficiencies than the other methods when the system made of silicon cylinders is considered, but r_1, r_4, r_P, and r_S estimate the reflection efficiency the best for the system made of glass cylinders. However, in systems where the transmission efficiency is close to unity, the estimations of the transmission efficiency t_2, t_3, and t_M should be always better than the other methods because the estimated coupling coefficients satisfy equation (2.44). Hence, the relative error of the estimation on the transmission efficiency has to be much smaller than the relative error on the reflection efficiency.

It is expected that the error of the estimations for the system made of silicon cylinders is several times larger than the estimation for the system made of glass cylinders. Since silicon is a high refractive index material, the refractive index variation inside the right medium is high. Therefore, the mode R differs significantly from a plane wave as shown in figs. 2.1b and 2.1c. In that case, the assumption that the coupling with others modes is negligible may not hold, which leads to the important error present in figs. 2.2a and 2.2b. When the system made of glass cylinders is considered, the variation of the refractive index is much smaller, which means that the mode R looks more like a plane wave as shown in fig. 2.1d. Therefore, the error in the estimation of the reflection and transmission coefficients is significantly smaller (figs. 2.2c and 2.2d).

In summary, the estimation of the coupling coefficients using the method provided in [105,106] is a subset of the estimation proposed in this work for the system presented in fig. 2.1a because the modes L and R are propagating. In terms of performance, the estimations (r_2, t_2) and (r_3, t_3) give a more accurate estimation of the transmission efficiency as shown in figs. 2.2b and 2.2d because the coupling coefficients satisfy equation (2.44). This is valid only when the transmission efficiency is high. Finally, the existence of multiple formula for the estimation of the coupling coefficients can give an indication on the accuracy of such estimations without the need to compute rigorously the coupling coefficients.

2.7.2 Fresnel coefficients for uniaxial media

In practice, the reformulation of the boundary condition presented in section 2.5 can be used in similar way to [106], where a small set of modes is considered on both sides of an interface and an approximation of the reflection and transmission coefficients is obtained. In this section, the Fresnel coefficients at an interface between homogeneous uniaxial media where the extraordinary axis are perpendicular to the interface, are derived. The objective is to show that the equations obtained in section 2.7.1 can be applied as it is formulated for a different case, and to get the expression of the Fresnel coefficients using the convention for the description of plane waves applied in the modified Fourier Modal Method presented in this work.

The first step is to describe the system and the modes propagating in it. We consider an interface at $z = 0$ between two uniaxial media where the extraordinary axis is normal to the interface. The permittivity and the permeability are given by

$$\hat{\epsilon}_m = \begin{pmatrix} \epsilon_{tm} & 0 & 0 \\ 0 & \epsilon_{tm} & 0 \\ 0 & 0 & \epsilon_{zm} \end{pmatrix}, \qquad \hat{\mu}_m = \begin{pmatrix} \mu_{tm} & 0 & 0 \\ 0 & \mu_{tm} & 0 \\ 0 & 0 & \mu_{zm} \end{pmatrix}, \tag{2.50}$$

where $m = 1$ for the medium at $z < 0$ and $m = 2$ for the medium at $z > 0$. Since the permittivity and the permeability have the form (2.24), both media are ZSI. Hence, a mode and its backward-propagating counterpart are related by the minus operator. Moreover, TE-modes and TM-modes are present. The TM-modes can be expressed as

$$\vec{E} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -R_{\epsilon m} k_{\parallel} \end{pmatrix}, \qquad \vec{H} = \begin{pmatrix} -\epsilon_{tm} k_0 s_y \\ \epsilon_{tm} k_0 s_x \\ 0 \end{pmatrix} \tag{2.51}$$

with the dispersion relation

$$R_{\epsilon m} k_{\parallel}^2 + k_{zm}^2 = \epsilon_{tm} \mu_{tm} k_0^2, \tag{2.52}$$

and the TE-modes as

$$\vec{E} = \begin{pmatrix} \mu_{tm} k_0 s_y \\ -\mu_{tm} k_0 s_x \\ 0 \end{pmatrix}, \qquad \vec{H} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -R_{\mu m} k_{\parallel} \end{pmatrix} \tag{2.53}$$

with the dispersion relation

$$R_{\mu m} k_{\parallel}^2 + k_{zm}^2 = \epsilon_{tm} \mu_{tm} k_0^2. \tag{2.54}$$

The implicit time and spatial dependance is $e^{i(\vec{k}_m \vec{x} - \omega t)}$, where $\vec{k}_m$ is given by

$$\vec{k}_m = \begin{pmatrix} k_x \\ k_y \\ k_{zm} \end{pmatrix} = \begin{pmatrix} k_{/\!/} s_x \\ k_{/\!/} s_y \\ k_{zm} \end{pmatrix} \tag{2.55}$$

with

$$k_{/\!/} = \sqrt{k_x^2 + k_y^2}, \qquad \begin{cases} s_x = 1,\ s_y = 0 & \text{if } k_{/\!/} = 0 \\ s_x = k_x / k_{/\!/},\ s_y = k_y / k_{/\!/} & \text{otherwise} \end{cases} \tag{2.56}$$

$R_{\epsilon m}$ and $R_{\mu m}$ are defined as

$$R_{\epsilon m} := \frac{\epsilon_{tm}}{\epsilon_{zm}}, \qquad R_{\mu m} := \frac{\mu_{tm}}{\mu_{zm}}. \tag{2.57}$$

The modes propagating in the medium 1 are called L_{TE} and L_{TM} and the modes propagating in the medium 2 are called R_{TE} and R_{TM}.

The second step is to define a Poynting operation. For homogenous media, a natural choice is

$$[\psi_1|\psi_2] := \lim_{S\to\mathbb{R}^2} \frac{1}{2|S|} \iint_S (\vec{E}_1 \times \vec{H}_2^* + \vec{E}_2 \times \vec{H}_1^*) \cdot \vec{n} ds. \tag{2.58}$$

With this Poynting operation, the TE and TM-modes propagating in the same medium are orthogonal with each other even for lossy media. Moreover, a mode L_{TE} is orthogonal to every R modes except the R_{TE} mode with the same k_x and k_y. Same can be said for the L_{TM} and L_{TE} modes.

The final step is to write the system of equations and solve it. Since the modes are mostly orthogonal and the system is illuminated by a single L mode with amplitude of one, the expressions (2.28) become the expression (2.42) in section 2.7.1 with the difference that no approximation has been done in this case. Hence, all the solutions presented in (2.43) are equivalent and, for the rest of this section, the solution (2.43b) is used:

$$r = \frac{|[L|R+R^-]|^2 - [L|L][R|R]^*}{|[L|R+R^-]|^2 + [L|L][R|R]^*}, \qquad t = \frac{2[L|L][L|R+R^-]^*}{|[L|R+R^-]|^2 + [L|L][R|R]^*}. \tag{2.59}$$

As said before, this solution is valid for any system where one mode excites at the interface the backward-propagating counterpart of the exciting mode and a single mode in the second medium. When L is a TM mode, the different terms present in the solution (2.59) become

$$\begin{aligned}
[L|L] &= \bar{\epsilon}_{t1} k_{z1} k_0, \\
[R|R] &= \bar{\epsilon}_{t2} k_{z2} k_0, \\
[L|R + R^{-}] &= \bar{\epsilon}_{t1} k_{z2} k_0.
\end{aligned} \tag{2.60}$$

Combining (2.59) and (2.60), the Fresnel coefficients r and t are obtained:

$$r = \frac{\epsilon_{t1} k_{z2} - \epsilon_{t2} k_{z1}}{\epsilon_{t1} k_{z2} + \epsilon_{t2} k_{z1}}, \qquad t = \frac{2\epsilon_{t1} k_{z1}}{\epsilon_{t1} k_{z2} + \epsilon_{t2} k_{z1}}. \tag{2.61}$$

In the TE case, the different terms are

$$\begin{aligned}
[L|L] &= \mu_{t1} \bar{k}_{z1} k_0, \\
[R|R] &= \mu_{t2} \bar{k}_{z2} k_0, \\
[L|R + R^{-}] &= \mu_{t2} \bar{k}_{z1} k_0,
\end{aligned} \tag{2.62}$$

and the Fresnel coefficients become

$$r = \frac{\mu_{t2} k_{z1} - \mu_{t1} k_{z2}}{\mu_{t2} k_{z1} + \mu_{t1} k_{z2}}, \qquad t = \frac{2\mu_{t1} k_{z1}}{\mu_{t2} k_{z1} + \mu_{t1} k_{z2}}. \tag{2.63}$$

Using the reformulation of the boundary condition, the coefficients r and t are expressed independently of the description of the modes itself. Moreover, only the expressions Q_L and Q_R are used to find the solution in this example, but one can check that the expressions Q_R and R_L are equal to zero since no approximation has been done. In other words, if approximations are introduced, the value of the expressions (2.42) gives an indication of the validity of the approximation and, at the same time, the partial derivatives of $[E|E]$.

2.8 Operations on the Gram matrix: Applications

2.8.1 Orthogonality of plane waves in lossy homogenous isotropic media

In lossless isotropic media, for any plane wave with a given propagation constant k_z, it exists another plane wave with the same propagation which is orthogonal to the first one. In this section, this statement is proven along with the demonstation that it is no more the case when the medium is lossless and the propagation direction is not parallel to the z-axis ($k_{\parallel} \neq 0$). Following the convention for plane waves used throughout this work, TM plane wave is given by

$$\vec{E} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -k_{\parallel} \end{pmatrix}, \qquad \vec{H} = \begin{pmatrix} -\epsilon k_0 s_y \\ \epsilon k_0 s_x \\ 0 \end{pmatrix} \tag{2.64}$$

and named ψ_{TM}, and TE plane wave is given by

$$\vec{E} = \begin{pmatrix} \mu k_0 s_y \\ -\mu k_0 s_x \\ 0 \end{pmatrix}, \qquad \vec{H} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -k_{\parallel} \end{pmatrix} \tag{2.65}$$

and named ψ_{TE}. The dispersion relation is the same for both polarizations and is given by the well known equation

$$k_{\parallel}^2 + k_z^2 = \epsilon\mu k_0^2, \tag{2.66}$$

For the vector $\vec{\psi}_0 = (\psi_{TM}, \psi_{TE})$, the associated Gram matrix is

$$\hat{G}_0 = \begin{pmatrix} \bar{\epsilon} k_z k_0 & 0 \\ 0 & \mu \bar{k}_z k_0 \end{pmatrix} = \begin{pmatrix} p_{TM} & 0 \\ 0 & p_{TE} \end{pmatrix}. \tag{2.67}$$

Since the gram matrix $\hat{G}_0$ is diagonal, TM and TE plane waves are orthogonal between each other whether the medium is lossy or lossless. Let assume that two other plane waves given by the vector $\vec{\psi}_1$ are also orthogonal between each other and is given by

$$\vec{\psi}_1 = \hat{R}\vec{\psi}_0, \qquad \hat{R} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{2.68}$$

The Gram matrix $\hat{G}_1$ associated to the vector $\vec{\psi}_1$ can be directly obtained from equation (2.38) by noticing that equation (2.68) is the equation (2.36) with $\hat{R} = \hat{A} = \hat{B}$. Hence, the Gram matrix $\hat{G}_1$ is given by

$$\hat{G}_1 = \hat{R}^* G_0 R^T = \begin{pmatrix} |a|^2 p_{TM} + |b|^2 p_{TE} & \bar{a} c p_{TM} + \bar{b} d p_{TE} \\ a \bar{c} p_{TM} + b \bar{d} p_{TE} & |c|^2 p_{TM} + |d|^2 p_{TE} \end{pmatrix} \tag{2.69}$$

Therefore, the two modes in $\vec{\psi}_1$ are orthogonal with each other if the off-diagonal elements of the Gram matrix $\hat{G}_1$ are zero, meaning that

$$\begin{aligned} a\bar{c}\bar{p}_{TM} + b\bar{d}\bar{p}_{TE} &= 0, \\ a\bar{c}p_{TM} + b\bar{d}p_{TE} &= 0. \end{aligned} \tag{2.70}$$

The system of equations (2.70) is satsified in different cases. If $a = b = 0$ or $c = d = 0$, it means that one of the mode in $\vec{\psi}_1$ is null and it is a trivial case since a null mode is orthogonal to any modes. If $a = d = 0$ or $c = b = 0$, $\vec{\psi}_1$ also contains the TM and TE mode. The last case is the case where the determinant of the system of equations (2.70) is zero, meaning that

$$\bar{p}_{TM}p_{TE} - p_{TM}\bar{p}_{TE} = \mathrm{Im}\{\bar{p}_{TM}p_{TE}\} = 0. \tag{2.71}$$

By replacing p_{TM} and p_{TE} by their corresponding expression (see equation (2.67)) and using the dispersion relation (2.66),the determinant of the system of equations (2.70) becomes

$$\begin{aligned} &\mathrm{Im}\left\{\epsilon\mu\bar{k}_z^2k_0^2\right\} = \mathrm{Im}\left\{\left(|\epsilon\mu|^2k_0^2 - \epsilon\mu k_{/\!/}^2\right)k_0^2\right\} = 0 \\ \Rightarrow\ &\mathrm{Im}\{\epsilon\mu\}k_{/\!/}^2 = 0 \end{aligned} \tag{2.72}$$

Hence, the determinant of the system of equations (2.70) is zero when the plane wave comes at normal incidence or when $\mathrm{Im}\{\epsilon\mu\}$ is zero, which is the case for lossless media but not for lossy media.

2.8.2 Orthonormalization of propagating and evanescent mode

In a lossless medium, three types of mode are present, namely propagating, evanescent, and complex modes. If the propagation constant is purely real, the mode is propagating. If it is purely imaginary, the mode is evanescent and, if it has a real and an imaginary part, the mode is complex. In this section and in section 2.8.3, a ZSI lossless medium which is invariant to a 90°-rotation around the z-axis is considered. The consequence of this symmetry is that, for most modes, there is another mode with the same propagation constant. For propagating and evanescent modes, those pair of modes with the same propagation constant may not be orthogonal. In this section, the property of the Gram matrix of propagating or evanescent mode pair is given along with their orthonormalizaion. Complex modes are treated in section 2.8.3.

The property of the Gram matrix of a pair of propagating or evanescent modes, called ψ_1 and ψ_2, is obtained from the conservation of power along the z direction. Let us have two other modes, ψ_F and ψ_B, which are a combination of the modes ψ_1 and ψ_2, meaning that

$$\begin{aligned} \psi_F &= f_1\psi_1 + f_2\psi_2 \\ \psi_B &= b_1\psi_1 + b_2\psi_2, \end{aligned} \tag{2.73}$$

where f_1, f_2, b_1 and b_2 are arbitrary complex numbers. ψ_F are the modes that propagate in the forward direction and ψ_B^- are the modes that propagate in the backward direction. If ψ_1 and ψ_2 are propagating modes with propagation constant γ, the power flow at any position z is given by

$$\begin{aligned} &\mathrm{Re}\{[e^{i\gamma z}\psi_F + e^{-i\gamma z}\psi_B^- | e^{i\gamma z}\psi_F + e^{-i\gamma z}\psi_B^-]\} \\ &\quad = \mathrm{Re}\{[\psi_F|\psi_F]\} - \mathrm{Re}\{[\psi_B|\psi_B]\} + 2\cos\big(2\gamma z\big)\,\mathrm{Re}\{[\psi_F|\psi_B^-]\} + 2\sin\big(2\gamma z\big)\,\mathrm{Im}\{[\psi_F|\psi_B]\}. \end{aligned} \tag{2.74}$$

Since the power flow is constant along the z direction, the following two equations are obtained:

$$\begin{aligned} \mathrm{Re}\{[\psi_F|\psi_B^-]\} &= \mathrm{Re}\{[f_1\psi_1 + f_2\psi_2 | b_1\psi_1^- + b_2\psi_2^-]\} = 0 \\ \mathrm{Im}\{[\psi_F|\psi_B]\} &= \mathrm{Im}\{[f_1\psi_1 + f_2\psi_2 | b_1\psi_1 + b_2\psi_2]\} = 0 \end{aligned} \tag{2.75}$$

Since f_1, f_2, b_1 and b_2 can be any complex numbers, the Poynting operation applied to a pair of propagating mode satisfies the following conditions:

$$\begin{aligned} &\mathrm{Im}\{[\psi_1|\psi_1]\} = 0 \qquad \mathrm{Im}\{[\psi_2|\psi_2]\} = 0 \\ &\mathrm{Im}\{[\psi_1|\psi_2]\} = 0 \qquad \mathrm{Re}\{[\psi_1|\psi_2^-]\} = 0. \end{aligned} \tag{2.76}$$

Since the Gram matrix $\hat{G}$ is defined as

$$G_{mn} = [\psi_m|\psi_n + \psi_n^-] \tag{2.77}$$

and, due to properties (2.12a) and (2.15a),

$$G_{nm} = [\psi_m|\psi_n - \psi_n^-], \tag{2.78}$$

the Gram matrix $\hat{G}_p$ of a pair of propagating modes has the form

$$\hat{G}_p = \begin{pmatrix} z_{11} & z_{12} \\ \bar{z}_{12} & z_{22} \end{pmatrix}, \tag{2.79}$$

where the diagonal elements z_{11} and z_{22} are purely real. Therefore, ψ_1 and ψ_2 are orthonormalized when the Gram matrix $\hat{G}_{p,o}$ is the identity matrix since the diagonal element must be real and the real part of the Poynting operation applied on a forward-propagating mode with himself is, by definition, positive.

If ψ_1 and ψ_2 are evanescent modes with propagation constant $i\gamma$, the power flow at any position z is given by

$$\begin{aligned}&\mathrm{Re}\{[e^{-\gamma z}\psi_F + e^{\gamma z}\psi_B^- | e^{-\gamma z}\psi_F + e^{\gamma z}\psi_B^-]\}\\ &\quad = e^{-2\gamma z}\,\mathrm{Re}\{[\psi_F|\psi_F]\} - e^{2\gamma z}\,\mathrm{Re}\{[\psi_B|\psi_B]\} + 2\,\mathrm{Re}\{[\psi_F|\psi_B^-]\}\}.\end{aligned} \tag{2.80}$$

The equations that must be satisfied in order to have a constant power flow along the z direction are

$$\begin{aligned}\mathrm{Re}\{[\psi_F|\psi_F]\} &= \mathrm{Re}\{[f_1\psi_1 + f_2\psi_2 | f_1\psi_1 + f_2\psi_2]\} \;\; = 0\\ \mathrm{Re}\{[\psi_B|\psi_B]\} &= \mathrm{Re}\{[b_1\psi_1 + b_2\psi_2 | b_1\psi_1 + b_2\psi_2]\} = 0\end{aligned} \tag{2.81}$$

Therefore, the Poynting operation applied to a pair of evanescent modes satisfies the following condition:

$$\begin{aligned}&\mathrm{Re}\{[\psi_1|\psi_1]\} = 0 \qquad \mathrm{Re}\{[\psi_2|\psi_2]\} \;\; = 0\\ &\mathrm{Re}\{[\psi_1|\psi_2]\} = 0 \qquad \mathrm{Im}\{[\psi_1|\psi_2^-]\} = 0.\end{aligned} \tag{2.82}$$

Hence, the Gram matrix $\hat{G}_e$ of a pair of evanescent modes is

$$\hat{G}_e = \begin{pmatrix} z_{11} & z_{12} \\ -\bar{z}_{12} & z_{22} \end{pmatrix}, \tag{2.83}$$

where the diagonal elements z_{11} and z_{22} are purely imaginary. The modes ψ_1 and ψ_2 are orthonormalized when the Gram matrix $\hat{G}_{e,o}$ is diagonal and its diagonal elements are $\pm i$.

In order to orthonormalize propagating and evanescent modes, the mode operations shown in table 2.1 are used. In general, the Gram matrix $\hat{G}_1$ for a pair of modes ψ_1 and ψ_2 is given by

$$\hat{G}_1 = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{pmatrix} \tag{2.84}$$

and the objective is to find the matrix $\hat{A}$ such that the orthonormalized modes $\vec{\psi}_o$ are given by

$$\vec{\psi}_o = \hat{A}\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} \tag{2.85}$$

First, the mode operations “Mode composition 1” and “Mode composition 2” are used to orthogonalize the modes:

$$\psi_2 \longleftarrow \psi_2 - \frac{z_{12}}{z_{11}}\psi_1,\ \hat{A}_2 = \begin{pmatrix} 1 & 0 \\ -\frac{z_{12}}{z_{11}} & 1 \end{pmatrix},\ \hat{G}_2 = \begin{pmatrix} z_{11} & 0 \\ z_{21} - \frac{\bar{z}_{12} z_{11}}{\bar{z}_{11}} & z_{22} - \frac{z_{21} z_{12}}{z_{11}} \end{pmatrix}. \tag{2.86}$$

For both propagating and evanescent modes, the off-diagonal element of the Gram matrix $\hat{G}_2$, given by

$$z_{21} - \frac{\bar{z}_{12} z_{11}}{\bar{z}_{11}}, \tag{2.87}$$

is zero, making the Gram matrix $\hat{G}_2$ diagonal. This is due to the property of the Gram matrix shown in (2.79) and (2.83): $z_{21} = \bar{z}_{12}$ and z_{11} is purely real for propagating modes, and $z_{21} = -\bar{z}_{12}$ and z_{11} is purely imaginary for evanescent modes. The normalization is done by applying the mode operation "Mode scaling":

$$\begin{gathered} \psi_1 \longleftarrow \sqrt{\frac{1}{|z_{11}|}}\psi_1 \\ \psi_2 \longleftarrow \sqrt{\frac{|z_{11}|}{|z_{11}z_{22} - z_{21}z_{12}|}}\psi_2 \\ \hat{A} = \begin{pmatrix} \sqrt{\frac{1}{|z_{11}|}} & 0 \\ -\sqrt{\frac{|z_{11}|}{|z_{11}z_{22} - z_{21}z_{12}|}}\frac{z_{12}}{z_{11}} & \sqrt{\frac{|z_{11}|}{|z_{11}z_{22} - z_{21}z_{12}|}} \end{pmatrix},\ \hat{G} = \begin{pmatrix} \frac{z_{11}}{|z_{11}|} & 0 \\ 0 & \frac{(z_{11}z_{22} - z_{21}z_{12})|z_{11}|}{z_{11}|z_{11}z_{22} - z_{21}z_{12}|} \end{pmatrix}. \end{gathered} \tag{2.88}$$

The matrix $\hat{A}$ in (2.88) allows to orthonormalize a pair of propagating or evanescent modes by using equation (2.85). If z_{11} is close to zero, preforming the mode operation "Mode swapping" before the mode operation shown in (2.86) may improve the stability of the orthonormalization. If the Gram matrix of a pair of propagating modes after orthonormalization has negative diagonal elements, the mode operation "Mode reversal" should be used in order to fulfill the definition of forward-propagating mode. In this case, the sign of the propagation constant is also changed.

2.8.3 Orthonormalization of complex modes

In this section, the orthonormalization procedure for complex modes, modes whose propagation constant has a real and an imaginary part, is shown. From the orthogonality relation (2.25) in section 2.4, complex modes have the property to be self-orthogonal. However, two complex modes with propagation constant γ and $\bar{\gamma}$ are not orthogonal. As proved in Chapter 7 of [99], for any mode with a complex propagation constant γ, there exists a mode with the propagation constant $\bar{\gamma}$. Therefore, in a ZSI lossless medium which is invariant to a 90°-rotation around the z-axis, the orthonormalization procedure involves four modes, two modes with propagation constant γ, called ψ_1 and ψ_3, and two modes with propagation constant $-\bar{\gamma}$, called $\psi_2 = \psi_{C1}$ and $\psi_4 = \psi_{C3}$. The sign of the propagation constant is chosen such that the amplitude of the mode decreases with z.

The Gram matrix for a group of four complex modes has a set of properties. First, the orthogonality relation (2.25) leads to

$$\begin{aligned} &[\psi_m|\psi_n] = 0, \qquad &[\psi_{Cm}|\psi_{Cn}] = 0,\\ &[\psi_m|\psi_n^-] = 0, \qquad &[\psi_{Cm}|\psi_{Cn}^-] = 0, \end{aligned} \tag{2.89}$$

where $m, n \in \{1, 3\}$. Second, the power flow is constant along the z direction. Let us have four other modes, ψ_F, ψ_{CF}, ψ_B and ψ_{CB}, which are a combination of the four complex modes ψ_m with $m \in [1, 4]$:

$$\begin{aligned} &\psi_F = f_1\psi_1 + f_3\psi_3 \qquad &\psi_{CF} = f_2\psi_{C1} + f_4\psi_{C3}\\ &\psi_B = b_1\psi_1 + b_3\psi_3 \qquad &\psi_{CB} = b_2\psi_{C1} + b_4\psi_{C3}, \end{aligned} \tag{2.90}$$

where f_m and b_m are arbitrary complex numbers. The power flow at any position z is given by

$$\begin{aligned} &\frac{1}{2}\mathrm{Re}\{[e^{i\gamma z}\psi_F + e^{-i\bar{\gamma} z}\psi_{CF} + e^{-i\gamma z}\psi_B^- + e^{i\bar{\gamma} z}\psi_{CB}^- | e^{i\gamma z}\psi_F + e^{-i\bar{\gamma} z}\psi_{CF} + e^{-i\gamma z}\psi_B^- + e^{i\bar{\gamma} z}\psi_{CB}^-]\}\\ &\quad = \mathrm{Re}\{e^{-2i\bar{\gamma} z}\}\,\mathrm{Re}\{[\psi_F|\psi_{CF}]\} - i\,\mathrm{Im}\{e^{-2i\bar{\gamma} z}\}\,\mathrm{Im}\{[\psi_F|\psi_{CF}^-]\} + \mathrm{Re}\{[\psi_F|\psi_{CF}]\}\\ &\quad + \mathrm{Re}\{[\psi_{CF}|\psi_B^-]\} - \mathrm{Re}\{e^{2i\bar{\gamma} z}\}\,\mathrm{Re}\{[\psi_B|\psi_{CB}]\} + \mathrm{Im}\{e^{2i\bar{\gamma} z}\}\,\mathrm{Im}\{[\psi_B|\psi_{CB}^-]\}. \end{aligned} \tag{2.91}$$

The power flow is z independent if the following quantities are zero:

$$\begin{aligned} &\mathrm{Re}\{[\psi_F|\psi_{CF}]\} = 0 \qquad &\mathrm{Im}\{[\psi_F|\psi_{CF}^-]\} = 0\\ &\mathrm{Re}\{[\psi_B|\psi_{CB}]\} = 0 \qquad &\mathrm{Im}\{[\psi_B|\psi_{CB}^-]\} = 0. \end{aligned} \tag{2.92}$$

After replacing ψ_F, ψ_{CF}, ψ_B and ψ_{CB} by their expressions in (2.90) and using the property (2.15d), the expressions in (2.92) are equal to zero if

$$\begin{aligned} &\mathrm{Re}\{[\psi_m|\psi_{Cn}]\} = 0\\ &\mathrm{Im}\{[\psi_m|\psi_{Cn}^-]\} = 0, \; m, n \in \{1, 3\}. \end{aligned} \tag{2.93}$$

Since $[\psi_m|\psi_{Cn}]$ is purely imaginary and $[\psi_m|\psi_{Cn}]$ is purely real along with the definition of the Gram matrix (2.77) and its property (2.78), the Gram matrix of a group of four complex modes has the form

$$\hat{G}_c = \begin{pmatrix} 0 & z_{12} & 0 & z_{14} \\ -\bar{z}_{12} & 0 & z_{23} & 0 \\ 0 & -\bar{z}_{23} & 0 & z_{34} \\ -\bar{z}_{14} & 0 & -\bar{z}_{34} & 0 \end{pmatrix}. \tag{2.94}$$

For the orthonormalization, we choose that two modes with different propagation constant cannot be combined, meaning that the Gram matrix $\hat{G}_{co}$ after orthonormalization cannot be diagonal because the self-orthogonality of complex modes must still hold. Therefore, it is not a standard orthonormalization and the Gram matrix $\hat{G}_{co}$ after orthonormalization is

$$\hat{G}_{co} = \begin{pmatrix} 0 & i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & i \\ 0 & 0 & i & 0 \end{pmatrix}. \tag{2.95}$$

To transform $\hat{G}_c$ into $\hat{G}_{co}$, the mode operations shown in table 2.1 are used and an example of orthonormalization is shown in table 2.2. To simplify the expression present in table 2.2, the following quantities are introduced:

$$y := z_{12}z_{34} + z_{14}\bar{z}_{23}, \qquad c_1 := \sqrt{|z_{12}|}, \qquad c_3 := \sqrt{\frac{|y|}{|z_{12}|}}. \tag{2.96}$$

In the orthogonalization procedure, accuracy issue may arise if z_{12} or y are close to zero. For the case where $|z_{12}|$ is close to zero or, in general, when $|z_{34}|$ is larger than $|z_{12}|$, swapping the mode ψ_1 and ψ_2 with, respectively, ψ_3 and ψ_4 beforehand improves the accuracy. For the case where y tends to zero, it can be shown that, after orthogonalization, ψ_1 and ψ_{C1} (or ψ_3 and ψ_{C3}) tend to be orthogonal. Hence, after the normalization, the amplitude of, at least, one of the mode tends to infinity.

From table 2.2, the relationship between the initial modes $\vec{\psi}_i$ and the orthonormalized modes $\vec{\psi}_o$ is

$$\vec{\psi}_o = \hat{A}\vec{\psi}_i, \tag{2.97}$$

Mode operation	$\hat{A} = \hat{B}$	$\hat{G}$
$\psi_3 \longleftarrow \psi_3 + \frac{z_{23}}{\bar{z}_{12}} \psi_1$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{z_{23}}{\bar{z}_{12}} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & z_{12} & 0 & z_{14} \\ -\bar{z}_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{y}{z_{12}} \\ -\bar{z}_{14} & 0 & -\frac{\bar{y}}{\bar{z}_{12}} & 0 \end{pmatrix}$
$\psi_4 \longleftarrow \psi_4 - \frac{z_{14}}{z_{12}} \psi_2$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{z_{23}}{\bar{z}_{12}} & 0 & 1 & 0 \\ 0 & -\frac{z_{14}}{z_{12}} & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & z_{12} & 0 & 0 \\ -\bar{z}_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{y}{z_{12}} \\ 0 & 0 & -\frac{\bar{y}}{\bar{z}_{12}} & 0 \end{pmatrix}$
$\psi_1 \longleftarrow \frac{1}{c_1} \psi_1$ $\psi_2 \longleftarrow \frac{i\bar{c}_1}{z_{12}} \psi_2$	$\begin{pmatrix} \frac{1}{c_1} & 0 & 0 & 0 \\ 0 & \frac{i\bar{c}_1}{z_{12}} & 0 & 0 \\ \frac{z_{23}}{\bar{z}_{12}} & 0 & 1 & 0 \\ 0 & -\frac{z_{14}}{z_{12}} & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{y}{z_{12}} \\ 0 & 0 & -\frac{\bar{y}}{\bar{z}_{12}} & 0 \end{pmatrix}$
$\psi_3 \longleftarrow \frac{1}{c_3} \psi_3$ $\psi_4 \longleftarrow \frac{i\bar{c}_3 z_{12}}{y} \psi_4$	$\begin{pmatrix} \frac{1}{\bar{c}_1} & 0 & 0 & 0 \\ 0 & \frac{ic_1}{z_{12}} & 0 & 0 \\ \frac{z_{23}}{\bar{z}_{12} c_3} & 0 & \frac{1}{c_3} & 0 \\ 0 & -\frac{iz_{14}\bar{c}_3}{y} & 0 & \frac{i\bar{c}_3 z_{12}}{y} \end{pmatrix}$	$\begin{pmatrix} 0 & i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & i \\ 0 & 0 & i & 0 \end{pmatrix}$

Table 2.2 – All the steps that transform $\hat{G}_c$ into $\hat{G}_{co}$ and their consequence on the matrices $\hat{A}$ and $\hat{B}$. c_1 and c_3 can be any quantity other than zero but they have been chosen such that both modes in the same group (ψ_1, ψ_2 or ψ_3, ψ_4) are multiplied by a constant with the same amplitude. In this table, "Mode composition 1" and "Mode composition 2" are combined.

with

$$\hat{A} = \begin{pmatrix} \frac{1}{\sqrt{|z_{12}|}} & 0 & 0 & 0 \\ 0 & i\frac{\sqrt{|z_{12}|}}{z_{12}} & 0 & 0 \\ \frac{z_{23}}{\bar{z}_{12}} \sqrt{\frac{|z_{12}|}{|y|}} & 0 & \sqrt{\frac{|z_{12}|}{|y|}} & 0 \\ 0 & -i\frac{z_{14}}{y}\sqrt{\frac{|y|}{|z_{12}|}} & 0 & i\frac{z_{12}}{y}\sqrt{\frac{|y|}{|z_{12}|}} \end{pmatrix}. \tag{2.98}$$

If the modes ψ_1 and ψ_2 are swapped with, respectively, the modes ψ_3 and ψ_4 at the beginning of the orthonomalization procedure, a similar orthonormalization can be done and the following matrix $\hat{A}$ is obtained:

$$\hat{A} = \begin{pmatrix} 0 & 0 & \frac{1}{\sqrt{|z_{34}|}} & 0 \\ 0 & 0 & 0 & i\frac{\sqrt{|z_{34}|}}{z_{34}} \\ \sqrt{\frac{|z_{34}|}{|y|}} & 0 & -\frac{\bar{z}_{14}}{\bar{z}_{34}}\sqrt{\frac{|z_{34}|}{|y|}} & 0 \\ 0 & i\frac{z_{34}}{y}\sqrt{\frac{|y|}{|z_{34}|}} & 0 & i\frac{\bar{z}_{23}}{y}\sqrt{\frac{|y|}{|z_{34}|}} \end{pmatrix}. \tag{2.99}$$

2.8.4 Rotation of propagating, evanescent and complex modes

In this section, a mode rotation is an operator given by a rotation matrix $\hat{R}$ that transforms a set of orthonormalized mode $\vec{\psi}_o$ into another set of orthonormalized mode $\vec{\psi}_r = \hat{R}\vec{\psi}_o$. As in sections 2.8.2 and 2.8.3, a ZSI lossless medium which is invariant to a 90°-rotation around the z-axis is considered here. A typical application is the case where an illumination excites two modes in a pair of propagating or evanescent modes (or all the four modes in a group of complex modes), and the modes are rotated such that the illumination excites only a single mode in a pair or two modes with different propagation constant in a group of four complex modes. In other words, the upper half of the rotation matrix $\hat{R}$ is known up to a scaling factor per lines, which ensure that the modes are normalized, and the objective is to obtain the other half. For propagating and evanescent modes, the equation (2.69) in section 2.8.1 is used:

$$\hat{G}_r = \hat{R}^* G_o R^T = \begin{pmatrix} |a|^2 p_1 + |b|^2 p_2 & \bar{a} c p_1 + \bar{b} d p_2 \\ a\bar{c} p_1 + b\bar{d} p_2 & |c|^2 p_1 + |d|^2 p_2 \end{pmatrix}, \tag{2.100}$$

where $\hat{G}_o$ is the Gramm matrix before rotation, which is diagonal and whose diagonal elements are p_1 and p_2, $\hat{G}_r$ is the Gram matrix after rotation and the rotation matrix $\hat{R}$ is given by

$$\hat{R} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{2.101}$$

The phase of p_1 and p_2 is directly related to the mode type and, for symmetry reason, two orthonormalized modes with the same propagation constant are of the same type, leading to p_1 and p_2 being equal. Therefore, the modes $\vec{\psi}_r$ are orthonormalized if the following equations are satisfied:

$$\begin{aligned} |a|^2 + |b|^2 &= 1 \\ |c|^2 + |d|^2 &= 1 \\ a\bar{c} + b\bar{d} &= 0. \end{aligned} \tag{2.102}$$

As mentioned earlier, the first line of the rotation matrix $\hat{R}$ is known up to a scaling factor, meaning that the first mode of the set $\vec{\psi}_r$, denoted ψ_{r1}, is given by

$$\psi_{r1} = k(q_1\psi_{o1} + q_2\psi_{o2}), \tag{2.103}$$

where q_1 and q_2 are given with the condition that $|q_1|^2 + |q_2|^2 \neq 0$, k is the scaling factor, and ψ_{o1} and ψ_{o2} are the modes in $\vec{\psi}_o$. From equation (2.102), the scaling factor k is

$$k = \frac{1}{\sqrt{|q_1|^2 + |q_2|^2}}, \tag{2.104}$$

and the rotation matrix $\hat{R}$ and its inverse become

$$\hat{R} = \frac{1}{\sqrt{|q_1|^2 + |q_2|^2}} \begin{pmatrix} q_1 & q_2 \\ -\bar{q}_2 & \bar{q}_1 \end{pmatrix} \qquad \hat{R}^{-1} = \frac{1}{\sqrt{|q_1|^2 + |q_2|^2}} \begin{pmatrix} \bar{q}_1 & -q_2 \\ \bar{q}_2 & q_1 \end{pmatrix}. \tag{2.105}$$

The matrix $\hat{R}$ is unitary and the coefficient has been chosen such that the determinant of $\hat{R}$ is one. For real coefficients, $\hat{R}$ is a rotation matrix, but, in general, $\hat{R}$ belongs to the special unitary group $SU(2)$.

For complex modes, the objective to find the matrix $\hat{R}$ that transforms the orthonormalized modes $\vec{\psi}_o$ into $\vec{\psi}_r$ such that

$$\begin{aligned} \psi_{r1} &= k_1 (q_1 \psi_{o1} + q_3 \psi_{o3}) \\ \psi_{r2} &= k_2 (q_2 \psi_{o2} + q_4 \psi_{o4}), \end{aligned} \tag{2.106}$$

where q_1, q_2, q_3 and q_4 are given. In order to get the matrix $\hat{R}$, the mode operations shown in table 2.3 are performed, where the following quantities are defined as:

$$y := \bar{q}_1 q_2 + \bar{q}_3 q_4 \qquad c := \sqrt{|q_1|^2 + |q_3|^2}. \tag{2.107}$$

The parameter c has been chosen such that the norm of the complex vector given by the first row of the matrix $\hat{R}$ is one. From table 2.3, the matrix $\hat{R}$ and its inverse are

$$\begin{aligned} \hat{R} &= \frac{1}{c} \begin{pmatrix} q_1 & 0 & q_3 & 0 \\ 0 & \frac{|c|^2 q_2}{y} & 0 & \frac{|c|^2 q_4}{y} \\ -\frac{|c|^2 \bar{q}_4}{\bar{y}} & 0 & \frac{|c|^2 \bar{q}_2}{\bar{y}} & 0 \\ 0 & -\bar{q}_3 & 0 & \bar{q}_1 \end{pmatrix} \\ \hat{R}^{-1} &= \frac{1}{\bar{c}} \begin{pmatrix} \frac{|c|^2 \bar{q}_2}{\bar{y}} & 0 & -q_3 & 0 \\ 0 & \bar{q}_1 & 0 & -\frac{|c|^2 q_4}{y} \\ \frac{|c|^2 \bar{q}_4}{\bar{y}} & 0 & q_1 & 0 \\ 0 & \bar{q}_3 & 0 & \frac{|c|^2 q_2}{y} \end{pmatrix} \end{aligned} \tag{2.108}$$

Even if the matrices $\hat{R}$ and $\hat{R}^{-1}$ are similar, $\hat{R}$ is not unitary. This is due to the pseudo-orthonormalization of the complex modes $\vec{\psi}_o$ and $\vec{\psi}_r$. In other words, $\hat{R}$ is not unitary because $\hat{G}_o$ and $\hat{G}_r$ are not diagonal.

Mode operation	$\hat{R}$	$\hat{G}$
$\psi_1 \longleftarrow q_1\psi_1$ $\psi_2 \longleftarrow q_2\psi_2$	$\begin{pmatrix} q_1 & 0 & 0 & 0 \\ 0 & q_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & i\bar{q}_1 q_2 & 0 & 0 \\ iq_1\bar{q}_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \\ 0 & 0 & i & 0 \end{pmatrix}$
$\psi_1 \longleftarrow \psi_1 + q_3\psi_3$	$\begin{pmatrix} q_1 & 0 & q_3 & 0 \\ 0 & q_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & i\bar{q}_1 q_2 & 0 & i\bar{q}_3 \\ iq_1\bar{q}_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & i \\ iq_3 & 0 & i & 0 \end{pmatrix}$
$\psi_2 \longleftarrow \psi_2 + q_4\psi_3$	$\begin{pmatrix} q_1 & 0 & q_3 & 0 \\ 0 & q_2 & 0 & q_4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & iy & 0 & i\bar{q}_3 \\ i\bar{y} & 0 & i\bar{q}_4 & 0 \\ 0 & iq_4 & 0 & i \\ iq_3 & 0 & i & 0 \end{pmatrix}$
$\psi_3 \longleftarrow \psi_3 - \frac{\bar{q}_4}{\bar{y}}\psi_1$	$\begin{pmatrix} q_1 & 0 & q_3 & 0 \\ 0 & q_2 & 0 & q_4 \\ -\frac{q_1\bar{q}_4}{\bar{y}} & 0 & \frac{q_1\bar{q}_2}{\bar{y}} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & iy & 0 & i\bar{q}_3 \\ i\bar{y} & 0 & 0 & 0 \\ 0 & 0 & 0 & i\frac{\bar{q}_1 q_2}{y} \\ iq_3 & 0 & i\frac{q_1\bar{q}_2}{\bar{y}} & 0 \end{pmatrix}$
$\psi_4 \longleftarrow \psi_4 - \frac{\bar{q}_3}{y}\psi_1$	$\begin{pmatrix} q_1 & 0 & q_3 & 0 \\ 0 & q_2 & 0 & q_4 \\ -\frac{q_1\bar{q}_4}{\bar{y}} & 0 & \frac{q_1\bar{q}_2}{\bar{y}} & 0 \\ 0 & \frac{q_2\bar{q}_3}{y} & 0 & \frac{\bar{q}_1 q_2}{y} \end{pmatrix}$	$\begin{pmatrix} 0 & iy & 0 & 0 \\ i\bar{y} & 0 & 0 & 0 \\ 0 & 0 & 0 & i\frac{\bar{q}_1 q_2}{y} \\ 0 & 0 & i\frac{q_1\bar{q}_2}{\bar{y}} & 0 \end{pmatrix}$
$\psi_1 \longleftarrow \frac{1}{c}\psi_1$ $\psi_2 \longleftarrow \frac{i\bar{c}}{y}\psi_2$	$\begin{pmatrix} \frac{q_1}{c} & 0 & \frac{q_3}{c} & 0 \\ 0 & \frac{\bar{c}q_2}{y} & 0 & \frac{\bar{c}q_4}{y} \\ -\frac{q_1\bar{q}_4}{\bar{y}} & 0 & \frac{q_1\bar{q}_2}{\bar{y}} & 0 \\ 0 & \frac{q_2\bar{q}_3}{y} & 0 & \frac{\bar{q}_1 q_2}{y} \end{pmatrix}$	$\begin{pmatrix} 0 & i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & i\frac{\bar{q}_1 q_2}{y} \\ 0 & 0 & i\frac{q_1\bar{q}_2}{\bar{y}} & 0 \end{pmatrix}$
$\psi_3 \longleftarrow \frac{c}{q_1}\psi_3$ $\psi_4 \longleftarrow \frac{y}{\bar{c}q_2}\psi_4$	$\begin{pmatrix} \frac{q_1}{c} & 0 & \frac{q_3}{c} & 0 \\ 0 & \frac{\bar{c}q_2}{y} & 0 & \frac{\bar{c}q_4}{y} \\ -\frac{c\bar{q}_4}{\bar{y}} & 0 & \frac{c\bar{q}_2}{\bar{y}} & 0 \\ 0 & \frac{\bar{q}_3}{\bar{c}} & 0 & \frac{\bar{q}_1}{\bar{c}} \end{pmatrix}$	$\begin{pmatrix} 0 & i & 0 & 0 \\ i & 0 & 0 & 0 \\ 0 & 0 & 0 & i \\ 0 & 0 & i & 0 \end{pmatrix}$

Table 2.3 – The steps that transform $\hat{G}_o$ into $\hat{G}_r$ and their consequence on the matrices $\hat{R}$. The mode operations in the three first rows are performed to satisfy equation (2.106). The scaling factor c in the two last rows can be different from each other and can take any value other than zero. In this table, "Mode composition 1" and "Mode composition 2" are combined.

2.9 Conclusion

In this chapter, we introduce the Poynting operation, provide the corresponding orthogonality relation while clearly stating the assumptions, and present its main uses, namely the reformulation of the boundary condition and the orthonormalization of modes.

In the reformulation of the boundary condition, the coupling coefficients can be found from the application of the Poynting operation on the modes present in both media. We show that such reformulation has several advantages. First, the estimation of the coupling coefficients between a subset of modes without taking into account the contribution of the other modes is as good or slightly better than similar operations presented in the literature [103, 105, 106, 117]. Second, different expressions of the coupling coefficients are provided. Therefore, if the estimates are not valid, the values resulting from these expressions are usually different. Third, the same expression of the coupling coefficients can be used for different systems.

An interesting point in the reformulation of the boundary condition is that, by choosing the appropriate equations, the obtained coupling coefficients satisfy the condition that the power flow towards the interface is equal to the power flow away of the interface. This property seems to hold if additional modes are taken into account, but this is only a conjecture.

For the orthonormalization of modes, different operations on the Gram matrix, a matrix which describe all the interactions between modes, are proposed. These operations can be used to implement an algorithm similar to the Gaussian elimination. They are used in the Fourier modal method presented in chapter 3 to orthonormalize and rotate the three different types of mode that are present in lossless ZSI z-invariant media.

In addition to its uses, the Poynting operation has also a meaning since it is related to the power flow through the Poynting vector. Moreover, the Poynting operation is defined only by its properties, which means that there is a certain flexibility in the exact definition of the Poynting operator. In other words, the work presented here can be used in many different contexts.

2.10 Proofs

2.10.1 Proof of the orthogonality relation

In section 2.4, the following orthogonality relations have been stated:

$$(\gamma_m - \bar{\gamma}_n) \iint_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n^* \times \vec{H}_m) \cdot \vec{n} \, ds = 0. \tag{2.109}$$

$$(\gamma_m^2 - \bar{\gamma}_n^2)[\psi_m | \psi_n] = 0, \tag{2.110}$$

$$(\gamma_m^2 - \bar{\gamma}_n^2)[\psi_m | \psi_n + \psi_n^-] = 0, \tag{2.111}$$

assuming that

$$\vec{X}_m(x,y,z) = \vec{X}_{m0}(x,y)e^{i\gamma_m z}, \tag{2.112}$$

$$\hat{\epsilon} = \hat{\epsilon}^H \qquad \hat{\mu} = \hat{\mu}^H \qquad \hat{\zeta} = \hat{\xi}^H, \tag{2.113}$$

$$\oint_{\partial S} \vec{n} \times (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n^* \times \vec{H}_m) \cdot d\vec{l} = 0, \tag{2.114}$$

$$\begin{aligned} \hat{\epsilon} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & 0 \\ \epsilon_{21} & \epsilon_{22} & 0 \\ 0 & 0 & \epsilon_{33} \end{pmatrix}, \qquad \hat{\mu} = \begin{pmatrix} \mu_{11} & \mu_{12} & 0 \\ \mu_{21} & \mu_{22} & 0 \\ 0 & 0 & \mu_{33} \end{pmatrix}, \\ \hat{\zeta} = \begin{pmatrix} 0 & 0 & \zeta_{13} \\ 0 & 0 & \zeta_{23} \\ \zeta_{31} & \zeta_{32} & 0 \end{pmatrix}, \qquad \hat{\xi} = \begin{pmatrix} 0 & 0 & \xi_{13} \\ 0 & 0 & \xi_{23} \\ \xi_{31} & \xi_{32} & 0 \end{pmatrix}, \end{aligned} \tag{2.115}$$

where $\vec{X}$ can be the electric or the magnetic field. The Poynting operation is defined as

$$[\psi_m|\psi_n] := \frac{1}{2}\int_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n \times \vec{H}_m^*) \cdot \vec{n}\, ds. \tag{2.116}$$

To prove the orthogonality relation (2.110), the following vector calculus identity is used:

$$\begin{aligned} \nabla \cdot (\vec{E}_m \times \vec{H}_n^*) &= \vec{H}_n^* \cdot (\nabla \times \vec{E}_m) - \vec{E}_m \cdot (\nabla \times \vec{H}_n^*), \\ \nabla \cdot (\vec{E}_n^* \times \vec{H}_m) &= \vec{H}_m \cdot (\nabla \times \vec{E}_n^*) - \vec{E}_n^* \cdot (\nabla \times \vec{H}_m), \end{aligned} \tag{2.117}$$

along with the time-independent ($e^{-i\omega t}$) Maxwell equations combined with the constitutive relation for bianisotropic media (2.19):

$$\begin{aligned} \nabla \times \vec{E}_m &= ik_0(\hat{\zeta}\vec{E}_m + \hat{\mu}\vec{H}_m), & \nabla \times \vec{E}_n^* &= -ik_0(\hat{\zeta}\vec{E}_n + \hat{\mu}\vec{H}_n)^*, \\ \nabla \times \vec{H}_m &= -ik_0(\hat{\epsilon}\vec{E}_m + \hat{\xi}\vec{H}_m), & \nabla \times \vec{H}_n^* &= ik_0(\hat{\epsilon}\vec{E}_n + \hat{\xi}\vec{H}_n)^*. \end{aligned} \tag{2.118}$$

After substituting the cross-products in the right-hand side of equations (2.117) by the Maxwell equations (2.118), we obtain

$$\begin{aligned} \nabla \cdot (\vec{E}_m \times \vec{H}_n^*) &= ik_0(\vec{H}_n^* \cdot (\hat{\zeta}\vec{E}_m + \hat{\mu}\vec{H}_m) - \vec{E}_m \cdot (\hat{\epsilon}\vec{E}_n + \hat{\xi}\vec{H}_n)^*), \\ \nabla \cdot (\vec{E}_n^* \times \vec{H}_m) &= -ik_0(\vec{H}_m \cdot (\hat{\zeta}\vec{E}_n + \hat{\mu}\vec{H}_n)^* - \vec{E}_n^* \cdot (\hat{\epsilon}\vec{E}_m + \hat{\xi}\vec{H}_m)). \end{aligned} \tag{2.119}$$

Due to the assumption that the medium is lossless (2.113) and to the property $\vec{x} \cdot (\hat{A}\vec{y})^* = \vec{y}^* \cdot (\hat{A}^H\vec{x})$, equations (2.119) become

$$\begin{aligned}\nabla\cdot(\vec{E}_m\times\vec{H}_n^*) &= \quad ik_0(\vec{H}_n^*\cdot(\hat{\zeta}\vec{E}_m)+\vec{H}_n^*\cdot(\hat{\mu}\vec{H}_m)-\vec{E}_n^*\cdot(\hat{\epsilon}\vec{E}_m)-\vec{H}_n^*\cdot(\zeta\vec{E}_m)),\\ \nabla\cdot(\vec{E}_n^*\times\vec{H}_m) &= -ik_0(\vec{E}_n^*\cdot(\hat{\xi}\vec{H}_m)+\vec{H}_n^*\cdot(\hat{\mu}\vec{H}_m)-\vec{E}_n^*\cdot(\hat{\epsilon}\vec{E}_m)-\vec{E}_n^*\cdot(\hat{\xi}\vec{H}_m)),\end{aligned} \tag{2.120}$$

leading to

$$\nabla\cdot(\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)=0. \tag{2.121}$$

After the integration on the surface S, equation (2.121) becomes

$$\iint_S \nabla_t\cdot(\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)ds+\iint_S \nabla_\perp\cdot(\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)ds=0. \tag{2.122}$$

If S is a plane, the divergence theorem can be used on the first term followed by the assumption (2.114), meaning that

$$\iint_S \nabla_t\cdot(\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)ds=\oint_{\partial S}(\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)\cdot\vec{n}dl=0. \tag{2.123}$$

Hence, equation (2.122) becomes

$$\iint_S \nabla_\perp\cdot(\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)ds=0. \tag{2.124}$$

By using equation (2.112), which describes the modes in a z-invariant medium, equation (2.124) becomes

$$(\gamma_m-\bar{\gamma}_n)\iint_S(\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)\cdot\vec{n}ds=0, \tag{2.125}$$

which is equation (2.109). Due to the assumption (2.115), if a mode described by ($\vec{E}_t$, $\vec{E}_\perp$, $\vec{H}_t$, $\vec{H}_\perp$, γ_m) fulfills the Maxwell equations, the mode described by ($\vec{E}_t$, $-\vec{E}_\perp$, $-\vec{H}_t$, $\vec{H}_\perp$, $-\gamma_m$) is still a solution of the Maxwell equations. Hence, the following equation is also true:

$$(\gamma_m+\bar{\gamma}_n)\iint_S(-\vec{E}_m\times\vec{H}_n^*+\vec{E}_n^*\times\vec{H}_m)\cdot\vec{n}ds=0. \tag{2.126}$$

By summing and subtracting equations (2.125) and (2.126), the following system of equations is obtained:

$$\begin{aligned} \gamma_m \iint_S (\vec{E}_n^* \times \vec{H}_m) \cdot \vec{n} ds - \bar{\gamma}_n \iint_S (\vec{E}_m \times \vec{H}_n^*) \cdot \vec{n} ds &= 0, \\ -\bar{\gamma}_n \iint_S (\vec{E}_n^* \times \vec{H}_m) \cdot \vec{n} ds + \gamma_m \iint_S (\vec{E}_m \times \vec{H}_n^*) \cdot \vec{n} ds &= 0. \end{aligned} \tag{2.127}$$

If $\gamma_m^2 \neq \bar{\gamma}_n^2$:

$$\begin{aligned} &\iint_S (\vec{E}_n^* \times \vec{H}_m) \cdot \vec{n} ds = 0, \\ &\iint_S (\vec{E}_m \times \vec{H}_n^*) \cdot \vec{n} ds = 0. \end{aligned} \tag{2.128}$$

Hence, equations (2.110) and (2.111) are retrieved:

$$(\gamma_m^2 - \bar{\gamma}_n^2) \iint_S (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n \times \vec{H}_m^*) \cdot \vec{n} ds = 0, \tag{2.129}$$

$$(\gamma_m^2 - \bar{\gamma}_n^2) \iint_S (\vec{E}_n \times \vec{H}_m^*) \cdot \vec{n} ds = 0. \tag{2.130}$$

2.10.2 Proof of the reformulation of the boundary condition

The main equations and expressions stated in section 2.5 are

$$\begin{aligned} S_{Lu} := \sigma(L_u, E) &= \sum_{m=1}^{M} (p_m + r_m) \sigma(L_u, L_m) - \sum_{n=1}^{N} (t_n + q_n) \sigma(L_u, R_n), \\ S_{Rv} := \sigma(R_v, E) &= \sum_{m=1}^{M} (p_m + r_m) \sigma(R_v, L_m) - \sum_{n=1}^{N} (t_n + q_n) \sigma(R_v, R_n), \\ T_{Lu} := \sigma(E, L_u)^* &= \sum_{m=1}^{M} (p_m - r_m) \sigma(L_m, L_u)^* - \sum_{n=1}^{N} (t_n - q_n) \sigma(R_n, L_u)^*, \\ T_{Rv} := \sigma(E, R_v)^* &= \sum_{m=1}^{M} (p_m - r_m) \sigma(L_m, R_v)^* - \sum_{n=1}^{N} (t_n - q_n) \sigma(R_n, R_v)^*, \end{aligned} \tag{2.131}$$

$$\begin{aligned} \sum_{m=1}^{M} (\bar{p}_m - \bar{r}_m) S_{Lm} &= \sum_{n=1}^{N} (\bar{t}_n - \bar{q}_n) S_{Rn}, \\ \sum_{m=1}^{M} (\bar{p}_m + \bar{r}_m) T_{Lm} &= \sum_{n=1}^{N} (\bar{t}_n + \bar{q}_n) T_{Rn}, \end{aligned} \tag{2.132}$$

$$\begin{aligned}
\frac{\partial}{\partial \bar{p}_u}[E|E] &= S_{Lu}, & \frac{\partial}{\partial \bar{p}_u}[E|E]^* &= T_{Lu},\\
\frac{\partial}{\partial \bar{r}_u}[E|E] &= -S_{Lu}, & \frac{\partial}{\partial \bar{r}_u}[E|E]^* &= T_{Lu},\\
\frac{\partial}{\partial \bar{t}_v}[E|E] &= -S_{Rv}, & \frac{\partial}{\partial \bar{t}_v}[E|E]^* &= -T_{Rv},\\
\frac{\partial}{\partial \bar{q}_v}[E|E] &= S_{Rv}, & \frac{\partial}{\partial \bar{q}_v}[E|E]^* &= -T_{Rv}.
\end{aligned} \tag{2.133}$$

E is defined as

$$E := \sum_{m=1}^{M} p_m L_m + r_m L_m^- - \sum_{n=1}^{N} q_n R_n^- + t_n R_n. \tag{2.134}$$

To prove equations (2.131), the properties of a sesquilinear form are used on the right-hand side of the equations:

$$\begin{aligned}
S_{Lu} &= \sum_{m=1}^{M} (\sigma(L_u, p_m L_m) + \sigma(L_u, r_m L_m)) - \sum_{n=1}^{N} (\sigma(L_u, t_n R_n) + \sigma(L_u, q_n R_n)),\\
S_{Rv} &= \sum_{m=1}^{M} (\sigma(R_v, p_m L_m) + \sigma(R_v, r_m L_m)) - \sum_{n=1}^{N} (\sigma(R_v, t_n R_n) + \sigma(R_v, q_n R_n)),\\
T_{Lu} &= \sum_{m=1}^{M} (\sigma(p_m L_m, L_u)^* - \sigma(r_m L_m, L_u)^*) - \sum_{n=1}^{N} (\sigma(t_n R_n, L_u)^* - \sigma(q_n R_n, L_u)^*),\\
T_{Rv} &= \sum_{m=1}^{M} (\sigma(p_m L_m, R_v)^* - \sigma(r_m L_m, R_v)^*) - \sum_{n=1}^{N} (\sigma(t_n R_n, R_v)^* - \sigma(q_n R_n, R_v)^*).
\end{aligned} \tag{2.135}$$

After using the properties (2.17) and recognizing the expression of E, equations (2.135) become

$$\begin{aligned}
S_{Lu} &= \sum_{m=1}^{M} \sigma(L_u, p_m L_m + r_m L_m^-) - \sum_{n=1}^{N} \sigma(L_u, t_n R_n + q_n R_n^-) = \sigma(L_u, E),\\
S_{Rv} &= \sum_{m=1}^{M} \sigma(R_v, p_m L_m + r_m L_m^-) - \sum_{n=1}^{N} \sigma(R_v, t_n R_n + q_n R_n^-) = \sigma(R_v, E),\\
T_{Lu} &= \sum_{m=1}^{M} \sigma(p_m L_m + r_m L_m^-, L_u)^* - \sum_{n=1}^{N} \sigma(t_n R_n + q_n R_n^-, L_u)^* = \sigma(E, L_u)^*,\\
T_{Rv} &= \sum_{m=1}^{M} \sigma(p_m L_m + r_m L_m^-, R_v)^* - \sum_{n=1}^{N} \sigma(t_n R_n + q_n R_n^-, R_v)^* = \sigma(E, R_v)^*.
\end{aligned} \tag{2.136}$$

Hence, when E is null, the expressions (2.131) are equal to zero. To prove equations (2.132), they can be written as

$$\begin{aligned} \sum_{m=1}^{M}(\bar{p}_m-\bar{r}_m)\sigma(L_m,E) &-\sum_{n=1}^{N}(\bar{t}_n-\bar{q}_n)\sigma(R_n,E) &&=0,\\ \sum_{m=1}^{M}(\bar{p}_m+\bar{r}_m)\sigma(E,L_m)^* &-\sum_{n=1}^{N}(\bar{t}_n+\bar{q}_n)\sigma(E,R_n)^* &&=0. \end{aligned} \tag{2.137}$$

In a similar procedure to the precedent proof, the properties (2.17) along with the properties of sesquilinear forms are used on equations (2.137):

$$\begin{aligned} &\sum_{m=1}^{M}(\bar{p}_m-\bar{r}_m)\sigma(L_m,E)-\sum_{n=1}^{N}(\bar{t}_n-\bar{q}_n)\sigma(R_n,E)\\ &\quad=\sum_{m=1}^{M}\sigma(p_mL_m+r_mL_m^-,E)-\sum_{n=1}^{N}\sigma(t_nR_n+q_nR_n^-,E),\\ &\sum_{m=1}^{M}(\bar{p}_m+\bar{r}_m)\sigma(E,L_m)^*-\sum_{n=1}^{N}(\bar{t}_n+\bar{q}_n)\sigma(E,R_n)^*\\ &\quad=\sum_{m=1}^{M}\sigma(E,p_mL_m+r_mL_m^-)^*-\sum_{n=1}^{N}\sigma(E,t_nR_n+q_nR_n^-)^*. \end{aligned} \tag{2.138}$$

The expression of E can be recognized in the right-hand side of equations (2.138). Hence, the following equalities are obtained:

$$\begin{aligned} \sum_{m=1}^{M}(\bar{p}_m-\bar{r}_m)S_{Lm}-\sum_{n=1}^{N}(\bar{t}_n-\bar{q}_n)S_{Rn}&=\sigma(E,E),\\ \sum_{m=1}^{M}(\bar{p}_m+\bar{r}_m)T_{Lm}-\sum_{n=1}^{N}(\bar{t}_n+\bar{q}_n)T_{Rn}&=\sigma(E,E)^*. \end{aligned} \tag{2.139}$$

By setting E to null, equations (2.132) are proved. To prove equations (2.133), the following properties of the Wirtinger derivative on a sesquilinear form are used:

$$\begin{aligned} \frac{\partial}{\partial\bar{x}_u}\sigma\left(\sum_{k=1}^{K}x_k\psi_k,\sum_{k=1}^{K}x_k\psi_k\right)&=\sigma\left(\psi_u,\sum_{k=1}^{K}x_k\psi_k\right),\\ \frac{\partial}{\partial\bar{x}_u}\sigma\left(\sum_{k=1}^{K}x_k\psi_k,\sum_{k=1}^{K}x_k\psi_k\right)^*&=\sigma\left(\sum_{k=1}^{K}x_k\psi_k,\psi_u\right)^*. \end{aligned} \tag{2.140}$$

Moreover, due to property (2.15b),

$$[E|E]=\sigma(E,E). \tag{2.141}$$

By combining the properties (2.17), equations (2.140) and equation (2.141),

$$
\begin{aligned}
\frac{\partial}{\partial \bar{q}_v}\sigma(E,E) &= \sigma(-R_v^-,E) &&= S_{Rv},\\
\frac{\partial}{\partial \bar{q}_v}\sigma(E,E)^* &= \sigma(E,-R_v^-)^* &&= -T_{Rv}.
\end{aligned}
\tag{2.142}
$$

The other 6 equations in equations (2.133) can be proven in a similar way.

2.10.3 Proof of the operations on the Gram matrix

In this section, the different operations presented in table 2.4, which is a copy of the table 2.1 in section 2.6, are proven.

Operation name	Mode operation
Mode swapping	$\psi_m \longleftrightarrow \psi_n$
Mode scaling	$\psi_m \longleftarrow k\psi_m$
Mode reversal	$\psi_m \longleftarrow \psi_m^-$
Mode composition 1	$\psi_m \longleftarrow \psi_m + \frac{k}{2}(\psi_n + \psi_n^-)$
Mode composition 2	$\psi_m \longleftarrow \psi_m + \frac{k}{2}(\psi_n - \psi_n^-)$

Operation name	Operation on $\hat{A}$, $\hat{B}$	Operation on $\hat{G}$
Mode swapping	$L_{Am} \longleftrightarrow L_{An}$ $L_{Bm} \longleftrightarrow L_{Bn}$	$C_{Gm} \longleftrightarrow C_{Gn}$ $L_{Gm} \longleftrightarrow L_{Gn}$
Mode scaling	$L_{Am} \longleftarrow kL_{Am}$ $L_{Bm} \longleftarrow kL_{Bm}$	$C_{Gm} \longleftarrow kC_{Gm}$ $L_{Gm} \longleftarrow \bar{k}L_{Gm}$
Mode reversal	$L_{Bm} \longleftarrow -L_{Bm}$	$L_{Gm} \longleftarrow -L_{Gm}$
Mode composition 1	$L_{Am} \longleftarrow L_{Am} + kL_{An}$	$C_{Gm} \longleftarrow C_{Gm} + kC_{Gn}$
Mode composition 2	$L_{Bm} \longleftarrow L_{Bm} + kL_{Bn}$	$L_{Gm} \longleftarrow L_{Gm} + \bar{k}L_{Gn}$

Table 2.4 – A set of operations that can be done on $\hat{G}$ and its consequences on different quantities.

Before proving the different operations present in table 2.4, the general formula is derived with the use of the vectorial and matrix representation of the modes, the sesquilinear form σ and the minus operator. The vectorial representation of the mode ψ is the vector $\vec{v}$ and the matrix representation of the set of modes at the r-th iteration $\vec{\psi}_r$ is $\hat{V}_r$ where each line represents a mode. The minus operator can be written as

$$
\psi^- \equiv \vec{v}^T\hat{M},
\tag{2.143}
$$

where $\hat{M}$ is the matrix representation of the minus operator. Because of the property (2.14c) of the minus operator, the matrix $\hat{M}$ is an involutory matrix, meaning that $\hat{M}^2 = I$. In section 2.6, the matrices $\hat{A}$ and $\hat{B}$ are introduced in the following way:

$$\vec{\psi}_r = \frac{1}{2}\hat{A}_r(\vec{\psi}_0 + \vec{\psi}_0^-) + \frac{1}{2}\hat{B}_r(\vec{\psi}_0 - \vec{\psi}_0^-). \tag{2.144}$$

Using the matrix representation of a set of modes and the minus operator, equation (2.144) becomes

$$\hat{V}_r = \frac{1}{2}(\hat{A}_r + \hat{B}_r)\hat{V}_0 + \frac{1}{2}(\hat{A}_r - \hat{B}_r)\hat{V}_0\hat{M}. \tag{2.145}$$

The sesquilinear form σ can be written as

$$\sigma(\psi_m, \psi_n) \equiv \vec{v}_m^H \hat{\Phi} \vec{v}_n, \tag{2.146}$$

where Φ is the matrix representation of σ. The properties (2.17) become

$$\begin{aligned} \hat{\Phi}\hat{M}^T &= \hat{\Phi}, \\ \hat{M}^*\hat{\Phi} &= -\hat{\Phi}. \end{aligned} \tag{2.147}$$

From section 2.6, the matrix $\hat{G}$ has been introduced as

$$G_{mn} = \sigma(\psi_m, \psi_n), \; m, n \in [1, M]. \tag{2.148}$$

Using equation (2.146), $\hat{G}$ can be written as

$$\hat{G} = \hat{V}^* \hat{\Phi} \hat{V}^T. \tag{2.149}$$

Let us introduce the transformation matrices $\hat{P}_1$ and $\hat{P}_2$ such that:

$$\begin{aligned} \vec{\psi}_r &= \hat{P}_1\vec{\psi}_{r-1} + \hat{P}_2\vec{\psi}_{r-1}^-, \\ \hat{V}_r &= \hat{P}_1\hat{V}_{r-1} + \hat{P}_2\hat{V}_{r-1}\hat{M}. \end{aligned} \tag{2.150}$$

Both equalities are equivalent and $\hat{P}_1$ and $\hat{P}_2$ represent the mode operations listed in table 2.4. In order to find the relationships between $\hat{A}_{r-1}$, $\hat{B}_{r-1}$ and $\hat{A}_r$, $\hat{B}_r$, equations (2.145) and (2.150) are combined:

$$\begin{aligned}\hat{V}_r =& \frac{1}{2}(\hat{A}_r + \hat{B}_r)\hat{V}_0 + \frac{1}{2}(\hat{A}_r - \hat{B}_r)\hat{V}_0\hat{M}\\ =& \frac{1}{2}\hat{P}_1((\hat{A}_{r-1} + \hat{B}_{r-1})\hat{V}_0 + (\hat{A}_{r-1} - \hat{B}_{r-1})\hat{V}_0\hat{M})\\ &+ \frac{1}{2}\hat{P}_2((\hat{A}_{r-1} + \hat{B}_{r-1})\hat{V}_0 + (\hat{A}_{r-1} - \hat{B}_{r-1})\hat{V}_0\hat{M})\hat{M}\\ =& \frac{1}{2}((\hat{P}_1 + \hat{P}_2)\hat{A}_{r-1} + (\hat{P}_1 - \hat{P}_2)\hat{B}_{r-1})\hat{V}_0\\ &+ \frac{1}{2}((\hat{P}_1 + \hat{P}_2)\hat{A}_{r-1} + (\hat{P}_2 - \hat{P}_1)\hat{B}_{r-1})\hat{V}_0\hat{M}.\end{aligned} \tag{2.151}$$

Hence, $\hat{A}_r$ and $\hat{B}_r$ are given by

$$\begin{aligned}\hat{A}_r &= (\hat{P}_1 + \hat{P}_2)\hat{A}_{r-1},\\ \hat{B}_r &= (\hat{P}_1 - \hat{P}_2)\hat{B}_{r-1}.\end{aligned} \tag{2.152}$$

The Gram matrix at the r-th iteration $\hat{G}_r$ is given by

$$\begin{aligned}\hat{G}_r &= \hat{V}_r^*\hat{\Phi}\hat{V}_r^T\\ &= (\hat{P}_1^*\hat{V}_{r-1}^* + \hat{P}_2^*\hat{V}_{r-1}^*\hat{M}^*)\hat{\Phi}(\hat{V}_{r-1}^T\hat{P}_1^T + \hat{M}^T\hat{V}_{r-1}^T\hat{P}_2^T)\\ &= \frac{1}{2}[(\hat{A}_r^* + \hat{B}_r^*)\hat{V}_0^* + (\hat{A}_r^* - \hat{B}_r^*)\hat{V}_0^*\hat{M}^*]\hat{\Phi}\frac{1}{2}[\hat{V}_0^T(\hat{A}_r^T + \hat{B}_r^T) + \hat{M}^T\hat{V}_0^T(\hat{A}_r^T - \hat{B}_r^T)].\end{aligned} \tag{2.153}$$

Using the properties (2.147) and recognizing $\hat{G}_{r-1}$ and $\hat{G}_0$, $\hat{G}_r$ becomes

$$\begin{aligned}\hat{G}_r &= (\hat{P}_1 - \hat{P}_2)^*\hat{G}_{r-1}(\hat{P}_1 + \hat{P}_2)^T\\ &= \hat{B}_r^*\hat{G}_0\hat{A}_r^T.\end{aligned} \tag{2.154}$$

For the proof of the operations called "Mode swapping", "Mode composition 1", and "Mode composition 2", the matrices $\hat{P}_1$ and $\hat{P}_2$ are defined for each cases and the operation on matrices $\hat{A}$, $\hat{B}$, and $\hat{G}$ are derived using equations (2.152) and (2.154). Then, the "Mode composition 1" is combined with "Mode composition 2" in order to prove the "Mode scaling" and "Mode reversal" operations.

For the "Mode swapping" operation, $\hat{P}_2$ is null and $\hat{P}_1$ is a permutation matrix that permutes the rows m and n. $\hat{P}_1$ is symmetric and real. Hence,

$$\begin{aligned} L_{Am} &\longleftrightarrow L_{An}, \\ L_{Bm} &\longleftrightarrow L_{Bn}, \\ C_{Gm} &\longleftrightarrow C_{Gn}, \\ L_{Gm} &\longleftrightarrow L_{Gn}. \end{aligned} \tag{2.155}$$

For the "Mode composition 1" operation, $\hat{P}_1$ and $\hat{P}_2$ are

$$\begin{aligned} \hat{P}_1 &:= \hat{I} + \frac{1}{2}\hat{K}, \\ \hat{P}_2 &:= \frac{1}{2}\hat{K}, \end{aligned} \tag{2.156}$$

with $\hat{K}$ defined as

$$K_{st} = \begin{cases} k & \text{if } s = m \cap t = n, \\ 0 & \text{otherwise.} \end{cases} \tag{2.157}$$

Hence,

$$\begin{aligned} L_{Am} &\longleftarrow L_{Am} + kL_{An}, \\ C_{Gm} &\longleftarrow C_{Gm} + kC_{Gn}. \end{aligned} \tag{2.158}$$

For the "Mode composition 2" operation, $\hat{P}_1$ and $\hat{P}_2$ are

$$\begin{aligned} \hat{P}_1 &:= \hat{I} + \frac{1}{2}\hat{K}, \\ \hat{P}_2 &:= -\frac{1}{2}\hat{K}. \end{aligned} \tag{2.159}$$

Hence,

$$\begin{aligned} L_{Bm} &\longleftarrow L_{Bm} + kL_{Bn}, \\ L_{Gm} &\longleftarrow L_{Gm} + \bar{k}L_{Gn}. \end{aligned} \tag{2.160}$$

To prove "Mode scaling" and "Mode reversal", the operations "Mode composition 1" and "Mode composition 2" are combined:

$$\begin{aligned}\psi_m &\longleftarrow \psi_m + \frac{k_1}{2}(\psi_m + \psi_m^-),\\ \psi_m &\longleftarrow \psi_m + \frac{k_2}{2}(\psi_m - \psi_m^-),\end{aligned} \tag{2.161}$$

which is equivalent to

$$\begin{aligned}\psi_m &\longleftarrow \psi_m + \frac{k_1}{2}(\psi_m + \psi_m^-) + \frac{k_2}{2}(\psi_m + \frac{k_1}{2}(\psi_m + \psi_m^-) - \psi_m^- - \frac{k_1}{2}(\psi_m + \psi_m^-))\\ \Rightarrow \psi_m &\longleftarrow \psi_m + \frac{k_1}{2}(\psi_m + \psi_m^-) + \frac{k_2}{2}(\psi_m - \psi_m^-).\end{aligned} \tag{2.162}$$

Moreover, "Mode composition 1" commutes with "Mode composition 2" because "Mode composition 1" and "Mode composition 2" modify only the matrix $\hat{A}$ and $\hat{B}$ respectively. The "Mode scaling" operation is a combination of "Mode composition 1" and "Mode composition 2" and it is equivalent to the operations (2.161) with $k_1 = k_2 = k - 1$. From the operations (2.158) and (2.160) on $\hat{A}$, $\hat{B}$, and $\hat{G}$, the following operations are obtained:

$$\begin{aligned}L_{Am} &\longleftarrow kL_{An},\\ L_{Bm} &\longleftarrow kL_{Bn},\\ C_{Gm} &\longleftarrow kC_{Gn},\\ L_{Gm} &\longleftarrow \bar{k}L_{Gn}.\end{aligned} \tag{2.163}$$

The "Mode reversal" is equivalent to the operations (2.161) with $k_1 = 0$ and $k_2 = -2$. Hence,

$$\begin{aligned}L_{Bm} &\longleftarrow -L_{Bm},\\ L_{Gm} &\longleftarrow -L_{Gn}.\end{aligned} \tag{2.164}$$

3 Improved Fourier modal method

3.1 Introduction

Optical components simulated and designed in this work are mainly binary dielectric metasurfaces, which can usually be considered as two-dimensional gratings, also known as crossed gratings. In order to be considered as a metasurface, the structures dimensions should be ideally deeply sub-wavelength, but, in this work, optical components composed of structures smaller than the wavelength of the light are still considered as metasurfaces. An example of such metasurface is shown in fig. 3.1a, which is an array of silicon cylinders on a glass substrate. In the near-infrared regime, the lattice constant is usually below one micron.

Different rigorous methods can be used to simulate metasurfaces. The well-known methods include the Finite-Difference Time-Domain method (FDTD) [75], the Finite-Difference Frequency-Domain method (FDFD) [118], Finite Element Method (FEM) [76–78], Boundary Element Method (BEM), also known as Method of Moments (MoM) [119], and the Fourier Modal Method (FMM) [61, 79–81], also known as the Rigorous Coupled Wave Analysis (RCWA) [82]. Ref. [95] revisits many of those methods. The method used in this work is the Fourier Modal Method and this method has been modified and improved in order to facilitate the design and analysis of metasurfaces as shown in chapters 4 to 6. In this work, we divide the Fourier modal method into three operations: the computation of the eigen-modes inside a layer, the computation of the S-matrix at an interface between two layers and the reduction of a layer into an interface. A layer is defined as a z-invariant medium between two interfaces and an interface is a plane perpendicular to the z-axis that separates two z-invariant media. In order to use the Fourier modal method on a multi-layer structure, the dimensions of the unit cell is the same for each layer. In this chapter, we consider that the z-invariant media are also Z-Symmetry Invariant (ZSI). The ZSI property is defined in section 2.4. Z-invariant media composed of isotropic materials are ZSI.

As an example, the metasurface shown in fig. 3.1a consists of one layer and two interfaces. The first interface is between the glass substrate and the metasurface and the second interface is between the metasurface and air. The Fourier modal method can also be used to simulate an array of structures that vary continuously along the z dimension as the one shown in fig. 3.1c by approximating it by a multi-layer structure as shown in fig. 3.1d. However, the benefits of the Fourier modal method presented in this chapter are lost if the layers that compose the system are so thin that many evanescent and complex modes need to be considered in order to get the response of the layer.

The discussion is organized as follows: The three operations mentioned earlier are introduced in section 3.1.1 along with the main differences between the Fourier modal methods proposed in the literature and the one presented in this work. In order to use the full potential of the Fourier modal method for the design of optical structures, it is important to consider a multi-layer structure as a collection of objects and those operations allow to get another kind of object or transform them. The list of those objects are described in section 3.1.2 along with their representation in a diagram. Each object contains information that might be useful for the design of a structure.

For a thick layer, only a few eigen-modes have an impact on the overall response of the structure. The action of reducing the number of eigen-modes, called mode filtering, is given in section 3.5. In section 3.1.2, the impact of mode filtering on the different operations is summarized. Section 3.5 also provides the equations needed for the computation of the contribution of the modes to the power flow assuming that the modes are orthonormalized (sections 2.8.2 and 2.8.3).

3.1.1 Overview of the improved Fourier modal method

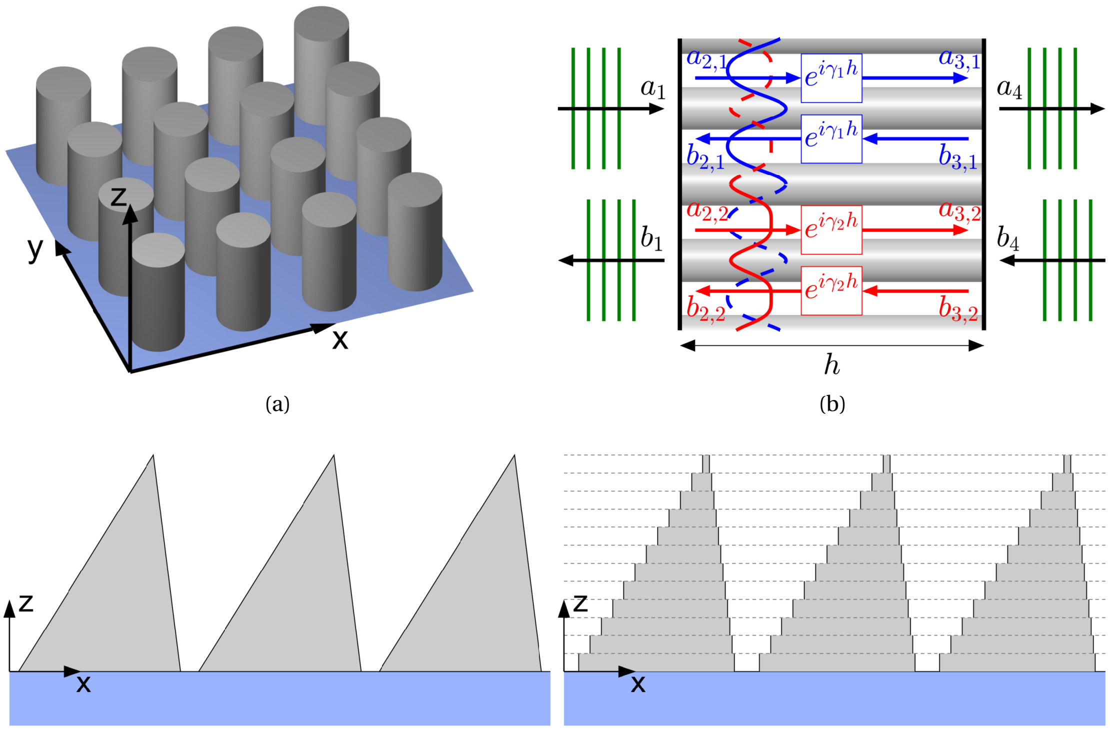
The Fourier modal method can be divided into three operations. The first operation is the computation of the eigen-modes in a layer and it is based on the work of L. Li [61]. Eigen-modes and their properties have been discussed in depth in chapter 2. As a reminder, eigen-modes are solutions of the source-free Maxwell equations in the time-harmonic regime for a z-invariant medium in the form

$$(\vec{E}(x,y,z),\vec{H}(x,y,z)) = (\vec{E}_0(x,y),\vec{H}_0(x,y))e^{i\gamma z}, \tag{3.1}$$

where γ is the propagation constant. Different methods can be used to compute $\vec{E}_0$ and $\vec{H}_0$. In this work, the Fourier modal method proposed by L. Li [61] is used and the summary of the Fourier modal method along with the state of the art is given in section 3.2.

In homogeneous media, the eigen-modes are simply plane waves and equation (3.1) simplifies to

$$(\vec{E}(x,y,z),\vec{H}(x,y,z)) = (\vec{E}_0,\vec{H}_0)e^{i(k_x x+k_y y+\gamma z)}, \tag{3.2}$$



(a)

(b)

(c)

(d)

Figure 3.1 – a) Metasurface composed of an array of silicon cylinders on a glass substrate surrounded by air. b) A single layer composed of thickness h of a ZSI z-invariant medium surrounded by two homogeneous media, where two pairs of eigen-modes propagating inside this layer are represented by arrows. Each pair is composed of a forward-propagating mode, meaning that the arrow goes from left to right, and a backward-propagating mode, meaning that the arrow goes from right to left. Because the medium is ZSI, each pair of eigen-modes share the same field profile, which are represented by the blue and red curves. The z-axis is from left to right. a and b are the weights just before or after the interfaces. In order to get $a_{3,m}$ and $b_{2,m}$ from respectively $a_{2,m}$ and $b_{3,m}$, $a_{2,m}$ and $b_{3,m}$ are multiplied by $e^{i\gamma_m h}$. c) Array of structures that vary along the z dimension. d) Approximation of the structure shown in fig. 3.1c into a multi-layer structure that can be simulated using the Fourier modal method. The dashed lines are interfaces.

where k_x and k_y are the tangential components of the wave vector $\vec{k}$. The fields $\vec{E}_0$ and $\vec{H}_0$ can be found analytically and are a function of k_x and k_y. Whether the medium is homogeneous or z-invariant heterogeneous, the expressions of $\vec{E}_0$ and $\vec{H}_0$ for a given propagation constant γ are not unique because a multiplication of an eigen-mode by a complex constant or the addition of two eigen-modes with the same propagation constant also represents an eigen-modes. For homogeneous isotropic media, the convention chosen in this work is given in section 3.2.1 and it has the advantage to be valid for any permittivity ϵ, permeability μ, k_x and k_y with the exception of the case $k_x = k_y = \epsilon\mu = 0$. In practice, the medium is non-magnetic ($\mu = 1$) and the permittivity ϵ is equal to or higher than one, so the limitation of this convention is not an issue. For heterogeneous isotropic media, the eigen-modes are normalized using the equations provided in sections 2.8.2 and 2.8.3 and, if there are multiple eigen-modes with the same propagation constant, they can be combined as shown in section 2.8.4.

Once the eigen-modes are computed, they are divided into forward-propagating and backward-propagating modes based on the imaginary part of their propagation constant and, for propagating modes, the z component of the Poynting vector. The result can be illustrated by fig. 3.1b, where two forward-propagating modes and two backward-propagating modes are represented. The coefficients $a_{m,n}$ and $b_{m,n}$ are the weights of the modes just before and after the interfaces and they are, at this point, unkowns and depend on the illumination condition. By convention, the weights a and b are the weights of the forward-propagating and backward-propagating modes respectively.

The relationship between the weights of the eigen-modes just before and after a given interface is obtained from the boundary condition, stating that the transverse eletric and magnetic fields are continuous across the interface, and can be expressed by the S-matrix or the T-matrix. Since they are multiple eigen-modes on both sides of the interface, the weights are represented by vectors, where the m-th element is the weight of the mode m. If $\vec{a}_1$ and $\vec{b}_1$ are the weights of the eigen-modes at the left-hand side of the interface and $\vec{a}_2$ and $\vec{b}_2$ are the weights of the eigen-modes at the right-hand side of the interface, the weights are related with each others by the equation

$$\hat{S}\begin{pmatrix} \vec{a}_1 \\ \vec{b}_2 \end{pmatrix} = \begin{pmatrix} \hat{T}_1 & \hat{R}_1 \\ \hat{R}_2 & \hat{T}_2 \end{pmatrix}\begin{pmatrix} \vec{a}_1 \\ \vec{b}_2 \end{pmatrix} = \begin{pmatrix} \vec{a}_2 \\ \vec{b}_1 \end{pmatrix}, \tag{3.3}$$

where $\hat{S}$ is the S-matrix, which is composed of four sub-matrices $\hat{R}_1$, $\hat{R}_2$, $\hat{T}_1$ and $\hat{T}_2$ that we call coupling matrices. The other way to represent the relationships between the weights is through the T-matrix $\hat{T}$:

$$\hat{T}\begin{pmatrix} \vec{a}_1 \\ \vec{b}_1 \end{pmatrix} = \begin{pmatrix} \vec{a}_2 \\ \vec{b}_2 \end{pmatrix}. \tag{3.4}$$

The T-matrix is more straightforward to compute than the S-matrix, and the T-matrix of a multi-layer structure can be obtained through a multiplication of the T-matrices that represents the different interfaces in the structure and the separations between them. However, such multiplication is usually unstable because the T-matrix that describes the separation between two interfaces, usually contains both extremely large and small terms due to the evanescent and complex modes in the layer [94]. The use of the S-matrix solves this issue and the S-matrix is also more related to physics because the weights on the left-hand side of equation (3.3) are the weights of the modes going toward the interface and the weights on the right-hand side of equation (3.3) are the weights of the modes going away of the interface. Since the T-matrix is easier to get and the S-matrix is more stable, the T-matrix is computed first and, then, this T-matrix is converted into a S-matrix

As shown in this work, the S-matrix of a multi-layer system can be computed without going through the T-matrix and, if the system is composed of ZSI media, the computation of a S-matrix can be two times faster than the computation of the T-matrix in [94], assuming that the matrix inversion is done through the Gauss-Jordan elimination algorithm. The equations used to compute the S-matrix are provided in section 3.3. The same equations could be used for the case of an interface between a homogeneous and a heterogeneous medium, but, if the homogeneous medium is lossless and a plane wave in the homogeneous medium is between a propagating and an evanescent wave, meaning that the propagation constant of this wave is zero, it is possible that one of the matrix cannot be inverted. This is a typical issue in some implementations of the Fourier modal method and this issue can be seen in the system of equations (B17) in [79]. It is explained in section 3.3 how to avoid this issue. Since the convention for the plane waves is specific to this work, the trivial case of an interface between two homogeneous media is provided in section 3.3.1 and is a simplification of the equations given in section 2.7.2.

In order to find the amplitudes of the transmitted and reflected plane waves for a given illumination condition, the S-matrix representing the whole structure is required. The usual strategy to get the S-matrix of a multi-layer structure is to initialize the S-matrix of the structure to an identity matrix and, by recursion, compute the T-matrix for the next interface and update the S-matrix of the structure while taking into account the separation between the interface and the next. More details about this strategy are given, for a simplified case, in Section 3.5.1 of [80] and, for a more general case, in [94] and in Annex 7.A and 7.B of [95]. The operation introduced in this work and given in section 3.4 allows to reduce a layer into an interface, meaning that one can choose the order at which the reductions of the layers are done. Moreover, the S-matrix of the multi-layer structure obtained using the Fourier modal method presented in this work can give the amplitude of transmitted and reflected plane waves when the structure is illuminated from both sides, which makes the adjoint method presented in chapter 6 more efficient. This characteristic is also present in [94, 95], but not in [80].

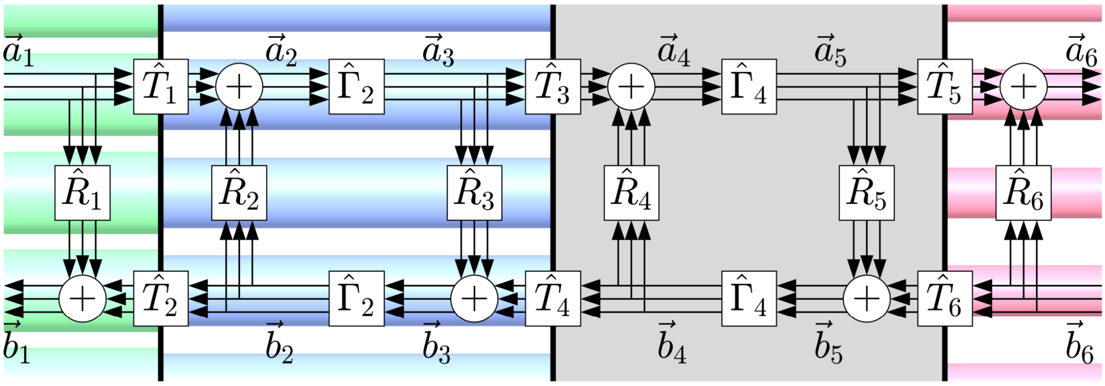


Figure 3.2 – A multi-layer structure composed of homogeneous and heterogeneous z-invariant media, where the objects obtained by the Fourier modal method are represented. The arrows represent the eigen-modes. $\vec{a}_m$ and $\vec{b}_m$ are the weights of the modes of the forward and backward-propagating modes just before and after the interfaces respectively. The matrices $\hat{\Gamma}$ are the propagation operators. The four matrices $\hat{R}_m$ and $\hat{T}_m$ at each interface are the coupling matrices and they are the sub-matrices that compose the S-matrix.

3.1.2 Representation of a multi-layer structure and design strategy

The multi-layer structure under consideration is a stack of layers composed of a z-invariant medium separated by interfaces, which is represented by the cylinders in the background in fig. 3.2. Each layer is composed of either a homogeneous medium, described by a permettivity ϵ and a permeability μ, or a heterogeneous medium described by a permittivity profile $\epsilon(x, y)$. Another input to the method is the Bloch phase, which is explained in section 3.2.

For every layer, the eigen-modes can be computed and, in fig. 3.2, they are represented by arrows. Because the media are ZSI, only the transverse electric and magnetic fields and the propagation constant of the forward-propagating eigen-modes are computed since the forward and backward-propagating modes are related by the minus operator defined in (2.9). That means that the fields and propagation constant of a backward-propagating mode is the same as its forward-propagating counterpart except that the sign of the propagation constant, the transverse components of the magnetic field and the z-component of the electric field is flipped. The z-component of the electric and magnetic fields is not used in the Fourier modal method, so it is computed only if one needs to get the electric and magnetic fields in a layer, which is typically the case for the adjoint method in chapter 6. The electric and magnetic fields of the eigen-modes are represented by respectively the matrix $\hat{E}$ and $\hat{H}$, where the m-th column describes the field of the mode m. The field of a mode allows to get an idea on how the mode is confined and, for example, allows to understand the behavior of the metasurface presented in section 5.3.4.

The propagation constants are represented by the vector $\vec{\gamma}$, where the m-th element is the propagation constant of the mode m. The propagation constant is an important parameter because it indicates how the mode decays in a layer. For a lossless medium, it indicates if the eigen-mode is propagating, evanescent or complex. Moreover, it gives the diagonal matrix $\hat{\Gamma}$, which is called the propagation operator and is shown in fig. 3.2, and relates the weight of the modes, given by $\vec{a}$ and $\vec{b}$ in fig. 3.2, on both sides of a layer. The diagonal element Γ_{mm} is $e^{i\gamma_m h}$. The main advantage of the Fourier modal method for the design of binary metasurfaces is that the thickness of the layers appears only in the matrix $\hat{\Gamma}$ and it is trivial to compute the matrix Γ for different thicknesses. It is explained later in this section how to use this advantage.

Once the eigen-modes of two adjacent layers are known, the coupling matrices, which are the sub-matrices of the S-matrix, at the interface between those two layers can be computed and are given by the matrices $\hat{R}$ and $\hat{T}$. Each interface is represented by four matrices as shown in fig. 3.2. With those matrices, all the elements in fig. 3.2 are explained except the indices of the matrices and vectors. As mentioned earlier, the weigths $\vec{a}$ and $\vec{b}$ are the weight of the eigen-modes just before and after the interfaces and they are numbered from left to right starting by one. Hence, the weights at the first interface are $\vec{a}_1$, $\vec{a}_2$, $\vec{b}_1$ and $\vec{b}_2$, and the weights at the m-th interface are $\vec{a}_{2m-1}$, $\vec{a}_{2m}$, $\vec{b}_{2m-1}$ and $\vec{b}_{2m}$. In other words, if the indice m is odd, $\vec{a}_m$ and $\vec{b}_m$ are the weights of the eigen-modes just before an interface. Otherwise, they are the weights of the eigen-modes just after an interface. The indice of the matrices is the same as the indice of the vector which is multiplied by this matrix. The matrices Γ_m with m odd is not present in fig. 3.2 because Γ_{m+1} is equal to Γ_m due to the ZSI property of the media.

At this stage, the weights $\vec{a}$ and $\vec{b}$ are unknowns, but the transmission and reflection of a multi-layer structure are found without getting the values of those weights since, after successive reduction of the layers, the multi-layer structure will be reduced to a single interface. If those weights are needed in order to get the fields or the power flow inside a layer, additional manipulations are necessary as given in section 3.4. If the eigen-modes in a layer composed of a lossless medium have been orthonormalized as described in sections 2.8.2 and 2.8.3, the contribution of the modes on the power flow is given directly by the mode weights with the condition that the mode type (propagating, evanescent, complex) is known. A mode that does not contribute to the power flow, which is the case for most evanescent and complex modes in a binary metasurface, can be neglected.

If the layer is not too thin, only a few modes need to be taken into account in order to get the response of the whole structure. The action of neglecting those modes are called in this work mode filtering, and it is discussed in section 3.5. Due to the mode filtering, the computational time of the S-matrix for a single interface can be reduced by up to a factor two. The main advantage of the mode filtering is the reduction of a layer into an interface because the size of the coupling matrices and the propagation operators, which are all the matrices required for reducing a layer, are greatly reduced. In other words, after mode filtering, the operation of reducing a layer can be extremely fast. For the metasurfaces designed in this work, keeping a hundred of eigen-modes is a common practice and this approximation has no meaningful impact on the final results. Hence, reducing a layer consists of ten matrix multiplications and a matrix inversion with square matrices of size hundred, leading to a computation time for reducing a layer in the order of the milliseconds.

For the design of a binary metasurface, the main design strategy used here is to compute the coupling matrices and the propagation constants, and to store them after mode filtering, reducing memory usage. With this data, the contribution of the main modes on the power flow and the response of the metasurface for any thickness can be computed nearly instantaneously. Moreover, the coupling matrices are used for the design of single-mode metasurfaces (chapter 4) and for the computation of the self-coupled modes. The concept of self-coupled mode is presented in chapter 5 and greatly facilitates the design and the analysis of resonant metasurfaces.

3.2 Computation of the eigen-modes

This section gives the key elements of the computation of the eigen-modes using the Fourier modal method and how the Fourier modal method has been improved when compared with the work of M. G. Moharam and T. K. Gaylord [120] to the current state of the art given in chapters 7 and 13 of [95]. In order to simplify the equations, the z-invariant periodic medium is considered isotropic and non-magnetic and it has a square lattice. The equations for the general case are provided in [95], but the ideas behind the derivation of those equations are the same.

The objective of the Fourier modal method is to describe any fields, given by $\vec{E}$ and $\vec{H}$, that satisfy the Maxwell equations as a sum of the eigen-modes. Due to the simplification mentioned earlier in this section and the normalization of the magnetic field, the Maxwell equations are given by

$$\begin{aligned} \nabla \times \vec{E} &= i k_0 \vec{H} \\ \nabla \times \vec{H} &= -i k_0 \epsilon(x,y) \vec{E}, \end{aligned} \tag{3.5}$$

where $\vec{E}$ and $\vec{H}$ depend on x, y and z. The Maxwell equation (3.5) is a system of six equations:

$$\begin{aligned} \frac{\delta}{\delta y}E_z - \frac{\delta}{\delta z}E_y &= i k_0 H_x & \frac{\delta}{\delta y}H_z - \frac{\delta}{\delta z}H_y &= -i k_0 \epsilon(x,y) E_x \\ \frac{\delta}{\delta z}E_x - \frac{\delta}{\delta x}E_z &= i k_0 H_y & \frac{\delta}{\delta z}H_x - \frac{\delta}{\delta x}H_z &= -i k_0 \epsilon(x,y) E_y \\ \frac{\delta}{\delta x}E_y - \frac{\delta}{\delta y}E_x &= i k_0 H_z & \frac{\delta}{\delta x}H_y - \frac{\delta}{\delta y}H_x &= -i k_0 \epsilon(x,y) E_z. \end{aligned} \tag{3.6}$$

The field components E_m and H_m, where m can be either x, y or z, can be expressed with the other field components without solving a differential equation. This is true even in the general case where the medium is bianisotropic. In all the Fourier modal methods presented in the literature, it is the field components E_z and H_z that are replaced by their respective expression because, in order to fulfill the boundary condition, only the tangential components of the fields are required. After replacing E_z and H_z, the system of equations (3.6) reduces to

$$\begin{aligned}
\frac{\delta}{\delta z}E_x &= ik_0H_y - \frac{1}{ik_0}\frac{\delta}{\delta x}\left[\frac{1}{\epsilon(x,y)}\left(\frac{\delta}{\delta x}H_y - \frac{\delta}{\delta y}H_x\right)\right]\\
\frac{\delta}{\delta z}E_y &= -ik_0H_x - \frac{1}{ik_0}\frac{\delta}{\delta y}\left[\frac{1}{\epsilon(x,y)}\left(\frac{\delta}{\delta x}H_y - \frac{\delta}{\delta y}H_x\right)\right]\\
\frac{\delta}{\delta z}H_x &= -ik_0\epsilon(x,y)E_y + \frac{1}{ik_0}\frac{\delta}{\delta x}\left(\frac{\delta}{\delta x}E_y - \frac{\delta}{\delta y}E_x\right)\\
\frac{\delta}{\delta z}H_y &= ik_0\epsilon(x,y)E_x + \frac{1}{ik_0}\frac{\delta}{\delta y}\left(\frac{\delta}{\delta x}E_y - \frac{\delta}{\delta y}E_x\right)
\end{aligned} \tag{3.7}$$

The eigen-modes are solutions of the system of equations (3.7) in the form

$$(\vec{E},\vec{H}) = (\vec{E}_0(x,y),\vec{H}_0(x,y))e^{i\gamma z}, \tag{3.8}$$

where γ is called the propagation constant and $\vec{E}_0$ and $\vec{H}_0$ are the field profile of the eigen-mode. Hence, the term $\delta/\delta z$ in (3.7) is replaced with $i\gamma$ and the system of equations (3.7) does not depend of z anymore.

Since ϵ is a periodic function, the solution of (3.7) is a Bloch wave, meaning that

$$(\vec{E}_0(x,y),\vec{H}_0(x,y)) = (\vec{E}_l(x,y),\vec{H}_l(x,y))e^{i(k_{x,0}x+k_{y,0}y)}, \tag{3.9}$$

where $\vec{E}_l$ and $\vec{H}_l$ are periodic functions with the same unit cell as ϵ. The k-vector components $k_{x,0}$ and $k_{y,0}$ are related to the Bloch phase and are given by the tangential k-vector of the illumination. The Bloch phase is the phase difference between $f(\vec{x})$ and $f(\vec{x}+\vec{a})$ where $\vec{a}$ is a lattice vector and f is a Bloch wave.

The specifity of the Fourier modal method is to describe the periodic functions $\vec{E}_l$, $\vec{H}_l$, ϵ and $1/\epsilon$ with a truncated Fourier serie, meaning that $\vec{E}$, $\vec{H}$ and ϵ are expressed as

$$\begin{aligned}
\vec{E} &= \sum_{m=-M}^{M}\sum_{n=-N}^{N}\vec{E}_{mn}e^{i\left(k_{x,mn}x+k_{y,mn}y\right)}e^{i\gamma z}\\
\vec{H} &= \sum_{m=-M}^{M}\sum_{n=-N}^{N}\vec{H}_{mn}e^{i\left(k_{x,mn}x+k_{y,mn}y\right)}e^{i\gamma z}\\
\epsilon(x,y) &= \sum_{m}\sum_{n}\epsilon_{mn}e^{2\pi i\left(\frac{mx}{d_x}+\frac{ny}{d_y}\right)}\\
\frac{1}{\epsilon(x,y)} &= \sum_{m}\sum_{n}(1/\epsilon)_{mn}e^{2\pi i\left(\frac{mx}{d_x}+\frac{ny}{d_y}\right)},
\end{aligned} \tag{3.10}$$

where $\vec{E}_{mn}$, $\vec{H}_{mn}$, ϵ_{mn} and ζ_{mn} are the Fourier coefficients of respectively $\vec{E}$, $\vec{H}$, ϵ and $1/\epsilon$, and

d_x and d_y are the dimensions of the unit cell. The spacial frequencies $k_{x,mn}$ and $k_{y,mn}$ are given by

$$k_{x,mn} = 2\pi \frac{mx}{d_x} + k_{x,0} \qquad k_{y,mn} = 2\pi \frac{ny}{d_y} + k_{y,0}. \tag{3.11}$$

The number M and N are related to the number of Fourier coefficients that is taken into account and are the variables that determine the accuracy of the method and the number of eigen-modes. For each component of the electric and magnetic fields, there are K unknown Fourier coefficients, where K is given by

$$K = (2M+1)(2N+1). \tag{3.12}$$

Since E_z and H_z are expressed in term of E_x, E_y, H_x and H_y, the total numbers of unknowns per eigen-mode are reduced to $4K$, which, in the general case, is also the number of eigen-modes. Since those eigen-modes include forward and backward-propagating modes and, in ZSI media, the fields and the propagation constant of a backward and forward-propagating modes are directly related to each other, the number of eigen-modes that need to be computed is two times smaller. Hence, it is expected that the eigen-modes that satisfy the system of equations (3.7) are the eigen-vector of a square matrix of size $2K$.

The natural representation of the Fourier coefficients for a given component of a field and the spatial frequencies k_x and k_y given in (3.11) is a matrix of size $2M+1$-by-$2N+1$. However, it is more practical to represent them in a vector of size K. In this work, this vector is denoted $\widetilde{X}$, where X can be the spatial frequency k_x or k_y, or any component of the electric or magnetic field. If p is the index of the p-th element of the vector $\widetilde{X}$, p is given by

$$p = m + M + 1 + (2M+1)(n+N), \; m \in [-M, M], \; n \in [-N, N]. \tag{3.13}$$

Replacing the expressions of $\vec{E}$, $\vec{H}$, ϵ and $1/\epsilon$ in equations (3.7) by their expressions given in (3.10) is equivalent to taking the Fourier transform of equations (3.7) except a small modification due to the Bloch phase. The main difficulty of the Fourier modal method is that the system of equations (3.7) contains the multiplication of two terms that depend of x and y, which becomes a convolution in the Fourier domain.

Let f, g and h be three periodic functions that depend on x and y such that $h = f \cdot g$. Then, the Fourier transform of h, denoted $\mathscr{F}(h)$, is given by

$$\mathscr{F}(h) = \mathscr{F}(f) * \mathscr{F}(g), \tag{3.14}$$

where $*$ is the convolution. If g and h are expressed as truncated Fourier series whose $\widetilde{g}$ and $\widetilde{h}$ are their Fourier coefficients, $\widetilde{h}$ is given by the Laurent's rule:

$$\widetilde{h} = [[f]]\widetilde{g}, \tag{3.15}$$

where $[[f]]$ is a Teoplitz matrix composed with the Fourier coefficients $\widetilde{f}$. Since the Fourier coefficients are represented in a vector, the Teoplitz matrix should be seen as a Teoplitz matrix composed of $2N+1$-by-$2N+1$ Toeplitz matrices of size $2M+1$-by-$2M+1$. If the Teoplitz matrix $[[f]]_{p_1 q_1}$, which is the matrix in the row p_1 and column q_1 of the matrix $[[f]]$, is denoted $[[f]]_{q_1-p_1}$, the element $f_{q_1-p_1, p_2 q_2}$ is the Fourier coefficient $f_{(q_2-p_2)(q_1-p_1)}$.

It is expected that, by retaining more Fourier coefficients, $\widetilde{h}$ converges to $\mathscr{F}(h)$, but, as pointed out by L. Li [121], this is not true if f and g are discontinuous and h is continuous. In most gratings, the permittivity profile ϵ and, therefore, the electric field $\vec{E}$ are discontinuous. To avoid this convergence issue and let f be discontinuous, it is important to use the Laurent's rule, given by equation (3.15), if g is continuous and h is discontinuous, and the inverse rule, introduced by L. Li [121] and given by

$$\widetilde{h} = [[1/f]]^{-1}\widetilde{g}, \tag{3.16}$$

if g is discontinuous and h is continuous.

If $\vec{E}$ in the system of equations (3.7) is replaced by its expression in (3.10) and the convergence issue pointed out by L. Li is ignored, the following system of equations is obtained:

$$\begin{aligned}
k_0\gamma\widetilde{E}_x &= k_0^2\widetilde{H}_y - \mathrm{diag}(\widetilde{k}_x)[[1/\epsilon]]\left(\mathrm{diag}(\widetilde{k}_x)\widetilde{H}_y - \mathrm{diag}(\widetilde{k}_y)\widetilde{H}_x\right)\\
k_0\gamma\widetilde{E}_y &= -k_0^2\widetilde{H}_x - \mathrm{diag}(\widetilde{k}_y)[[1/\epsilon]]\left(\mathrm{diag}(\widetilde{k}_x)\widetilde{H}_y - \mathrm{diag}(\widetilde{k}_y)\widetilde{H}_x\right)\\
k_0\gamma\widetilde{H}_x &= -ik_0^2[[\epsilon]]\widetilde{E}_y + \mathrm{diag}(\widetilde{k}_x)\left(\mathrm{diag}(\widetilde{k}_x)\widetilde{E}_y - \mathrm{diag}(\widetilde{k}_y)\widetilde{E}_x\right)\\
k_0\gamma\widetilde{H}_y &= ik_0^2[[\epsilon]]\widetilde{E}_x + \mathrm{diag}(\widetilde{k}_y)\left(\mathrm{diag}(\widetilde{k}_x)\widetilde{E}_y - \mathrm{diag}(\widetilde{k}_y)\widetilde{E}_x\right),
\end{aligned} \tag{3.17}$$

where $\mathrm{diag}(\vec{v})$ is a diagonal matrix whose diagonal is the vector $\vec{v}$. This is the system of equations given in [79], but a very similar system of equations have been given earlier in [120].

From the work of L. Li [121], it is known that some terms in (3.17) does not converge well. Typically, the z component of the D-field given by

$$D_z = \frac{i}{k_0}\left(\frac{\delta}{\delta x}H_y - \frac{\delta}{\delta y}H_x\right) \tag{3.18}$$

is discontinuous and E_z is continuous, meaning that the term $[[1/\epsilon]]\left(\mathrm{diag}(\widetilde{k}_x)\widetilde{H}_y - \mathrm{diag}(\widetilde{k}_y)\widetilde{H}_x\right)$

in (3.17) should be replaced by $[[\epsilon]]^{-1}\left(\operatorname{diag}(\widetilde{k}_x)\widetilde{H}_y - \operatorname{diag}(\widetilde{k}_y)\widetilde{H}_x\right)$. It is more tricky for the terms $[[\epsilon]]\widetilde{E}_x$ and $[[\epsilon]]\widetilde{E}_y$ because, in the general case, E_x, E_y, D_x and D_y are all discontinuous since the boundary between two materials has different orientations. To solve this issue, two strategies are proposed. The first strategy, proposed by L. Li in [61], is to assume that the boundaries are always parallel to one of the two lattice vectors. For a square lattice, the boundaries are parallel to either the x or the y axis. From this assumption, the terms $[[\epsilon]]\widetilde{E}_x$ and $[[\epsilon]]\widetilde{E}_y$ in (3.17) are replaced by terms that use the Laurent's rule (3.15) in one direction and the inverse rule (3.16) in the other direction. In [61], those terms are expressed as $\lfloor\lceil\epsilon\rceil\rfloor\widetilde{E}_x$ and $\lceil\lfloor\epsilon\rfloor\rceil\widetilde{E}_y$. $\lfloor\lceil\epsilon\rceil\rfloor$ means that the Fourier transform along the x dimension is applied to f, then, the Teoplitz matrix, whose elements depend on y, is computed and inverted, and, finally, the Fourier transform along the y dimension is applied to each of the element of the inverted matrix. The operation $\lceil\lfloor\cdot\rfloor\rceil$ is the same except that the dimensions are swapped. The formal definition of the operations $\lfloor\lceil\cdot\rceil\rfloor$ and $\lceil\lfloor\cdot\rfloor\rceil$ is given in [61].

The second strategy, proposed by E. Popov and M. Nevière and fully developed in Chapter 7 of [95], is to decompose the electric and magnetic fields into components that are either normal or tangential to the boundaries, meaning that, at every point in space, the components of the electric and magnetic fields are oriented differently.

In this work, the implementation given by L. Li [61] has been chosen because it can be applied to any permittivity profile without having to determine how the components of the fields are oriented, even if the convergence is lower as shown in Section 7.6.5 in [95]. The system of equations given by L. Li [61] applied to a square lattice is similar to (3.17) and is

$$\begin{aligned}
k_0\gamma\widetilde{E}_x &= k_0^2\widetilde{H}_y - \operatorname{diag}(\widetilde{k}_x)[[\epsilon]]^{-1}\left(\operatorname{diag}(\widetilde{k}_x)\widetilde{H}_y - \operatorname{diag}(\widetilde{k}_y)\widetilde{H}_x\right)\\
k_0\gamma\widetilde{E}_y &= -k_0^2\widetilde{H}_x - \operatorname{diag}(\widetilde{k}_y)[[\epsilon]]^{-1}\left(\operatorname{diag}(\widetilde{k}_x)\widetilde{H}_y - \operatorname{diag}(\widetilde{k}_y)\widetilde{H}_x\right)\\
k_0\gamma\widetilde{H}_x &= -ik_0^2\lceil\lfloor\epsilon\rfloor\rceil\widetilde{E}_y + \operatorname{diag}(\widetilde{k}_x)\left(\operatorname{diag}(\widetilde{k}_x)\widetilde{E}_y - \operatorname{diag}(\widetilde{k}_y)\widetilde{E}_x\right)\\
k_0\gamma\widetilde{H}_y &= ik_0^2\lfloor\lceil\epsilon\rceil\rfloor\widetilde{E}_x + \operatorname{diag}(\widetilde{k}_y)\left(\operatorname{diag}(\widetilde{k}_x)\widetilde{E}_y - \operatorname{diag}(\widetilde{k}_y)\widetilde{E}_x\right).
\end{aligned} \tag{3.19}$$

The equations in [61] are generalized for any lattice and the equations in Chapter 13 of [95] are for any media composed of anisotropic materials and also for any lattice.

The system of equations (3.19) can be expressed as

$$k_0\gamma\begin{pmatrix}\widetilde{E}_x\\ \widetilde{E}_y\end{pmatrix} = \hat{F}\begin{pmatrix}\widetilde{H}_x\\ \widetilde{H}_y\end{pmatrix} \qquad k_0\gamma\begin{pmatrix}\widetilde{H}_x\\ \widetilde{H}_y\end{pmatrix} = \hat{G}\begin{pmatrix}\widetilde{E}_x\\ \widetilde{E}_y\end{pmatrix}. \tag{3.20}$$

Hence, the following eigen-value equation is obtained:

$$\hat{F}\hat{G}\begin{pmatrix}\widetilde{E}_x\\ \widetilde{E}_y\end{pmatrix} = k_0^2\gamma^2\begin{pmatrix}\widetilde{E}_x\\ \widetilde{E}_y\end{pmatrix}. \tag{3.21}$$

It is expected that the eigen-values of (3.21) are proportional to γ^2, since, in a ZSI medium, if an eigen mode has a propagation constant γ, it exists another eigen-mode with the propagation constant $-\gamma$. Moreover, from equations (3.20), flipping the sign of the propagation constant changes the sign of the tangential components of one of the field which is also the property of the eigen-modes in a ZSI medium.

As mentioned earlier, the equation (3.21) admits $2K = 2(2M+1)(2N+1)$ eigen-modes, whose tangential components of the electric and magnetic fields are described by K Fourier components each. In section 3.2.1, the matrices that contain the Fourier coefficients of the electric and magnetic field of all the eigen-modes, are denoted $\hat{E}$ and $\hat{H}$ and their columns are given by

$$\hat{E}_{:m} = \begin{pmatrix} \widetilde{E}_{x,m} \\ \widetilde{E}_{y,m} \end{pmatrix} \qquad \hat{H}_{:m} = \begin{pmatrix} \widetilde{H}_{x,m} \\ \widetilde{H}_{y,m} \end{pmatrix}, \tag{3.22}$$

where $\widetilde{E}_{x,m}$, $\widetilde{E}_{y,m}$, $\widetilde{H}_{x,m}$ and $\widetilde{H}_{y,m}$ are the Fourier coefficients of respectively the fields E_x, E_y, H_x and H_y that describe the eigen-mode m.

3.2.1 Convention for plane waves

Plane waves in an isotropic medium can be described in different ways. In this work, the chosen polarizations are Transvers Magnetic (TM) and Transverse Electric (TE). As shown in section 2.8.1, only TE and TM-polarized plane waves are always orthogonals with each other, but the disadvantage of such choice is the presence of a singularity at $k_x = k_y = 0$.

The description of the plane waves given in this section is a simplification of their description for uniaxial media given in section 2.7.2. The TM-polarized plane waves are described as

$$\vec{E}_{TM} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -k_{/\!/} \end{pmatrix} e^{i(k_x x + k_y y + k_z z)}, \qquad \vec{H}_{TM} = \begin{pmatrix} -\epsilon k_0 s_y \\ \epsilon k_0 s_x \\ 0 \end{pmatrix} e^{i(k_x x + k_y y + k_z z)} \tag{3.23}$$

and the TE-polarized plane waves are given as

$$\vec{E}_{TE} = \begin{pmatrix} \mu k_0 s_y \\ -\mu k_0 s_x \\ 0 \end{pmatrix} e^{i(k_x x + k_y y + k_z z)}, \qquad \vec{H}_{TE} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -k_{/\!/} \end{pmatrix} e^{i(k_x x + k_y y + k_z z)}, \tag{3.24}$$

where k_z follows the dispertion relation

$$k_z^2 + k_{/\!/}^2 = \epsilon\mu k_0^2. \tag{3.25}$$

ϵ and μ are the permettivity and the permeability of the medium respectively. The parameters $k_{/\!/}$, s_x and s_y are given by

$$k_{/\!/} = \sqrt{k_x^2 + k_y^2}, \qquad (s_x, s_y) = \begin{cases} (1,0) & \text{if } k_{/\!/} = 0 \\ \frac{1}{k_{/\!/}}(k_x, k_y) & \text{otherwise} \end{cases} \tag{3.26}$$

The advantage of describing the TM and TE-polarizaed plane wave as done in (3.23) and (3.24) is that those expressions are valid for any k_x, k_y, ϵ and μ with the exception of the case $\epsilon\mu = 0$ with $k_{/\!/} = 0$.

An important quantity for the computation of the reflection and transmission efficiencies is the z-component of the Poynting vector, denoted $P_{z,TM}$ and $P_{z,TE}$ for respectively TM and TE-polarized plane waves, and they are given by

$$P_{z,TM} = \epsilon k_0 k_z \qquad\qquad P_{z,TE} = \mu k_0 k_z. \tag{3.27}$$

If the plane waves described in (3.23) and (3.24) are forward-propagating modes and the medium is passive, the sign of k_z has to be chosen such that the real part of $P_{z,TM}$ and $P_{z,TE}$, and the imaginary part of k_z are positives. For active media, it is more difficult to find the sign of k_z [122].

Usually, it is more convenient to use x and y-polarized plane waves, meaning that the electric field is polarized along x and y respectively, since there is no singularity at normal incidence. If the amplitude of the tangential electric field of the x and y-polarized plane waves is one, x-polarized plane waves are described as

$$\vec{E}_X = \begin{pmatrix} 1 \\ 0 \\ -\frac{k_x}{k_z} \end{pmatrix} e^{i(k_x x + k_y y + k_z z)}, \qquad \vec{H}_X = \frac{1}{\mu k_0 k_z} \begin{pmatrix} -k_x k_y \\ k_x^2 + k_z^2 \\ -k_z k_y \end{pmatrix} e^{i(k_x x + k_y y + k_z z)} \tag{3.28}$$

and y-polarized plane waves are described as

$$\vec{E}_Y = \begin{pmatrix} 0 \\ 1 \\ -\frac{k_y}{k_z} \end{pmatrix} e^{i(k_x x + k_y y + k_z z)}, \qquad \vec{H}_Y = \frac{1}{\mu k_0 k_z} \begin{pmatrix} -(k_y^2 + k_z^2) \\ k_x k_y \\ k_z k_x \end{pmatrix} e^{i(k_x x + k_y y + k_z z)} \tag{3.29}$$

The relationships between the weights of the TM and TE-polarized plane waves, p_{TE} and p_{TM}, and the weights of the x and y-polarized plane waves, q_X and q_Y, are given by

$$\begin{aligned}\begin{pmatrix} p_X \\ p_Y \end{pmatrix} &= \begin{pmatrix} k_z s_x & \mu k_0 s_y \\ k_z s_y & -\mu k_0 s_x \end{pmatrix}\begin{pmatrix} p_{TM} \\ p_{TE} \end{pmatrix} \\ \begin{pmatrix} p_{TM} \\ p_{TE} \end{pmatrix} &= \frac{1}{\mu k_0 k_z}\begin{pmatrix} \mu k_0 s_x & \mu k_0 s_y \\ k_z s_y & -k_z s_x \end{pmatrix}\begin{pmatrix} p_X \\ p_Y \end{pmatrix}\end{aligned} \tag{3.30}$$

X and y-polarized plane waves are orthogonals only if $s_x s_y = 0$.

For section 3.3, it is required to provide for each medium the matrices $\hat{E}$ and $\hat{H}$ that describe the electric and magnetic fields of the eigen-modes. If the medium is homogeneous, the convention chosen in this work is that the eigen-modes 1 to K, where K is the number of Fourier coefficients that describe the fields and is given in (3.12), are TM-polarized plane waves, and the eigen-modes $K+1$ to $2K$ are TE-polarized plane waves. The x and y components of the k-vector of the eigen-modes p and $K+p$ are $\widetilde{k}_{x,p}$ and $\widetilde{k}_{y,p}$. Due to the description of the plane waves given in (3.23) and (3.24), the matrices $\hat{E}$ and $\hat{H}$ that describe the fields of the plane waves propagating in a medium described by the permettivity ϵ and permeability μ are tri-diagonal and are given by

$$\begin{aligned}\hat{E} &= \begin{pmatrix} \operatorname{diag}(\widetilde{k}_z \odot \widetilde{s}_x) & \mu k_0 \operatorname{diag}(\widetilde{s}_y) \\ \operatorname{diag}(\widetilde{k}_z \odot \widetilde{s}_y) & -\mu k_0 \operatorname{diag}(\widetilde{s}_x) \end{pmatrix} \\ \hat{H} &= \begin{pmatrix} -\epsilon k_0 \operatorname{diag}(\widetilde{s}_y) & \operatorname{diag}(\widetilde{k}_z \odot \widetilde{s}_x) \\ \epsilon k_0 \operatorname{diag}(\widetilde{s}_x) & \operatorname{diag}(\widetilde{k}_z \odot \widetilde{s}_y) \end{pmatrix},\end{aligned} \tag{3.31}$$

where $\odot$ is the element-wise product also known as the Hadamard product. The relationship between $\widetilde{s}_x$, $\widetilde{s}_y$, $\widetilde{k}_x$, $\widetilde{k}_y$ and $\widetilde{k}_z$ are equivalent to the ones given in (3.25) and (3.26).

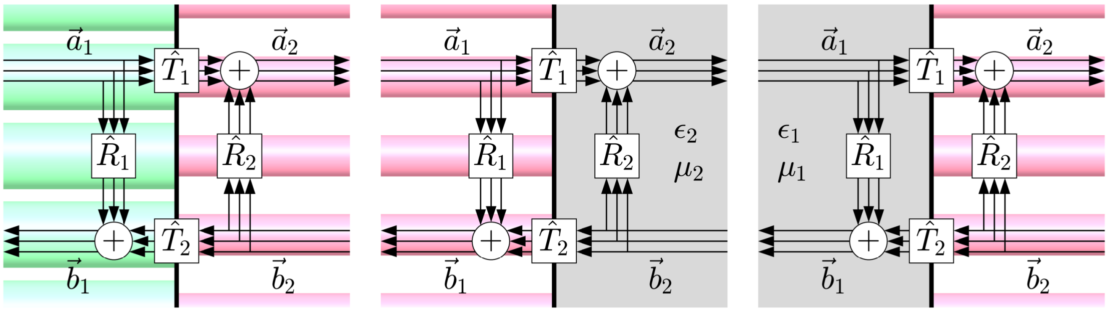
3.3 Computation of the coupling matrices

In this section, the equations for the coupling matrices, that compose the S-matrix, at an interface between two ZSI z-invariant media are developped for the systems illustrated in fig. 3.3. The coupling matrices $\hat{R}_1$, $\hat{R}_2$, $\hat{T}_1$ and $\hat{T}_2$ describe the relationships between the weights of the eigen-modes propagating toward and away of the interface in the following way:

$$\begin{aligned}\hat{T}_1\vec{a}_1 + \hat{R}_2\vec{b}_2 &= \vec{a}_2 \\ \hat{R}_1\vec{a}_1 + \hat{T}_2\vec{b}_2 &= \vec{b}_1,\end{aligned} \tag{3.32}$$

where, as shown in fig. 3.3, $\vec{a}_1$ and $\vec{b}_1$ are the weights of the eigen-modes just before the interface, and $\vec{a}_2$ and $\vec{b}_2$ are the weights of the eigen-modes just after the interface.

The coupling matrices are obtained from the boundary condition stating that the tangential components of the electric and magnetic fields are continuous at the interface. In order to write the boundary condition into a system of equations, the tangential components of the fields of the eigen-modes in each medium are described by two matrices, where the m-th column of those matrices describes the field of the eigen-mode m. For forward-propagating eigen-modes, the matrix $\hat{E}$ describes the tangential components of the electric field and the matrix $\hat{H}$ describes the tangential components of the magnetic field. Since the media considered in this work are ZSI, the backward-propagating eigen-modes are described by $\hat{E}$ and $-\hat{H}$. Sections 3.2 and 3.2.1 give how those matrices are computed for respectively a heterogeneous and a homogeneous medium.



(a) (b) (c)

Figure 3.3 – a) Interface between two heterogenous z-invariant media. $\hat{R}_1$, $\hat{R}_2$, $\hat{T}_1$ and $\hat{T}_2$ are the coupling matrices and $\vec{a}_1$, $\vec{a}_2$, $\vec{b}_1$ and $\vec{b}_1$ are the weights of the eigen-modes. b) Same as fig. 3.3a except that the medium on the right-hand side of the interface is homogeneous. ϵ_2 and μ_2 are respectively the permittivity and the permeability of the homogeneous medium. c) Same as fig. 3.3a except that the medium on the left-hand side of the interface is homogeneous. ϵ_1 and μ_1 are respectively the permittivity and the permeability of the homogeneous medium.

If $\hat{E}_1$ and $\hat{H}_1$ describe the eigen-modes in the medium on the left-hand side of the interface and $\hat{E}_2$ and $\hat{H}_2$ describe the eigen-modes in the medium on the right-hand side of the interface, the boundary condition can be written as

$$\begin{aligned} \hat{E}_1(\vec{a}_1+\vec{b}_1) &= \hat{E}_2(\vec{a}_2+\vec{b}_2) \\ \hat{H}_1(\vec{a}_1-\vec{b}_1) &= \hat{H}_2(\vec{a}_2-\vec{b}_2). \end{aligned} \tag{3.33}$$

As proved in section 3.7.1, the coupling matrices are given by

$$\begin{aligned} \hat{T}_1 &= 2\hat{C}_3\hat{H}_1 & \hat{R}_1 &= \hat{C}_1\hat{T}_1-\hat{I} \\ \hat{R}_2 &= \hat{C}_3(\hat{H}_2-\hat{C}_2) & \hat{T}_2 &= \hat{C}_1(\hat{I}+\hat{R}_2) \end{aligned} \tag{3.34}$$

with

$$\hat{C}_1 = \hat{E}_1^{-1}\hat{E}_2 \qquad \hat{C}_2 = \hat{H}_1\hat{C}_1 \qquad \hat{C}_3 = (\hat{H}_2 + \hat{C}_2)^{-1}. \tag{3.35}$$

Usually, computational instabilites come from the inversion of an ill-conditioned matrix. In equation (3.35), two inversions are present. The first one is the inversion of the matrix $\hat{E}_1$, but, if $\hat{E}_1$ is obtained from the eigen-value equation (3.21), $\hat{E}_1$ is well-conditioned. However, if $\hat{E}_1$ describes the plane waves in a homogeneous medium and k_z of one of the TM-polarized plane waves is zero, the tangential components of the electric field of this plane wave is zero and $\hat{E}_1$ is ill-conditioned. Hence, the equations (3.34) and (3.35) should not be used when the medium on the left-hand side of the interface is homogeneous, which is the case shown in fig. 3.3c. However, those equations can be safely used when the right-hand side of the interface is homogenous, which is the case shown in fig. 3.3b. Since the case shown in fig. 3.3b is the same as the case shown in fig. 3.3c except that the forward and backward directions are reversed, the S-matrix for both cases are composed of the same coupling matrices, but the coupling matrix $\hat{R}_1$ is swapped with $\hat{R}_2$ and $\hat{T}_1$ is swapped with $\hat{T}_2$. For the trivial case when both media are homogeneous, the coupling matrices are diagonal matrices composed of the Fresnel coefficients and they are given in section 3.3.1.

The other matrix inversion in (3.35) is the inversion of the matrix $\hat{H}_2 + \hat{C}_2$. If the smallest eigen-value of this matrix is close to zero, the coupling matrices become very sensitive to a change of the parameters of both media and it means that a surface resonance is present. Metal usually needs to be present in order to have surface resonances and, in this case, such resonance is called surface plasmon resonance.

In equations (3.34) and (3.35), the coupling matrices are computed using two matrix inversions and six matrix multiplications. If most eigen-modes in both media can be neglected, meaning that the coupling matrices are much smaller than the matrices $\hat{E}$ and $\hat{H}$, the computational time of the four multiplications in (3.34) becomes negligible. Hence, the computation of the S-matrix is reduced by approximately a factor two. For comparison, the computation of the T-matrix $\hat{T}$ as presented in [94] can be written as

$$\hat{T} = \begin{pmatrix} \hat{E}_2 & \hat{E}_2 \\ \hat{H}_2 & -\hat{H}_2 \end{pmatrix}^{-1} \begin{pmatrix} \hat{E}_1 & \hat{E}_1 \\ \hat{H}_1 & -\hat{H}_1 \end{pmatrix}, \tag{3.36}$$

which is an inversion and multiplication of matrices that are two times larger. Assuming that the matrix inversion is done through the Gauss-Jordan elimination algorithm, the complexity of both matrix inversion and multiplication is $O(n^3)$ and the computation of the T-matrix using equation (3.36) is two times slower that the computation of the coupling matrices using equations (3.34) and (3.35). As a side note, if the medium on the right-hand side of the interface is homogeneous and k_z of one the plane waves is zero, the matrix that is inverted in (3.36) is ill-conditioned.

3.3.1 Coupling matrices at an interface between two homogeneous media

For an interface between two homogeneous media, the coupling matrices are diagonal matrices containing the Fresnel coefficients, which are given for uniaxial media by equations (2.61) and (2.63) in section 2.7.2. If the convention presented in section 3.2.1 is used, the coupling matrices are

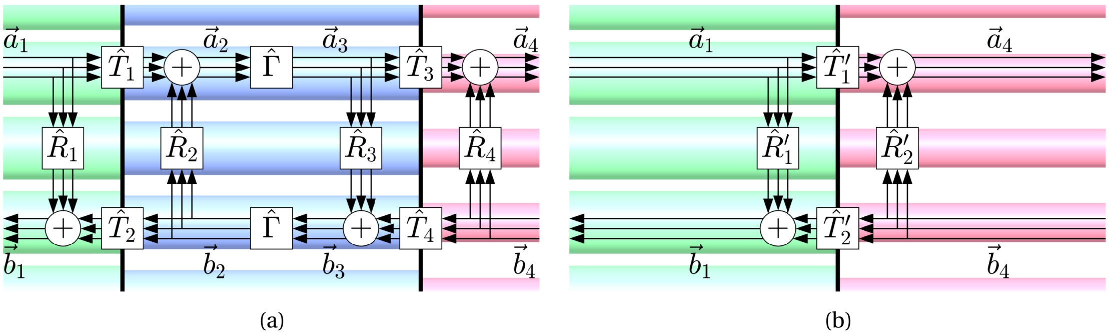
$$\begin{aligned} \hat{T}_1 &= \operatorname{diag}\left(\begin{pmatrix} \vec{t}_{1,TM} \\ \vec{t}_{1,TE} \end{pmatrix}\right) & \hat{R}_1 &= \operatorname{diag}\left(\begin{pmatrix} \vec{r}_{TM} \\ \vec{r}_{TE} \end{pmatrix}\right) \\ \hat{R}_2 &= -\hat{R}_1 & \hat{T}_2 &= \operatorname{diag}\left(\begin{pmatrix} \vec{t}_{2,TM} \\ \vec{t}_{2,TE} \end{pmatrix}\right), \end{aligned} \tag{3.37}$$

where the m-th element of $\vec{t}_{1,TM}$, $\vec{t}_{1,TE}$, $\vec{t}_{2,TM}$, $\vec{t}_{2,TE}$, $\vec{r}_{TM}$ and $\vec{r}_{TE}$ are given by

$$\begin{aligned} t_{1,TM,m} &= \frac{2\epsilon_1 k_{z,1,m}}{\epsilon_1 k_{z,2,m} + \epsilon_2 k_{z,1,m}} & t_{1,TE,m} &= \frac{2\mu_1 k_{z,1,m}}{\mu_2 k_{z,1,m} + \mu_1 k_{z,2,m}} \\ t_{2,TM,m} &= \frac{2\epsilon_2 k_{z,2,m}}{\epsilon_1 k_{z,2,m} + \epsilon_2 k_{z,1,m}} & t_{2,TE,m} &= \frac{2\mu_2 k_{z,2,m}}{\mu_2 k_{z,1,m} + \mu_1 k_{z,2,m}} \\ r_{TM,m} &= \frac{\epsilon_1 k_{z,2,m} - \epsilon_2 k_{z,1,m}}{\epsilon_1 k_{z,2,m} + \epsilon_2 k_{z,1,m}} & r_{TE,m} &= \frac{\mu_2 k_{z,1,m} - \mu_1 k_{z,2,m}}{\mu_2 k_{z,1,m} + \mu_1 k_{z,2,m}}. \end{aligned} \tag{3.38}$$

ϵ_p and μ_p is respectively the permittivity and permeability of medium p and $k_{z,p,m}$ is the z component of the k-vector of the m-th plane wave in medium p. Medium 1 and 2 are respectively on the left and right-hand side of the interface.

3.4 Layer reduction



(a) (b)

Figure 3.4 – a) A multi-layer structure before layer reduction with all the coupling matrices and the propagation operators that are needed for the computation of the coupling matrices shown in fig. 3.4b. b) The same multi-layer structure as in fig. 3.4a except that the layer at the center of fig. 3.4a is reduced into a single interface.

The operation, called layer reduction, transforms a layer into an interface, meaning it transforms the system shown in fig. 3.4a into the system shown in fig. 3.4b. During this operation, three groups of matrices are computed. The first group is composed of the coupling matrices at the interface after layer reduction and those matrices are denoted by $\hat{T}'_1$, $\hat{R}'_1$, $\hat{T}'_2$ and $\hat{R}'_2$ as shown in fig. 3.4b. If the fields inside the reduced layer are needed, the relationships between the weights of the eigen-modes propagating toward the interface, which are given by $\vec{a}_1$ and $\vec{b}_4$, and the weights of the eigen-modes propagating inside the reduced layer, given by $\vec{a}_2$ and $\vec{b}_3$, are required. Those relationships are described by the second group of matrices, denoted $\hat{A}_1$, $\hat{B}_1$, $\hat{A}_2$ and $\hat{B}_2$, and are written as

$$\begin{aligned} \hat{A}_1\vec{a}_1 + \hat{A}_2\vec{b}_4 &= \vec{a}_2 \\ \hat{B}_1\vec{a}_1 + \hat{B}_2\vec{b}_4 &= \vec{b}_3. \end{aligned} \tag{3.39}$$

As shown in section 3.4, those eight matrices are given by

$$\begin{aligned} \hat{A}_1 &= (\hat{I} - \hat{M})^{-1}\hat{T}_1 & \hat{A}_2 &= (\hat{I} - \hat{M})^{-1}\hat{R}_2\hat{\Gamma}\hat{T}_4 \\ \hat{B}_1 &= \hat{R}_3\hat{\Gamma}\hat{A}_1 & \hat{B}_2 &= \hat{T}_4 + \hat{R}_3\hat{\Gamma}\hat{A}_2 \\ \hat{T}'_1 &= \hat{T}_3\hat{\Gamma}\hat{A}_1 & \hat{T}'_2 &= \hat{T}_2\hat{\Gamma}\hat{B}_2 \\ \hat{R}'_1 &= \hat{R}_1 + \hat{T}_2\hat{\Gamma}\hat{B}_1 & \hat{R}'_2 &= \hat{R}_4 + \hat{T}_3\hat{\Gamma}\hat{A}_2 \end{aligned} \tag{3.40}$$

with

$$\hat{M} = \hat{R}_2\hat{\Gamma}\hat{R}_3\hat{\Gamma}. \tag{3.41}$$

The only matrix that is inverted is the matrix $\hat{I} - \hat{M}$, meaning that, if one of the eigen-values of the matrix $\hat{I} - \hat{M}$ is close to zero, the layer is resonant. It is more meaningful to consider the eigen-values of the matrix $\hat{M}$, known as the round-trip matrix [123, 124], since the matrix $\hat{M}$ describes the loops in fig. 3.4a. The eigen-vectors of the matrix $\hat{M}$ describe the self-coupling modes presented in chapter 5 and, in this work, a self-coupling mode is considered resonant if the associated eigen-value is 0.5 or above. If the eigen-value is one, the light is perfectly trapped inside the layer.

For the design of a structure, it is sometimes needed to get the response of the structure for multiple layer thicknesses, meaning that the layer reduction operation is applied multiple times while only changing the propagation operator $\hat{\Gamma}$. In this case, it greatly shortens the computation time if only the eigen-modes that contribute to the power flow inside the layer are kept. The numbers of eigen-modes kept in the first and second layer shown in fig. 3.4a are respectively u and v. The size of the matrices $\hat{T}_1$, $\hat{R}_1$, $\hat{T}_2$ and $\hat{R}_2$ is respectively v-by-u, u-by-u, u-by-v and v-by-v. The same can be done for the interface between the second and third medium. Those reduced coupling matrices can be directly used in equations (3.40) and (3.41) and the obtained matrices have the expected dimensions without any additional manipulation.

It is possible that the two interfaces present in fig. 3.4a were a stack of layers that has been previously reduced and the weights of the eigen-modes inside those layers are needed later. Let $\vec{c}_1$ and $\vec{c}_3$ be the weights of the eigen-modes in a reduced layer replaced by respectively the first and second interface in fig. 3.4a and they are obtained through the equations

$$\begin{aligned} \hat{C}_1\vec{a}_1 + \hat{C}_2\vec{b}_2 &= \vec{c}_1 \\ \hat{C}_3\vec{a}_3 + \hat{C}_4\vec{b}_4 &= \vec{c}_2, \end{aligned} \tag{3.42}$$

where $\hat{C}_m$ are equivalent to the matrices $\hat{A}$ and $\hat{B}$ in (3.39). If the matrices $\hat{A}_1$, $\hat{A}_2$, $\hat{B}_1$ and $\hat{B}_2$ are stored, no additional operations are required. However, the weights of the eigen-modes present in the reduced layer are not needed later, it may be advantageous to transforms the matrices $\hat{C}_m$ into the matrices $\hat{C}'_m$ such that

$$\begin{aligned} \hat{C}'_1\vec{a}_1 + \hat{C}'_2\vec{b}_4 &= \vec{c}_1 \\ \hat{C}'_3\vec{a}_1 + \hat{C}'_4\vec{b}_4 &= \vec{c}_2. \end{aligned} \tag{3.43}$$

By combining equations (3.39) and (3.42), the matrices $\hat{C}'_m$ are given by

$$\begin{aligned} \hat{C}'_1 &= \hat{C}_1 + \hat{C}_2\hat{\Gamma}\hat{B}_1 & \qquad \hat{C}'_2 &= \hat{C}_2\hat{\Gamma}\hat{B}_2 \\ \hat{C}'_3 &= \hat{C}_3\hat{\Gamma}\hat{A}_1 & \qquad \hat{C}'_4 &= \hat{C}_3\hat{\Gamma}\hat{A}_2 + \hat{C}_4 \end{aligned} \tag{3.44}$$

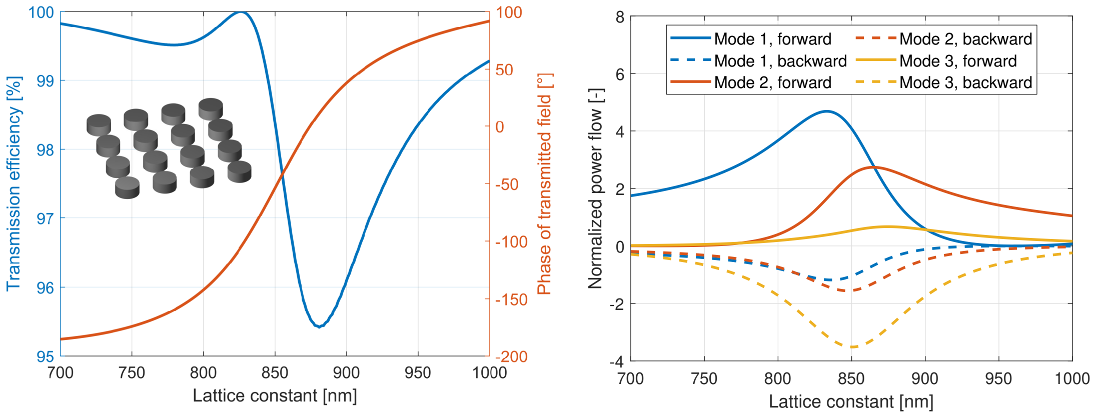
The matrices $\hat{C}'_1$ are the last group of matrices that can be obtained during layer reduction.

3.5 Mode filtering and contribution of the eigen-modes to the power flow

Mode filtering is a key element of the Fourier modal method provided in this chapter since it divides the computation time required for getting the coupling matrices describing an interface by a factor two. But, more importantly, the computation time required for the reduction of a layer into an interface becomes negligible. Since the coupling matrices and the eigen-modes does not depend on the layer thickness, a change of the layer thickness requires only the layer reduction operation, which can be done very fast due to mode filtering. The computation of the coupling matrices and the layer reduction are given in sections 3.3 and 3.4 respectively.

In this section, two criteria for mode filtering is proposed. The first one is based on the contribution of the eigen-modes to the power flow. This contribution is expressed differently if the mode is propagating, evanescent or complex and the equations are provided later in this section. The second criterion is based on the imaginary part of the propagation constant, which determines how fast the amplitude of the eigen-mode decreases along z. The mode filtering is applied to the Huygens' metasurface proposed in [37,44]. The Huygens' metasurface is composed of an array of silicon cylinders embedded in glass. The cylinders diameter and height are respectively 524 nm and 243 nm. The response for different lattice constants shown in fig. 3.5a is obtained for a wavelength of 1477 nm at normal incidence. The refractive index of glass and silicon is 1.44 and 3.48 respectively.

The Huygens' metasurface is an interesting candidate to apply mode filtering for several reasons. First, the Huygens' metasurface is a double resonant multi-mode metasurface, meaning that multiple propagating eigen-modes are present in the metasurface and, because it is resonant, the mode weights change rapidly as shown in fig. 3.5b, where the contribution of the propagating eigen-modes to the power flow is plotted. It is not possible to identify the two resonances in this figure, but the self-coupling mode presented in chapter 5 can. Second, the thickness of the metasurface is small compared to the wavelength. Hence, evanescent and complex modes are needed in order to accurately compute the response of the metasurface. Finally, the metasurface is made of silicon cylinders, which is a high refractive index material, and the importance of complex modes is higher in a medium with large refractive index difference [99]. Complex modes are more difficult to analysis than propagating and evanescent modes.



(a) (b)

Figure 3.5 – (a) Transmission efficiency and the phase of the transmitted plane wave for the Huygens' metasurface shown in the inset. The Huygens' metasurface, which is the same as in [37, 44], is composed of silicon cylinders with a diameter of 534 nm and a height of 243 nm embedded in glass. The incident plane wave is x-polarized and comes at normal incidence. The wavelength is 1477 nm. b) Power flow contribution of the main propagating eigen-modes. If the contribution is one, it means that the contribution is equivalent to the power flow of the incident plane wave. The propagation constant of modes 1, 2 and 3 is around 12.5, 5.9 and 5.1 respectively.

The impact of an eigen-mode propagating in a lossless layer on the response of the whole system is the contribution of the eigen-mode to the power flow since the eigen-modes are orthogonal in an lossless medium. Hence, an eigen-mode which has a negligible contribution to the power flow can be neglected without impacting the response the system. This contribution is obtained throught the Poynting operation (chapter 2) and defined as

$$[\psi_m|\psi_n] = \frac{1}{2|\Lambda|}\iint_\Lambda (\vec{E}_m \times \vec{H}_n^* + \vec{E}_n \times \vec{H}_m^*)\cdot \vec{n}\,ds, \tag{3.45}$$

where ψ_m is an eigen-mode described by the fields $\vec{E}_m$ and $\vec{H}_m$, Λ is the unit cell and $\vec{n}$ is the unit vector perpendicular to the surface. The expression $[\psi_m|\psi_m]$ is the integration of the z-component of the Poynting operation over the unit cell and $\mathrm{Re}\{[\psi_m|\psi_m]\}$ is the contribution of the eigen-modes ψ_m to the power flow assuming that the eigen-mode is orthogonal to the other eigen-modes present in the layer. If the eigen-mode is propagating or evanescent, the contribution of the eigen-mode to the power flow is given by

$$\mathrm{Re}\{[a\psi + b\psi^-|a\psi + b\psi^-]\} = \mathrm{Re}\left\{\left(|a|^2 - |b|^2 + 2i\,\mathrm{Im}\{\bar{a}b\}\right)[\psi|\psi]\right\}, \tag{3.46}$$

where a and b are the weigths of, respectively, the forward and backward-propagating modes at the same z position, and the operator $(\cdot)^-$ flips the propagation direction of a mode. The derivation of equation (3.46) is based on the properties of the Poynting operation (section 2.3) and, if it exists two modes with the same propagation constant, both modes are assumed to be orthogonal. If the mode is normalized (section 2.8.2), meaning that $[\psi|\psi]$ is 1 for a propagating mode and $\pm i$ for an evanescent mode, equation (3.46) becomes

$$\mathrm{Re}\{[a\psi + b\psi^-|a\psi + b\psi^-]\} = |a|^2 - |b|^2 \tag{3.47}$$

for a propagating mode, and

$$\mathrm{Re}\{[a\psi + b\psi^-|a\psi + b\psi^-]\} = -2\,\mathrm{Im}\{\bar{a}b\}\,\mathrm{Im}\{[\psi|\psi]\} \tag{3.48}$$

for an evanescent mode. The term $\mathrm{Im}\{[\psi|\psi]\}$ in (3.48) is either 1 or -1. It is possible to get the sign of $\mathrm{Im}\{[\psi|\psi]\}$ by looking at the reflection coefficient r at the interface that relates the evanescent eigen-modes with its backward-propagating counterpart. If the evanescent mode is the only modes propagating toward the interface, the power flow related to this mode has to go toward the interface, meaning that the sign of $\mathrm{Im}\{[\psi|\psi]\}$ is the inverse of the sign of $\mathrm{Im}\{r\}$. For example, if $b = ra$, meaning that the evanescent mode is on the left-hand side of the interface, equation (3.48) becomes

$$\mathrm{Re}\{[a\psi + b\psi^- | a\psi + b\psi^-]\} = -2|a|^2\,\mathrm{Im}\{r\}\,\mathrm{Im}\{[\psi|\psi]\}, \tag{3.49}$$

which has to be positive.

In lossless heterogeneous media, if a complex mode, called ψ_1, with the propagation constant γ exists, a complex mode, called ψ_2, with propagation constant $-\bar{\gamma}$ also exists. Since ψ_1 and ψ_2 are not orthogonal with each other, the contribution of the eigen-modes ψ_1 and ψ_2 to the power flow cannot be decoupled. When complex modes are orthonormalized as shown in section 2.8.3, the Poynting operation applied to those modes gives

$$[\psi_1|\psi_2] = i \qquad [\psi_1|\psi_2^-] = 0. \tag{3.50}$$

Moreover, complex modes are self-orthogonal meaning that

$$[\psi_1|\psi_1] = 0 \qquad [\psi_2|\psi_2] = 0. \tag{3.51}$$

Hence, the contribution of the eigen-modes ψ_1 and ψ_2 to the power flow is given by

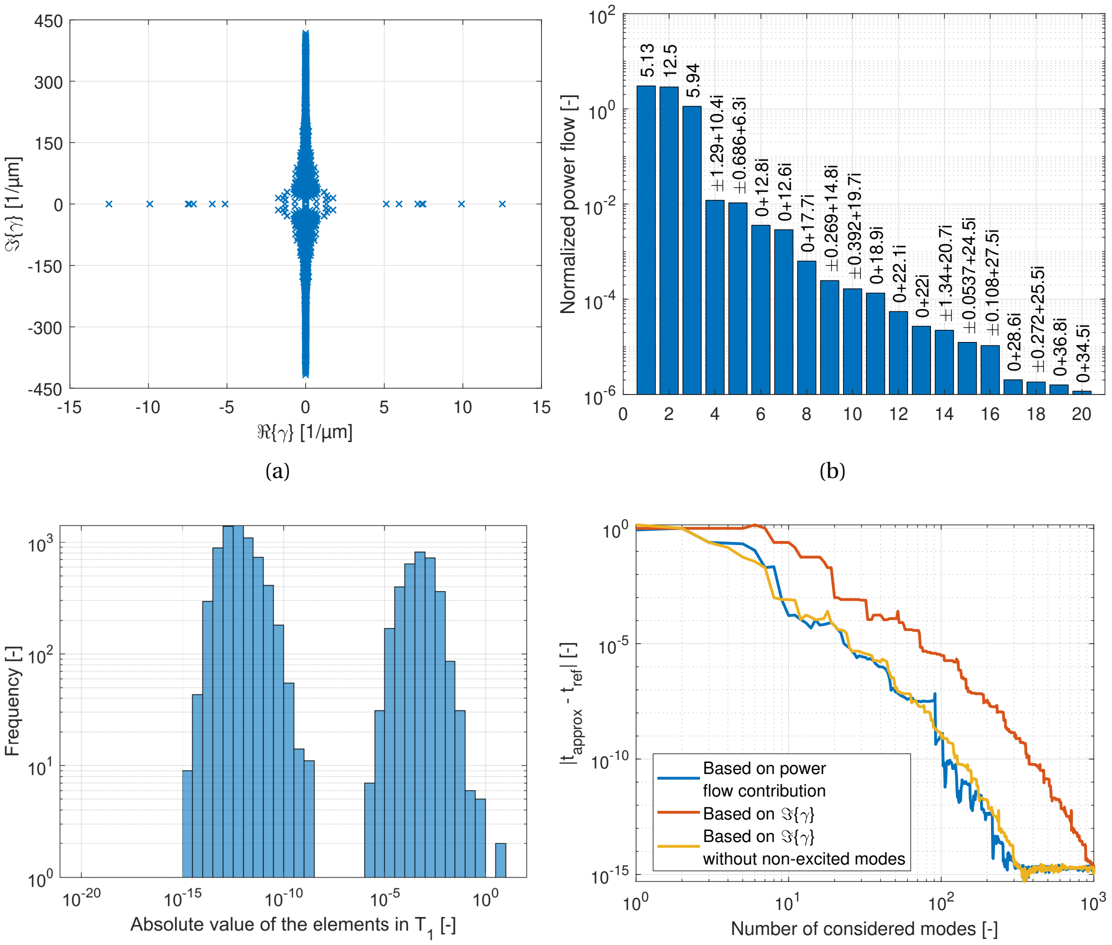
$$\mathrm{Re}\{[a_1\psi_1 + a_2\psi_2 + b_1\psi_1^- + b_2\psi_2^- | a_1\psi_1 + a_2\psi_2 + b_1\psi_1^- + b_2\psi_2^-]\} = -2\,\mathrm{Im}\{\bar{a}_1 b_2 + \bar{a}_2 b_1\}. \tag{3.52}$$

As mentioned earlier, the mode filtering is applied to the Huygens' metasurface of [37, 44]. For the computation of the eigen-modes, 81 Fourier coefficients per dimension are taken into account, meaning that the x and y compononents of the fields for an eigen-mode are described by 6 561 Fourier coefficients each and 13 122 forward-propagating eigen-modes are obtained. For a lattice constant of 852 nm, which is also used in section 5.3.1, the propagation constant of the eigen-modes, both backward and forward-propagating, is given in fig. 3.6a. The large majority of the modes are either evanescent or complex, leaving only 11 propagating modes per propagation direction, which includes three pairs of eigen-modes with the same propagationg constant. It is expected to have pairs of eigen-modes due to the symmetry of the permittivity profile. After rotating the eigen-modes as shown in section 2.8.4 such that x-polarized light excites only one mode of the pair and the y-polarized light excites only the other, the contribution of the eigen-modes to the power flow is computed using equations (3.47), (3.48) and (3.52) and the results for the main modes are shown in fig. 3.6b. The incident plane wave is x-polarized. Each bar is the contribution of a forward-propagating mode and its backward-propagating counterpart, if the mode is propagating or evanescent, or the contribution of two forward and two backward-propagating complex mode. A contribution of one means that the contribution is equal to the power flow of the incident plane wave. The three main contributions are from propagating modes, which makes this metasurface a multi-mode metasurface. The two following main contributions are from complex modes, meaning they can be more important than the evanescent modes in some cases.

The blue curve in fig. 3.6d is the distance in the complex plane between the amplitude of the transmitted plane wave when only the n eigen-modes that contribute the most to the power flow are retained and and when all the modes are considered. By retaining only 10 eigen-modes, the error on the transmitted amplitude is in the order of 10^{-4}. With 100 eigen-modes, the error drops to 10^{-10} and it does not decrease when retaining more than 300 modes. However, the criterion for mode filtering based on the contribution of the eigen-modes to the power flow suffers from two issues. The main one is that the mode weights are required, meaning that the mode can be filtered once the response of the system is known. In the case when the mode filtering is done for a specific case and, then, the response of the system is computed for larger layer thicknesses, which means that the impact of the evanescent and complex modes on the response of the system is reduced, it is possible that an eigen-mode which needs to be considered is filtered out because it had no impact in the specific case. As an example, for the Huygens' metasurface, the weight of the mode 1 for both propagation directions for a lattice constant of around 960 nm is very close to zero, meaning that, as shown in fig. 3.5b, the contribution of mode 1 to the power flow is nearly zero and mode 1 may be filtered out even if this mode is required to compute the response of the system for a different layer thickness.

A more simple way of choosing which eigen-modes can be neglected is to look at the imaginary part of the propagation constant, which gives how the amplitude of the eigen-modes decreases while propagating. By considering only the n eigen-modes with the lowest imaginary part of the propagationg constant, the obtained error is given by the red curve in fig. 3.6d. For the metasurface considered in this section, the number of modes required to have the same error as the case when the criteria is based on the power flow contribution is increased by a factor three. The advantage of this approach is that it does not depend on the coupling matrices.

By looking how the incident x-polarized plane wave excites the eigen-modes in the layer, which is given by the vector $\vec{T}_1$ shown in fig. 3.4a, it is possible to reduce the number of considered eigen-modes for the same error. Because only the x-polarized plane wave is considered, $\vec{T}_1$ is a vector. As shown in fig. 3.6c, the eigen-modes can be divided into two groups, where one of them are excited in a negligible way. By neglecting those eigen-modes that are not excited and, then, considering the eigen-modes whose propagation constant has the lowest imaginary part, the number of eigen-modes that need to be considered in order get a similar error is equivalent to the case where the mode filtering is based on the power flow contribution. It is expected to get such results because the contribution to the power flow of complex and evanescent waves are proportional to $e^{-\mathrm{Im}\{\gamma\}h}$, where γ is the propagation constant of the mode and h is the thickness of the layer, since the weights of the modes present in (3.48) and (3.52) are obtained at the same position on the z-axis.



(a) (b)

(c) (d)

Figure 3.6 – a) Propagation constant of the eigen-modes present in the Huygens' metasurface shown in fig. 3.5a with a lattice constant of 852 nm. b) Contribution of the main eigen-modes to the power-flow. The number at the top of the bars is the propagation constant of the eigen-mode. c) Histogram of the absolute value of the elements in $\vec{T}_1$, which is the vector that relates the incident x-polarized plane wave to the excitation of the eigen-modes inside the metasurface. d) Error on the transmitted field when the n more important eigen-modes are considered. The importance of an eigen-mode is based on three different criteria and the error is defined as the distance on the complex plane between the transmitted field with and without approximation.

3.6 Conclusion

The Fourier modal method proposed in this chapter is divided into three operations: the computation of the eigen-modes, the computation of the S-matrix at an interface and the reduction of a layer, which is a z-invariant heterogeneous medium delimited by two interfaces. The computation of the eigen-modes follows what has been done in [61] and the last two operations are specific to this work.

For the second operation, we provide a set of equations that allows us to efficiently compute the S-matrix, composed by the coupling matrices, by taking into account the property of the eigen-modes in a ZSI medium. Many implementations of the Fourier modal method do not compute the S-matrix for each interface [80,94,95], even though the coupling matrices at the different interfaces contain information that can greatly facilitate the design of metasurfaces. In this work, this information is used for the design of anti-reflective metasurfaces in chapter 4 and the computation of the self-coupling modes in chapter 5.

The third operation is the reduction of a layer and this is an important feature of our Fourier modal method. Since the propagation of an eigen-mode is fully described by its propagation constant, it is possible to compute the response of a metasurface for different thicknesses by reducing the layer each time, assuming that the coupling matrices at the interfaces and the propagation constant of the eigen-modes are known. Moreover, we show that the computational effort for the layer reduction can be drastically reduced by filtering the eigen-modes, which means that the eigen-modes that do not contribute to the power flow within the metasurface are neglected. As a result, after simulating the metasurface once, the response of the metasurface with a different thickness can be obtained in a few milliseconds.

The operations provided by our Fourier modal method can be used in different orders, allowing greater flexibility in its implementation. Hence, a multi-layer system can be simulated in such way that only the information necessary for the design process is obtained. Such information includes the coupling matrices and the propagation constant of the eigen-modes when the response of the system for different layer thicknesses is desired, or the field profile of the modes when the system is optimized using the adjoint method (chapter 6).

3.7 Proofs

3.7.1 Proof of the coupling matrices at an interface between two heterogeneous media

In this section, the equations (3.34) and (3.35), which are

$$\begin{aligned} \hat{T}_1 &= 2\hat{C}_3\hat{H}_1 & \hat{R}_1 &= \hat{C}_1\hat{T}_1 - \hat{I} \\ \hat{R}_2 &= \hat{C}_3(\hat{H}_2 - \hat{C}_2) & \hat{T}_2 &= \hat{C}_1(\hat{I} + \hat{R}_2) \\ \hat{C}_1 &= \hat{E}_1^{-1}\hat{E}_2 & \hat{C}_2 &= \hat{H}_1\hat{C}_1 \\ \hat{C}_3 &= (\hat{H}_2 + \hat{C}_2)^{-1}, \end{aligned} \tag{3.53}$$

are proven from the boundary condition (3.33):

$$\begin{aligned} \hat{E}_1(\vec{a}_1 + \vec{b}_1) &= \hat{E}_2(\vec{a}_2 + \vec{b}_2) \\ \hat{H}_1(\vec{a}_1 - \vec{b}_1) &= \hat{H}_2(\vec{a}_2 - \vec{b}_2). \end{aligned} \tag{3.54}$$

In order to get the matrices in (3.53), the objective is to transform the equations (3.54) in the form

$$\begin{aligned} \hat{T}_1\vec{a}_1 + \hat{R}_2\vec{b}_2 &= \vec{a}_2 \\ \hat{R}_1\vec{a}_1 + \hat{T}_2\vec{b}_2 &= \vec{b}_1, \end{aligned} \tag{3.55}$$

which is equation (3.32).

From the first equation in (3.54), $\vec{b}_1$ is given by

$$\vec{b}_1 = \hat{C}_1(\vec{a}_2 + \vec{b}_2) - \vec{a}_1. \tag{3.56}$$

Replacing $\vec{b}_1$ in the second equation in (3.54) by its expression in (3.56) gives

$$2\hat{H}_1\vec{a}_1 + (\hat{H}_2 - \hat{H}_1\hat{C}_1)\vec{b}_2 = (\hat{H}_2\vec{a}_2 + \hat{H}_1\hat{C}_1)\vec{a}_2. \tag{3.57}$$

By comparing this equation with the first equation in (3.55), the expressions of $\hat{T}_1$ and $\hat{R}_2$ given in (3.53) are found. The matrices $\hat{R}_1$ and $\hat{T}_2$ are obtained after replacing $\vec{a}_2$ in (3.56) by its expression given in (3.55), giving

$$\vec{b}_1 = (\hat{C}_1\hat{T}_1 - \hat{I})\vec{a}_1 + \hat{C}_1(\hat{R}_2 + \hat{I})\vec{b}_2, \tag{3.58}$$

and comparing the resulting equation with the second equation in (3.55).

3.7.2 Proof of the layer reduction

In this section, the expressions given in (3.40) and (3.41), which are

$$
\begin{aligned}
\hat{A}_1 &= (\hat{I}-\hat{M})^{-1}\hat{T}_1 & \hat{A}_2 &= (\hat{I}-\hat{M})^{-1}\hat{R}_2\hat{\Gamma}\hat{T}_4 \\
\hat{B}_1 &= \hat{R}_3\hat{\Gamma}\hat{A}_1 & \hat{B}_2 &= \hat{T}_4+\hat{R}_3\hat{\Gamma}\hat{A}_2 \\
\hat{T}_1' &= \hat{T}_3\hat{\Gamma}\hat{A}_1 & \hat{T}_2' &= \hat{T}_2\hat{\Gamma}\hat{B}_2 \\
\hat{R}_1' &= \hat{R}_1+\hat{T}_2\hat{\Gamma}\hat{B}_1 & \hat{R}_2' &= \hat{R}_4+\hat{T}_3\hat{\Gamma}\hat{A}_2 \\
\hat{M} &= \hat{R}_2\hat{\Gamma}\hat{R}_3\hat{\Gamma}
\end{aligned}
\tag{3.59}
$$

are proven knowing that

$$\hat{T}_1'\vec{a}_1+\hat{R}_2'\vec{b}_4 = \vec{a}_4 \tag{3.60a}$$
$$\hat{R}_1'\vec{a}_1+\hat{T}_2'\vec{b}_4 = \vec{b}_1 \tag{3.60b}$$
$$\hat{A}_1\vec{a}_1+\hat{A}_2\vec{b}_4 = \vec{a}_2 \tag{3.60c}$$
$$\hat{B}_1\vec{a}_1+\hat{B}_2\vec{b}_4 = \vec{b}_3 \tag{3.60d}$$
$$\hat{T}_1\vec{a}_1+\hat{R}_2\hat{\Gamma}\vec{b}_3 = \vec{a}_2 \tag{3.60e}$$
$$\hat{R}_1\vec{a}_1+\hat{T}_2\hat{\Gamma}\vec{b}_3 = \vec{b}_1 \tag{3.60f}$$
$$\hat{T}_3\hat{\Gamma}\vec{a}_2+\hat{R}_4\vec{b}_4 = \vec{a}_4 \tag{3.60g}$$
$$\hat{R}_3\hat{\Gamma}\vec{a}_2+\hat{T}_4\vec{b}_4 = \vec{b}_3. \tag{3.60h}$$

Equations (3.60c) and (3.60d) come from (3.39) and the other equations in (3.60) are represented in fig. 3.4 with $\vec{b}_2 = \hat{\Gamma}\vec{b}_3$ and $\vec{a}_3 = \hat{\Gamma}\vec{a}_2$.

By replacing $\vec{b}_3$ in equation (3.60e) by its expression in (3.60h), the following equation is obtained:

$$\hat{T}_1\vec{a}_1+\hat{R}_2\hat{\Gamma}\hat{T}_4\vec{b}_4 = (\hat{I}-\hat{M})\vec{a}_2. \tag{3.61}$$

By comparing this equation with equation (3.60c), the expressions of $\hat{A}_1$ and $\hat{A}_2$ are found. Replacing $\vec{a}_2$ in equations (3.60g) and (3.60h) by its expression in (3.60h) gives

$$
\begin{aligned}
\hat{T}_3\hat{\Gamma}\hat{A}_1\vec{a}_1+(\hat{T}_3\hat{\Gamma}\hat{A}_2\vec{+}\hat{R}_4)\vec{b}_4 &= \vec{a}_4 \\
\hat{R}_3\hat{\Gamma}\hat{A}_1\vec{a}_1+(\hat{R}_3\hat{\Gamma}\hat{A}_2+\hat{T}_4)\vec{b}_4 &= \vec{b}_3
\end{aligned}
\tag{3.62}
$$

By comparing those equations with equations (3.60b) and (3.60d), the expressions of $\hat{T}_1'$, $\hat{R}_2'$, $\hat{B}_1$ and $\hat{B}_2$ are obtained.

After replacing $\vec{b}_3$ in (3.60f) by its expression in (3.60d), the following equation is obtained:

$$(\hat{R}_1 + \hat{T}_2\hat{\Gamma}\hat{B}_1)\vec{a}_1 + \hat{T}_2\hat{\Gamma}\hat{B}_2\vec{b}_4 = \vec{b}_1, \tag{3.63}$$

The expressions of the matrices $\hat{R}_1'$ and $\hat{T}_2'$ are obtained by comparing this equation with equation (3.60b).

4 Single-mode metasurface

4.1 Introduction

Single-mode dielectric metasurfaces include periodic zeroth-order gratings where two eigenmodes, one per polarization, propagate inside it, and any aperiodic metasurfaces that use zeroth-order grating as building blocks. Three main groups of dielectric single-mode metasurfaces are present in the literature. The first group is about zeroth-order gratings and they can be used as an anti-reflective layer [125–127] or to change the polarization state of light [51, 128]. The second group is about aperiodic metasurfaces which are composed of cylinders with various dimensions [26–30]. The third group is about aperiodic metasurfaces composed of ellipses or rectangles and their response is based on the Pancharatnam–Berry phase [31–35, 38]. The second and third group are generally used as holograms or metalenses.

The design of the aperiodic metasurfaces mentioned earlier is based on the response of zeroth-order gratings. Those aperiodic metasurfaces are composed of cylinders with various cross-section and the assumption done during the first iteration of the design is that the response at the position of each cylinder is the same response as a zeroth-order grating composed with this cylinder. This assumption can be done because the eigen-modes are spatially localized into the cylinders, meaning that a variation of the dimensions of the neighboring cylinders has only a weak impact on the behavior of the eigen-mode propagating in the cylinder. In the case of an important variation of cylinder dimensions, which typically occurs when the light is deflected to a large angle, a second design iteration is performed, where a group of structure are simulated and optimized. Such design techniques are presented in [84].

In this chapter, different design techniques are proposed for the design of single-mode metasurfaces. The first set of design techniques is for the design of ideal metasurfaces composed of cylinders with an elliptical or rectangular cross-section. The properties of ideal metasurfaces are that the structures does not reflect or absorb light and the two orthogonal linear polarizations that are transmitted unchanged, called the eigenpolarization (section 8.5 in [129]), excite a single propagating eigen-mode inside the metasurface each. Ideal metasurfaces are equivalent to an ideal waveplate. The difference between an ideal metasurface and a real metasurface is discussed in detail in section 4.2.1. The principal result is that the main difference is the reflection occurring at the two interfaces of the metasurface, but the reflection is low enough such that the concept of ideal metasurface can be used during the first iteration in the design process.

As shown in section 4.2 and proven in sections 4.7.1 and 4.7.2, the parameters of the ideal metasurface can be directly obtained from the desired functionalities of the metasurface using a simple set of equations that greatly simplifies the design procedure and allows to probe efficiently the design space. The standard approach is to use the concept of the Poincaré sphere in the design process [36, 130] by looking at trajectories, which can be complicate in some cases [131]. In comparison, our approach is based on the solutions of a set of complex-valued equations, and those equations give all possible solutions.

The equations provided in section 4.2 are used in section 4.2.2 in order to review how to design many types of metasurface-based holograms. It includes phase-only holograms for a given polarization state, two different phase-only holograms for two orthogonal polarization states and phase and amplitude holograms.

The main constraint of using a single ideal metasurface is that the Jones matrix describing the system has to be unitary and symmetric. Hence, in order to design any system described by a unitary Jones matrix, equations that give all the possible pair of ideal metasurfaces are provided in section 4.3 and proved in section 4.7.3. In [132], it is proved that any system composed of any number of waveplates and rotators is optically equivalent to a system composed of one waveplate and one rotator. From our results, we demonstrate that such system is also equivalent to a system composed of two waveplates.

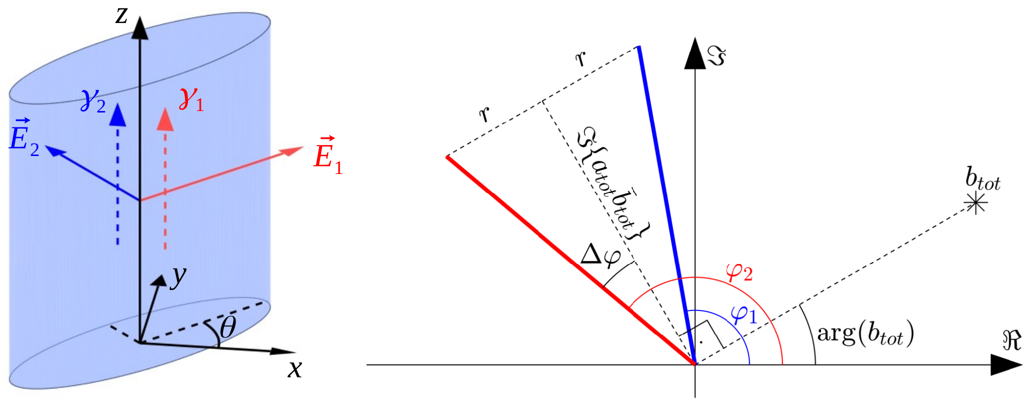
In section 4.3.1, those equations are applied for the design of a polarization rotator, which is actually a degenerate case, and of an optical element called in this work a pseudo-quarter-wave plate. A pseudo-quarter-wave plate can also transform a linear polarized beam into a circular polarized beam but it can also transform a x-polarized beam into a y-polarized beam. It is shown that a pseudo-quarter-wave plate can be composed of a quarter-wave plate and a half-wave plate, but there is also a non-obvious solution that minimizes the thickness of both ideal metasurfaces.

The second set of techniques is for the design of anti-reflective metasurfaces. Anti-reflective metasurfaces are already been demonstrated in the literature [125–127], but, in section 4.4, a systematic design technique is given using the Fourier Modal Method presented in Chapter 2. In section 4.4.1, the possible designs of an anti-reflective metasurface are provided for glass cylinders on a glass substrate and silicon cylinders on a silicon substrate. The main advantage of such anti-reflecting metasurface is that it requires a single etch of the substrate. High-power applications are a typical application for anti-reflective metasurface, where conventional anti-reflective coatings may burn due to absorption. A design of a glass metasurface on a glass substrate is given and analyzed in depth in section 4.4.1.

An important element in polarization optics, topic covered in sections 4.2 and 4.3, is half-wave plate and the design of a metasurface acting as a half-wave plate is provided in section 4.5, along with the design process. This section gives the limitations, advantages and challenges of such metasurface.

4.2 Design of ideal single-mode metasurfaces

Ideal metasurfaces are defined as zeroth-order gratings composed of cylinders with usually circular, elliptical or rectangular cross-section which have two properties. First, the metasurface do not absorb and reflect light and, second, the eigenpolarizations, the polarizations that are transmitted unchanged, are linear and excite a single eigen-mode of the metasurface each. Hence, ideal metasurfaces can be considered as ideal birefringent media or media with only a linear phase anisotropy [133], meaning that ideal wave plates can be treated the same way as ideal metasurfaces since they satisfy the same properties. A unit cell of an ideal metasurface is shown in fig. 4.1a, where the two eigen-modes of the metasurface are represented by the field $\vec{E}_1$ and $\vec{E}_2$ and by the propagation constant γ_1 and γ_2 respectively. For cylinders with cylindrical or rectangular cross-section, it is assumed that the eigen-modes of the metasurface share the same properties in terms of the condition of excitation as the waveguide modes propagating into a single cylinder, meaning that the angle of the linear polarization is parallel to the main axis of the ellipse or rectangle. The differences between an ideal and a real metasurface is discussed in section 4.2.1 and the main deviation of the response of a real metasurface to an ideal one comes from the reflection at the interfaces, which is usually below 10%.



(a) (b)

Figure 4.1 – a) Schema of a cylinder with an elliptical cross-section rotated by an angle θ from the x-axis. This cylinder is the building block of a single-mode metasurface, where the propagation constant of the eigen-modes is γ_1 and γ_2. b) Schema that illustrates the expression of φ_1, φ_2 and $\Delta\varphi$ with respect to a_{tot} and b_{tot} in the complex plane.

If the plane waves are orthonormalized, the response of the ideal metasurface described by the Jones matrix $\hat{T}_{tot}$ is given by

$$\hat{T}_{tot} = \begin{pmatrix} a_{tot} & b_{tot} \\ b_{tot} & d_{tot} \end{pmatrix} = \hat{R}_{\theta} \begin{pmatrix} e^{j\varphi_1} & 0 \\ 0 & e^{j\varphi_2} \end{pmatrix} \hat{R}_{-\theta}, \tag{4.1}$$

where the matrix $\hat{R}$ is the rotation matrix and φ_1 and φ_2 is the phase accumulation of the eigen-modes propagating inside the metasurface. The relationship between the phase accumulation φ and the propagation constant γ of the eigen-mode is $\varphi = \gamma h$, where h is the thickness of the metasurface which is equivalent to the height of the cylinders. For linear phase anisotropy, the Jones matrix $\hat{T}_{tot}$ is symmetric [133], which is the reason why $\hat{T}_{tot}$ is only expressed in terms of a_{tot}, b_{tot} and d_{tot}.

In general, if the plane waves before and after the metasurface are orthonormalized, the Jones matrix $\hat{T}_{tot}$ describing the transmission of an ideal metasurface or multiple ideal metasurfaces separated by lossless materials has to be unitary since no reflection and absorption are present in the system. The property of such system is that, if a polarization state described by the Jones vector $\vec{p}_1$ is transformed after going through the system into the polarization state $\vec{q}_1$, a polarization state $\vec{p}_2$ orthogonal to the polarization state $\vec{p}_1$, meaning that $\vec{p}_1^H \vec{p}_2 = 0$, is transformed into a polarization state $\vec{q}_2$ orthogonal to the polarization state $\vec{q}_1$. In other words, if the transformation $\vec{p}_1$ to $\vec{q}_1$ is known, the transformation $\vec{p}_2$ to $\vec{q}_2$ is known up to a phase factor. For an ideal metasurface, in most cases, this phase factor is imposed by the choice of the transformation from $\vec{p}_1$ to $\vec{q}_1$ due to $\hat{T}_{tot}$ being symmetric. For a given $\vec{p}_1 = (p_{x,1}, p_{y,1})$ and $\vec{q}_1 = (q_{x,1}, q_{y,1})$, the elements of the Jones matrix $\hat{T}_{tot}$ are

$$\begin{aligned} a_{tot} &= \frac{q_{x,1}\bar{q}_{y,1} - \bar{p}_{x,1}p_{y,1}}{s} \\ b_{tot} &= \frac{|p_{x,1}|^2 - |q_{x,1}|^2}{s} \\ d_{tot} &= \frac{p_{x,1}\bar{p}_{y,1} - \bar{q}_{x,1}q_{y,1}}{s} \\ s &= p_{x,1}\bar{q}_{y,1} - p_{y,1}\bar{q}_{x,1} \end{aligned} \tag{4.2}$$

When s is equal to zero, multiple solutions exist and are given by

$$\begin{aligned} a_{tot} &= \frac{q_{x,1}}{p_{x,1}}(1 - |p_{y,1}|^2(1 + e^{i\phi})) \\ b_{tot} &= \bar{p}_{x,1}q_{y,1}(1 + e^{i\phi}) \\ d_{tot} &= \frac{q_{y,1}}{p_{y,1}}(1 - |p_{x,1}|^2(1 + e^{i\phi})), \end{aligned} \tag{4.3}$$

where ϕ can be any real number. The coefficient s can be equal to zero only if $|p_{x,1}|$ and $|p_{y,1}|$ are equal to $|q_{x,1}|$ and $|q_{y,1}|$ respectively. Hence, if $p_{x,1}$ or $p_{y,1}$ is equal to zero, the expressions (4.3) are still valid. For example, if $p_{x,1}$ is zero, $q_{x,1}$ is also zero, $|p_{y,1}|^2$ is one and a_{tot} becomes

$a_{tot} = e^{j\phi}$.

The case $s = 0$ is an interesting case because the phase of the two orthogonal polarization states $\vec{q}_1$ and $\vec{q}_2$ can be controlled independently. This case has already been studied in the literature [36]. Compared to previous work, the formula provided here can be directly applied. If s is equal to zero, using the expressions (4.3), the matrix $\hat{T}_{tot}$ can be written as

$$\hat{T}_{tot} = e^{i(\beta-\alpha)} \begin{pmatrix} e^{i\alpha}[1-\sin^2(\sigma)(1+e^{i\phi})] & \sin(\sigma)\cos(\sigma)(1+e^{i\phi}) \\ \sin(\sigma)\cos(\sigma)(1+e^{i\phi}) & e^{-i\alpha}[1-\cos^2(\sigma)(1+e^{j\phi})] \end{pmatrix}, \tag{4.4}$$

for

$$\vec{p}_1 = \begin{pmatrix} \cos(\sigma) \\ e^{i\alpha}\sin(\sigma) \end{pmatrix} \qquad \vec{q}_1 = e^{i\beta} \begin{pmatrix} \cos(\sigma) \\ e^{-i\alpha}\sin(\sigma) \end{pmatrix}. \tag{4.5}$$

All the polarization state pairs $\vec{p}_1$ and $\vec{q}_1$ that satisfy the condition $s = 0$ with $p_{x,1}$ real, can be expressed in the form shown in equation (4.5). For a polarization state $\vec{p}_2$ orthogonal to $\vec{p}_1$, the output polarization state $\vec{q}_2$ is

$$\vec{p}_2 = \begin{pmatrix} \sin(\sigma) \\ -e^{i\alpha}\cos(\sigma) \end{pmatrix} \quad \Rightarrow \quad \vec{q}_2 = e^{i(\beta+\phi)} \begin{pmatrix} -\sin(\sigma) \\ e^{-i\alpha}\cos(\sigma) \end{pmatrix}. \tag{4.6}$$

The polarization states that are generally used in the literature and fulfill the condition $s = 0$ are linear polarized lights, whose polarization angle does not change after going through the metasurface [40–42], and circular polarized lights whose handedness changes while going through the metasurface [36, 40].

Once the desired matrix $\hat{T}_{tot}$ is found, the orientation of the cylinder, given by the angle θ, and the phase accumulations for both eigen-modes, given by φ_1 and φ_2, are needed. Those parameters can be obtained by performing a diagonalization of the matrix $\hat{T}_{tot}$ since $\hat{T}_{tot}$ can be described in the form shown in equation (4.1). However, it exists a direct expression of the parameters θ, φ_1 and φ_2, which are

$$\begin{aligned}
\theta &= \frac{1}{2}\,\mathrm{atan2}(s_1|b_{tot}|^2, |\mathrm{Re}\{a_{tot}\bar{b}_{tot}\}|) & & \\
\varphi_1 &= \frac{\pi}{2} + \arg(b_{tot}) - \Delta\varphi_1 & \varphi_2 &= \frac{\pi}{2} + \arg(b_{tot}) + \Delta\varphi_2 \\
\Delta\varphi_1 &= \mathrm{atan2}(s_1 r_1, \mathrm{Im}\{a_{tot}\bar{b}_{tot}\}) & \Delta\varphi_2 &= \mathrm{atan2}(s_1 r_2, \mathrm{Im}\{d_{tot}\bar{b}_{tot}\}) \\
r_1 &= \sqrt{\mathrm{Re}\{a_{tot}\bar{b}_{tot}\}^2 + |b_{tot}|^4} & r_2 &= \sqrt{\mathrm{Re}\{d_{tot}\bar{b}_{tot}\}^2 + |b_{tot}|^4} \\
s_1 &= \begin{cases} 1 & \text{if } \mathrm{Re}\{a_{tot}\bar{b}_{tot}\} \geq 0 \\ -1 & \text{otherwise} \end{cases} & &
\end{aligned} \tag{4.7}$$

In theory, $\Delta\varphi_1$ is equal to $\Delta\varphi_2$ and r_1 is equal to r_2. Hence, $\Delta\varphi_1$ and $\Delta\varphi_2$ are referred as $\Delta\varphi$ and r_1 and r_2 are referred as r. However, when b_{tot} is nearly zero, numerical instability may occur and the two separate definitions for $\Delta\varphi$ and r guarantee the convergence as b_{tot} goes to zero, where the parameters θ, φ_1 and φ_2 are chosen to be

$$\begin{aligned}
\theta &= 0 \\
\varphi_1 &= \arg(a_{tot}) \\
\varphi_2 &= \arg(d_{tot})
\end{aligned} \tag{4.8}$$

When changing continuously the elements in the matrix $\hat{T}_{tot}$, two types of discontinuities in the parameters θ, ϕ_1 and ϕ_2 can occur. The first type of discontinuity is the wrapping of the phase accumulations ϕ_1 and ϕ_2. The second type of discontinuity is when θ goes from $\pi/4$ to $-\pi/4$ or vice versa. When removing the discontinuity on θ by adding or subtracting $\pi/2$ to θ, the value of ϕ_1 and ϕ_2 must be swapped.

For metasurfaces, the coefficient $\Delta\varphi$ determines how difficult the fabrication of the metasurface is since the height h of the structure is given by

$$2\Delta\varphi = \varphi_2 - \varphi_1 = (\gamma_2 - \gamma_1)h, \tag{4.9}$$

where γ_1 and γ_2 are the propagation constants of the eigen-modes in the metasurface. Adding the same phase to φ_1 and φ_2 adds a constant phase to the transmitted polarized state, which is usually not taken into account for the design of ideal metasurfaces. For phase-only holograms, the situation is different and, in general, in order to vary the phase after the metasurface from 0 to 2π, the height is given by

$$h(\gamma_{max} - \gamma_{min}) = 2\pi \tag{4.10}$$

where γ_{max} and γ_{min} are the smallest and largest propagation constant that can be obtained. In the case $s = 0$, the right hand-side of equation (4.10) can be reduced down to π, which is the minimum reached by the metasurfaces based on the Pancharatnam-Berry phase if the phase of the transmitted light is designed for a single incident polarization. This is discussed in section 4.2.2. The proofs for the equations in this section are given in sections 4.7.1 and 4.7.2.

4.2.1 Comparison between ideal and real single-mode metasurfaces

In the section above, a set of equations is proposed, which allows to compute the orientation of a structure and the phase accumulation of the two eigen-modes present in the metasurface directly from a desired transformation of polarization states. However, those equations are valid only for ideal metasurfaces. In this section, two different metasurfaces are analyzed for a wavelength of 1550 nm, which is commonly used in telecommunication, in order to give an estimation on how a real metasurface differs from an ideal metasurface. By scaling the dimensions accordingly, the results are in the same range for other wavelengths.

Two sets of quantities are analyzed for the estimation of the difference between a real metasurface and an ideal metasurface. First, the reflections at the two interfaces of both metasurfaces are given. It should be zero for an ideal metasurface. Second, for each eigen-mode of the metasurface, the polarization of the incident plane wave that excites only this eigen-mode is computed along with the polarization of the transmitted plane wave. For an ideal metasurface, the incident and transmitted plane waves have the same linear polarization and the difference of the polarization angle of the incident plane waves that excite a single eigen-mode is 90°. From this analysis, the angle of polarization of the incident plane waves that excite a single eigen-mode is compared with the angle of rotation of the cylinders along with the effect of this geometrical rotation on the propagation constant of the eigen-modes and the reflection efficiency at the interfaces. This is important in the design point of view since it gives an estimation of the error due to the assumption that the angle θ given in section 4.2 is the same as the angle of rotation of the structure.

Both metasurfaces are composed of silicon cylinders on a glass substrate, but their cross-sections, shown in figs. 4.2a and 4.2b, are different. Those two metasurfaces are chosen for the following reasons. The structures are made of silicon because it is a common material with one of the highest refractive index for a dielectric material in the near-infrared region. Therefore, the use of silicon offers a large range of propagation constants due to its high refractive index, and, at the same time, it leads to higher reflection at the interfaces than a material with a lower refractive index. Hence, it behaves less as an ideal metasurface than a metasurface made of a material with a lower refractive index. For the substrate, it is a common practice to use a low refractive index material instead of a high refractive index material because the lattice constant can be larger before the first order propagates and it also increases the transmission efficiency of the metasurface. The dimension of the unit cell is set to 650 nm because, if it is larger, the metasurface may become a multi-mode metasurface, which behaves differently than a single-mode metasurface as shown in chapter 5.

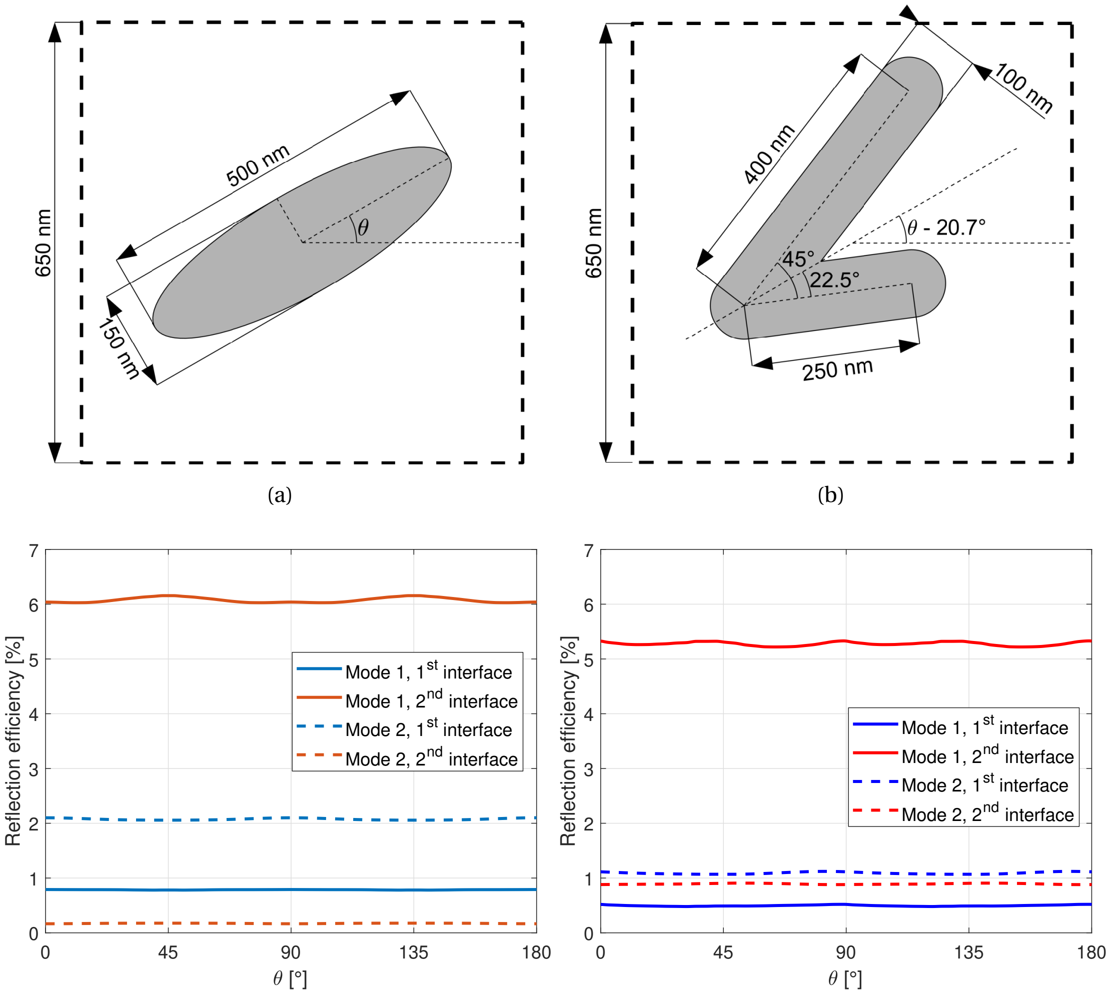
The cross-section of the first metasurface is an ellipse, where the length of the axis is 500 nm and 150 nm. This cross-section is shown in fig. 4.2a and it is one of the simplest structures that produces a difference in the propagation constant of the two eigen-modes in a metasurface. The cross-section of the second metasurface, shown in fig. 4.2b, is an asymmetric v-shape, where the arms length are 500 nm and 350 nm. The angle between those arms is 45° and they are 100 nm wide. This cross-section has been chosen to see how it can deviate from the second property of an ideal metasurface, which is the existence of linear eigenpolarizations that excite a single eigen-mode in the metasurface each. If it is possible to deviate sufficiently from this property, a binary metasurface can be used for applications that require an asymmetric Jones matrix. As shown in this section, this deviation is negligible.

An ideal metasurface has no reflection at the two interfaces of the metasurface. For the two considered metasurfaces, the reflection efficiencies at the two interfaces of the metasurface for each eigen-mode and for different angles of rotation of the cylinders are given in figs. 4.2c and 4.2d, where eigen-mode 1 is the eigen-mode with the highest propagation constant. The propagation constant of both eigen-modes is shown in fig. 4.3b. In order to compute the reflection efficiencies, an incident eigen-mode or plane wave is required. For the second interface, the incident eigen-mode used to compute the reflection efficiency related to the eigen-mode m is obviously the eigen-mode m. For the first interface, the incident plane wave used to compute the reflection efficiency related to the eigen-mode m is the plane wave that excites only the eigen-mode m. Those reflection efficiencies can be directly obtained from the coupling matrices at the two interfaces when the eigen-modes of the metasurface are orthonormalized. The reflection efficiencies shown in figs. 4.2c and 4.2d have the same magnitude. In order to estimate the effect of those reflections on the overall transmitted efficiency, the metasurface is described as two independent Fabry-Pérot cavities, one per polarization. This description is further discussed in section 4.4. In a Fabry-Pérot cavity, the minimum and maximum transmission efficiencies T_{min} and T_{max} are given by

$$T_{min} = \frac{(1-R_1)(1-R_2)}{(1+\sqrt{R_1R_2})^2} \qquad T_{max} = \frac{(1-R_1)(1-R_2)}{(1-\sqrt{R_1R_2})^2}, \tag{4.11}$$

where R_1 and R_2 are the reflection efficiencies at the first and second interface respectively. Therefore, in the case of the ellipse, the transmission efficiencies are between 89.2% and 97.4% if only the eigen-mode 1 is excited and 96.6% and 98.9% if only the eigen-mode 2 is excited. In the case of the v-shape, the transmission efficiencies are similar: between 91.3% and 97.4% for eigen-mode 1 and between 96.1% and nearly 100% for eigen-mode 2.

By comparing figs. 4.2c and 4.2d and fig. 4.3b, a relationship between the propagation constant of the eigen-modes and the reflection efficiencies can be seen. The propagation constant of the eigen-mode 1 for both metasurfaces is above the propagation constant of a plane wave propagating in glass, which is 5.88 1/µm. Therefore, the reflection efficiency at the second interface is higher than the reflection efficiency at the first interface. Moreover, the reflection efficiencies are higher in the case of the ellipse than in the case of the v-shape since the propagation constant in the case of the ellipse is higher. For the eigen-mode 2, its

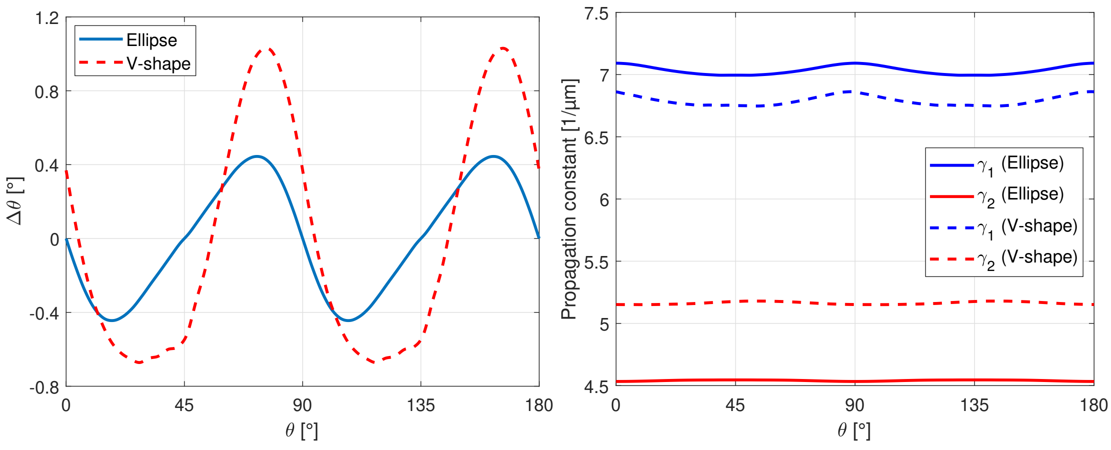


(a) (b)

(c) (d)

Figure 4.2 – a) Cross-section of the first metasurface. It is an ellipse made of silicon and the surrounding is air. The length of the axis is 500 nm and 150 nm. θ is the angle of rotation of the ellipse with respect to the x-axis. b) Cross-section of the second metasurface. It is an asymmetric v-shape made of silicon and the surrounding is air. The length of the arms is 500 nm and 350 nm and they are separated by an angle of 45°. The arms thickness is 100 nm. The angle between the x-axis and the small arm is $\theta + 43.2°$. c) Reflection efficiencies for different angles of rotation at the interfaces of a metasurface composed of cylinders with elliptical cross-section for both eigen-modes. The first interface is the interface between glass and the metasurface. The second interface is the interface between the metasurface and air. d) Reflection efficiencies for different angles of rotation at the interfaces of a metasurface composed of cylinders with v-shaped cross-section for both modes. The scales of the axis is the same as in fig. 4.2c.

propagation constant is between the propagation constant of a plane wave in air, which is 4.05 1/µm, and in glass. Hence, it is more difficult to predict which interface reflects the most, especially that, for the case of the v-shape, the reflection efficiency at the first interface is still higher even if the propagation constant of the eigen-mode is closer to the one for the glass. Since the propagation constant of the eigen-mode 2 in the case of the ellipse is lower than in the case of the v-shape, the reflection efficiency is lower at the first interface and higher at the second interface. This relationship can be understood with the concept of effective permittivity, but such concept may not be accurate since a metamaterial or a metasurface can be described accurately by effective parameters only if the lattice constant is much smaller than the wavelength, which is usually not the case for metasurfaces in the near-infrared regime. If a unit cell is designed in order to have an eigen-mode with a higher propagation constant than those presented here, the transmission efficiency related to this eigen-mode is expected to decrease.



(a) (b)

Figure 4.3 – a) Difference between the polarization angle of the plane wave that excites only eigen-mode 1 and the angle of rotation of the cylinders θ with respect to θ b) Propagation constant of the eigen-modes with respect to the angle of rotation of the cylinders θ.

The second property of an ideal metasurface is that the eigenpolarizations are linear and excite a single eigen-mode each. In order to check if a metasurface has this property, the incident plane waves that excite a single eigen-mode of the metasurface have to be orthogonal and linearly polarized. Moreover, the transmitted plane waves excited by the eigen-modes of the metasurface also have to be orthogonal and linearly polarized with the same polarization angle as the incident plane wave mentioned earlier. The polarization states are obtained from the coupling matrices at both interfaces, which is described in terms of the weight of the TM and the TE plane waves. Using the expressions (3.23) and (3.24) of the TM and TE plane waves, the complex amplitudes of the tangential electric field E_x and E_y are obtained. Then, the polarization state is given by

$$
\begin{aligned}
&\theta_p = \operatorname{atan2}(2\,\mathrm{Re}\{\bar{E}_x E_y\}, |E_x|^2 - |E_y|^2) && \\
&\tan(\chi) = \frac{b}{a} && \\
&a = \sqrt{\frac{r+\Delta}{2}} && b = \sqrt{\frac{r-\Delta}{2}} \\
&r = |E_x|^2 + |E_y|^2 && \Delta = |E_x^2 + E_y^2|,
\end{aligned}
\tag{4.12}
$$

where θ_p is the tilt of the polarization ellipse, which is the angle between the x-axis and the major axis, χ is the ellipticity angle and a and b are the length of the major and minor half-axis respectively. Usually, the expression of $\tan(2\theta_p)$ is given instead of θ_p, which creates practical difficulties since the computed θ_p can be the angle between the x-axis and the minor axis. The expression of θ_p provided in (4.12) solves this issue and it is based on the expression of the Stokes parameters (section 6.1 of [92]).

For the considered metasurfaces, the ellipticity angle is maximum 0.17° for the four different polarization states in the case of the ellipse and maximum 0.45° in the case of the v-shape. Hence, the polarization states can be considered as linearly polarized. Then, the difference between the polarization angle θ_p and the angle of rotation of the cylinder θ, called $\Delta\theta$ in fig. 4.3a, is computed. The reference for the rotation angle θ is chosen such that, when θ is zero, the plane wave that excites only the eigen-mode 1, the eigen-mode with the highest propagation constant, is approximately x-polarized. In fig. 4.3a, the difference $\Delta\theta$ between the polarization angle of the incident plane wave that excites only the eigen-mode 1, and the rotation angle is plotted. For the plane waves that excite a single eigen-mode of the metasurface, the difference of polarization angle between the incident polarization state and the transmitted polarization state is maximum 0.008° in the case of the ellipse and maximum 0.051° in the case of the v-shape. Such low values can be explained by the small difference of the refractive index between the substrate (glass) and the superstrate (air), since this value has to be zero for a symmetric system, meaning that the superstrate would be glass instead of air. For an ideal metasurface, the two plane waves that excite a single eigen-mode of the metasurface are orthogonal, meaning that the difference of polarization angle is 90°. In the considered metasurfaces, the deviation from this difference is maximum 0.01° in the case of the ellipse and maximum 0.19° in the case of the v-shape. Even if those values are small, the deviation from the orthogonal polarization state is larger by nearly an order of magnitude compared to the difference between the polarization angle of the incident and transmitted plane waves. However, those are negligible effects and the Jones matrix that describes a metasurface can be considered as symmetric.

The last point of this section is to discuss if it is reasonable to simulate a single metasurface and, then, to deduce the eigenpolarizations of the system, the propagation constants and the reflections at the interfaces of the metasurface with rotated cylinders. From figs. 4.2c, 4.2d, 4.3a, and 4.3b, it can be done if an error of a few percents is allowed. This error may increases if the dimensions of the cylinders are larger since the eigen-modes propagating inside the cylinders feel more the presence of their neighbors. Another conclusion from those plots is that there is no significant advantage to use complex cross-sections in the design of single-mode metasurfaces.

4.2.2 Design of metasurface-based hologram

A metasurface-based hologram is created by an array of cylinders with different cross-sections, where each cross-section is designed such that it gives a specific phase accumulation and change in polarization states. Such well-known dielectric metasurfaces that generate an hologram are based on cylinders with circular cross-section with varying diameters such as [26–30] or based on the Pancharatnam–Berry phase as shown in [31–35, 38]. It is also possible to design a metasurface that generates two different holograms for two orthogonal polarization states [36, 40–42], but it is more difficult to fabricate such metasurface due to higher aspect ratio. In this section, the important elements to consider when designing such metasurfaces are provided and the design of four different groups of hologram-generating metasurfaces are discussed. Those groups are:

- polarization-independent metasurfaces
- metasurfaces designed for a single polarization state for a phase-only hologram
- metasurfaces that generate two phase-only holograms for two orthogonal polarization states
- metasurfaces designed for a single polarization state for a phase and amplitude hologram

First of all, in the design of such optical devices, it is assumed that the transmission function at the location of the cylinder is the same as if the metasurface is periodic, meaning that the Jones matrix at the location of each cylinder is given by the Jones matrix of the periodic metasurface composed of this cylinder. Then, all the required Jones matrices are computed using, for the non-obvious case, equations (4.2) and (4.3), along with the associated ideal metasurface, which is described by the two phase delays φ_1 and φ_2 and the rotation angle θ and are given by equations (4.7) and (4.8).

From those sets of ideal metasurfaces, there are a few critical metasurfaces, which are those with the smallest phase accumulation, the largest phase accumulation and the largest phase accumulation difference $\Delta\varphi$. The metasurfaces that behave as those critical ideal metasurfaces are composed with cylinders with the largest aspect ratio, making them difficult to fabricate. For a metasurface based on cylinders with circular cross-section, the metasurface with the smallest phase accumulation is the cylinder with the smallest diameter, and the metasurface with the largest phase accumulation is the cylinder with the largest diameter, which means the metasurface which has the smallest gap with its neighbors and which is also the most likely to be multi-mode.

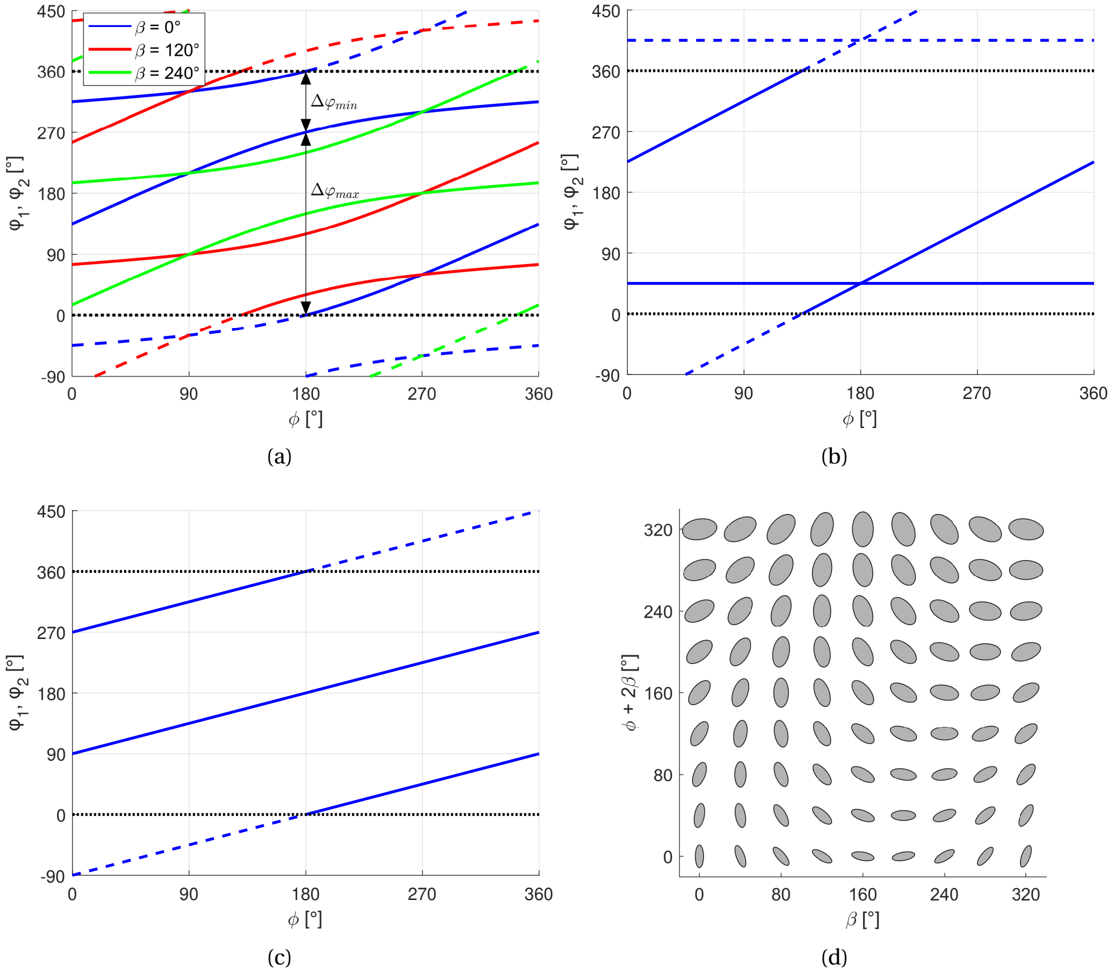
A large phase accumulation difference is also an issue. As an example, the cylinders with the elliptical cross-section from section 4.2.1, which is shown in fig. 4.2a, is considered. The propagation constants of the two eigen-modes are 4.54 1/µm and 7.03 1/µm. For cylinders with a diameter of 500 nm, which is the length of the long axis of the ellipse shown in fig. 4.2a, the propagation constant is 11.3 1/µm, but it is a multi-mode metasurface, meaning that the diameter and the lattice constant are too large for a single-mode metasurface. For the same lattice constant, the diameter has to be reduced around 360 nm in order to have a single-mode metasurface and the propagation constant is then 8.30 1/µm. For cylinders with a diameter of 150 nm, which is the length of the small axis of the ellipse shown in fig. 4.2a, the propagation constant is 4.24 1/µm. Therefore, the difference of the propagation constant between the modes in the cylinder with those two different diameters is larger than the difference of the propagation constant of the modes in the cylinder with elliptical cross-section. Those differences impact the height of the cylinder since, if the difference between the propagation constants is smaller, the cylinder has to be taller for the same desired difference in phase accumulation.

Polarization-independent metasurfaces

For the design of polarization-independent metasurfaces for phase-only holograms, any unitary symmetric Jones matrix can be chosen at first and the phase accumulation is created by adding this phase delay to φ_1 and φ_2. Of course, in order to minimize the fabrication difficulties, the difference between φ_1 and φ_2 has to be zero, which means that the metasurface composed of cylinders with circular cross-section with varying diameter is the best candidate. However, if the desired output is given by its angular spectrum and this angular spectrum is central symmetric, there is a better candidate as discussed below.

Phase-only hologram for a single polarization state

For metasurfaces designed for a single polarization state for phase-only holograms, if the input and output polarization states are fixed and the parameter s, given in equation (4.2), is not zero, then, equation (4.2) gives the only possible solution, which is also a polarization independent metasurface. However, if the parameter s is zero, the Jones matrix given in (4.4) can be used, where the parameter ϕ can be chosen arbitrarily since it does not affect the output polarization state $\vec{q}_1$. If the input polarization state $\vec{p}_1$ is elliptically polarized, the typical relationship between the parameter ϕ and the phase accumulations φ_1 and φ_2 is shown in fig. 4.4a, where the input polarization state $\vec{p}_1$ is given by equations (4.5) with $\sigma = \alpha = 45°$. The color of the plot is the same for φ_1 and φ_2 because they can be interchanged by adding 90° to the rotation of the cylinder. From the matrix in (4.4), as the phase of the output state $\vec{q}_1$, called β, increases, the phase β is added to the function $\varphi_1(\phi)$ and $\varphi_2(\phi)$ as shown in fig. 4.4a. Hence,



(a)

(b)

(c)

(d)

Figure 4.4 – a) The phase accumulations φ_1 and φ_2 with respect to the parameter ϕ for three different values of β when the input and output polarization states are given by the expressions in (4.5) with $\sigma = \alpha = 45°$, which correspond to elliptical polarized light. $\Delta\varphi_{min}$ and $\Delta\varphi_{max}$ are respectively the minimum and maximum difference of the phase accumulations of the two propagating modes. b) The phase accumulations φ_1 and φ_2 with respect to the parameter ϕ for $\beta = 45°$ when the input and output polarization states are given by the expressions in (4.5) with $\sigma = 45°$ and $\alpha = 0°$, which correspond to linear polarized light. c) The phase accumulations φ_1 and φ_2 with respect to the parameter ϕ for $\beta = 0°$ when the input and output polarization states are given by the expressions in (4.5) with $\sigma = 45°$ and $\alpha = 90°$, which correspond to circular polarized light. d) Cross-section of the cylinders that transforms a left-hand circular polarized beam into a right-hand circular polarized beam with a phase delay of β and a right-hand circular polarized beam into a left-hand circular polarized beam with a phase delay of $\beta + \phi$, assuming that the phase accumulation is an affine function of the length of the ellipse main axis, the phase accumulation is 0° for the smallest ellipse main axis and the phase accumulation is 360° for the largest ellipse main axis. The y-axis is chosen such that the cross-sections on a horizontal line are used in a metasurface based on the Pancharatnam-Berry phase.

for every phase β, the parameter is chosen such that φ_1 and φ_2 are as close as possible of the minimum phase accumulation that can be obtained and $\Delta\varphi$ is also minimized. In fig. 4.4a, the minimum phase accumulation is set to 0° but it can be any phase.

There are two extreme strategies in order to find the parameter ϕ for every β. The first strategy is to choose ϕ such that the phase accumulation difference $\Delta\varphi$ is minimum. In that case, ϕ is 180° (fig. 4.4a). This is obviously the best strategy when the incident light is linearly polarized, which gives a metasurface composed of cylinders with circular cross-section. This case is illustrated in fig. 4.4b where $\alpha = 0°$, which corresponds to linearly polarized light, and $\beta = 45°$. The second strategy is to chose ϕ such that one of the phase accumulation, φ_1 or φ_2, is kept to zero. In fig. 4.4a, the worst case for this strategy is for $\beta = 0°$, where the phase accumulation difference $\Delta\varphi$ is equal to $\Delta\varphi_{max}$.

This second strategy works best in the case shown in fig. 4.4c, where $\sigma = 45°$ and $\alpha = 90°$. In this case, the incident light is left-hand circular polarized, the transmitted light is right-hand circular polarized and ϕ is given by

$$\phi = -2\beta. \tag{4.13}$$

From this strategy, the phase accumulations φ_1 and φ_2 are independent of the phase β, meaning that the cylinder dimensions do not change. The phase β is produced by the rotation θ of the cylinder: β is proportional to 2θ. The cross-section of the cylinders for the different phase β is shown in the first horizontal line in fig. 4.4d. In the literature, it is known as the geometrical or the Pancharatnam–Berry phase [134, 135]. Metasurfaces based on the geometrical phase have another interesting property. If the transfer function for a left-hand circular polarized incident light is given by $t(x, y)$, the transfer function for a right-hand circular polarized incident light is $\bar{t}(x, y)$, due to equations (4.6) and (4.13). In many applications, the output is characterized by the intensity at the far-field. In other words, the function $I(k_x, k_y)$ used to characterized the output is the radiant intensity and is given by the Fourier transform of the transfer function $t(x, y)$:

$$I_1(k_x, k_y) = |\mathscr{F}\{t(x, y)\}(k_x, k_y)|^2. \tag{4.14}$$

The Fourier transform has the following property related to complex conjugation:

$$\mathscr{F}\{\bar{t}(x, y)\}(k_x, k_y) = \mathscr{F}\{t(x, y)\}^*(-k_x, -k_y). \tag{4.15}$$

Then, the radiant intensity $I_2(k_x, k_y)$ due to a metasurface described by the transfer function $\bar{t}(x, y)$ is related to $I_1(k_x, k_y)$ by

$$I_2(k_x, k_y) = I_1(-k_x, -k_y). \tag{4.16}$$

Hence, if the metasurface based on the geometrical phase produces a certain pattern in the far-field for left-hand polarized light, this metasurface produces a central symmetric version of this pattern for right-hand polarized light. If the pattern in the far-field is already central symmetric, the metasurface mimics a polarization-independent metasurface. Metasurfaces based on the geometrical phase can easily be designed to split left-hand polarized light from right-hand polarized light: by designing a metasurface that deflects a left-hand polarized beam to, for example, the left, this metasurface deflects a right-hand polarized beam to the right.

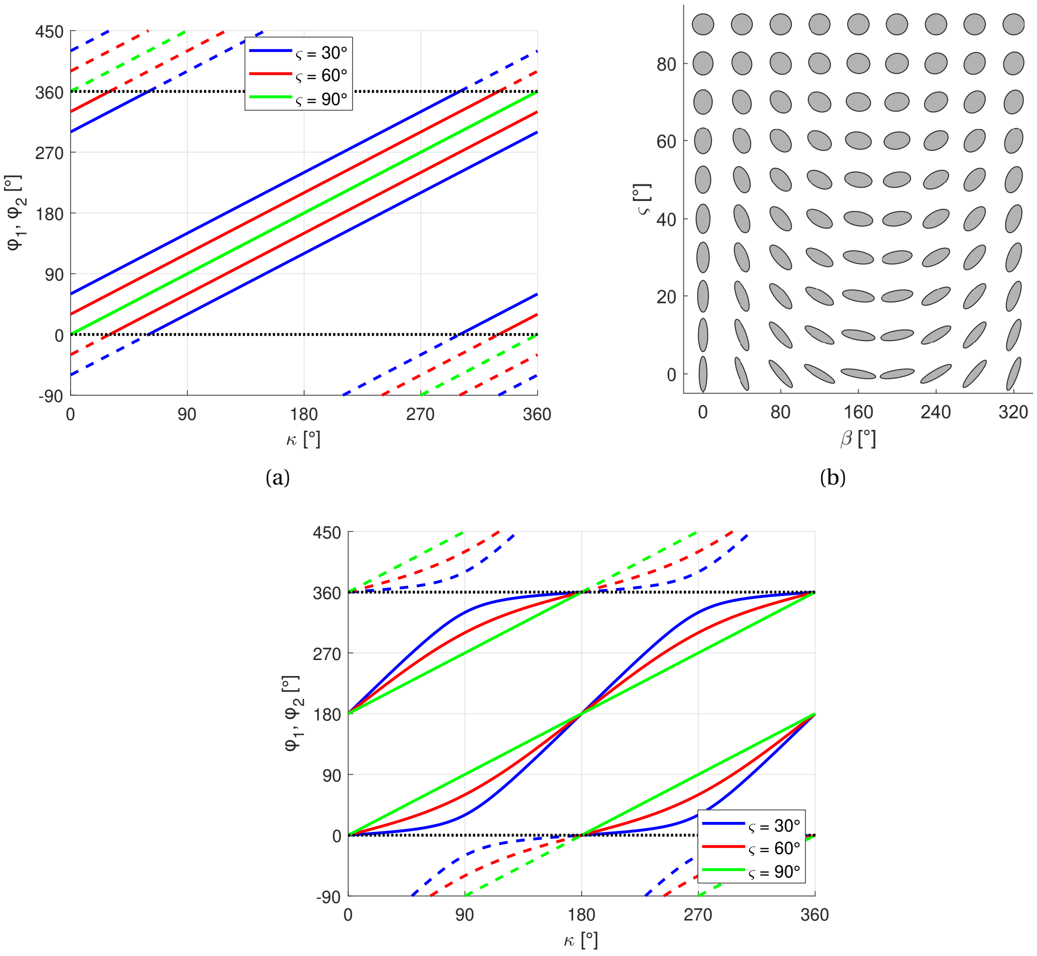
To conclude on metasurfaces designed for a single polarization state for a phase-only hologram, two different strategies have been discussed. The first strategy minimized $\Delta\varphi$ and it is optimal for linearly polarized light with metasurfaces composed of an array of cylinders with circular cross-section. The second strategy minimized the largest phase accumulation φ and it is optimal for circular polarized light with metasurfaces based on the geometrical phase. By comparing those two optimal solutions, the metasurface based on the geometrical phase is easier to fabricate since, deduced from the elements given at the beginning of this section, it is easier to produce a difference of phase accumulations $\Delta\varphi$ of 180° than varying both phase accumulations φ_1 and φ_2 from 0° to 360° even if $\Delta\varphi$ is zero.

Two phase-only holograms for two orthogonal polarization states

For metasurfaces that generate two holograms for two orthogonal polarization states, the only degree of freedom is the input polarization state since the parameter s given in equation (4.2) has to be zero, giving the output polarization state. Once an input polarization state is fixed, meaning that the parameters σ and α from equation (4.5) are known, the phase of the output polarization state $\vec{q}_1$ is given by β and the phase of the output polarization state $\vec{q}_2$ is given by $\beta - \phi$. Therefore, the Jones matrices describing the cylinders that compose the metasurface are obtained from (4.4). Metasurfaces that can generate two different holograms are particularly challenging to fabricate because the phase accumulation φ varies across the cylinders from 0° to 360° with the difference of phase accumulations $\Delta\varphi$ that can reach 180° or more. From figs. 4.4a to 4.4c, it can be deduced that the best choice of input polarization is the circular polarized one since the maximum $\Delta\varphi$ is 180° and the cross-section of the cylinders are given in fig. 4.4d. The other possibilities of input polarization state lead to a higher $\Delta\varphi$.

Phase and amplitude holograms for a single polarization state

In metasurfaces designed for a single polarization state for phase and amplitude holograms, one of the output polarization state is filtered out by a polarizer, giving the possibility of varying the amplitude of the transfer function of the whole system. Compared to metasurfaces



(a)

(b)

(c)

Figure 4.5 – a) The phase accumulations φ_1 and φ_2 with respect to the phase delay κ of the filtered polarization state for three different values of ς when the input beam is left-hand circular polarized and the kept output beam is right-hand circular polarized. b) Cross-section of the cylinders that transforms a left-hand circular polarized beam into a right-hand circular polarized beam with a phase delay of β and an amplitude of $\cos(\varsigma)$, assuming that phase accumulation is an affine function of the length of the ellipse main axis, the phase accumulation is 0° for the smallest ellipse main axis and the phase accumulation is 360° for the largest ellipse main axis. c) The phase accumulation φ_1 and φ_2 with respect to phase delay κ of the filtered polarization state for three different values of ς when the input beam the output beam kept is linear polarized with the same polarization angle.

discussed in this section, the parameter ϕ is not a degree freedom because it requires that the parameter s given in equation (4.2) remains zero, fixing the output polarization state. However, the phase of the output polarization state that is filtered out, called here κ, is now a degree of freedom which can be used in a similar way as the parameter ϕ illustrated by figs. 4.4a to 4.4c. The output polarization state $\vec{q}$ can be written as

$$\vec{q} = e^{i\beta}(cos(\varsigma)\vec{q}_1 + e^{i\kappa} sin(\varsigma)\vec{q}_2), \tag{4.17}$$

where $\vec{q}_1$ is the polarization state of the desired output, $\vec{q}_2$ is the polarization state that is filtered out and ς and β control respectively the amplitude and the phase after the metasurface for the transmitted light with polarization described by $\vec{q}_1$. The range of ς is from 0, where the incident light is fully transmitted to polarization state $\vec{q}_1$, to 90°, where the incident light is fully transmitted to polarization state $\vec{q}_2$.

If the incident light is left-hand circular polarized and the desired output is right-hand circular polarized, meaning that right-hand polarized light is filtered out using usually a circular polarizer, the phase accumulations φ_1 and φ_2 are given by fig. 4.4c when ς is zero, since the parameter s is zero in that case, and by fig. 4.5a otherwise. For $\varsigma = 90°$, φ_1 is equal to φ_2 meaning that the two lines representing φ_1 and φ_2 are superimposed. In both figures, the phase β is zero. For a different phase β, this phase is added to φ_1 and φ_2 and the plots shown in figs. 4.4c and 4.5a are shifted upward as shown in fig. 4.4a. In the general case, κ has to be determined for each value of β and ς. In the case where the incident light is left-hand circular polarized, the chosen strategy is to find κ such that

$$\frac{\varphi_2 + \varphi_1}{2} = \frac{\pi}{2} \tag{4.18}$$

Hence, κ is given by

$$\kappa = \frac{\pi}{2} - \beta \tag{4.19}$$

and the cross-section of the cylinders for different β and ς is given in fig. 4.5b.

If the incident light and the desired output are linearly polarized along the x-axis, the phase accumulations φ_1 and φ_2 in function of κ for different ς are given in fig. 4.5c. In that case, there is no simple strategy in order to minimize the thickness of the metasurface, but the obtained metasurface is more difficult to fabricate than in the case where the incident light is left-hand circular polarized.

4.3 Design of a pair of ideal single-mode metasurfaces

As shown in section 4.2, the Jones matrix describing an ideal metasurface is symmetric, which means a reduced set of optical functions can be performed. In order to perform any optical function described by a non-symmetric Jones matrix such as polarization rotation, two ideal metasurfaces are required. In this section, all the possible pairs of ideal metasurfaces are provided for a given Jones matrix. From the different solutions and the knowledge from section 4.2, it is possible to find the best starting point for a given functionality. In practice, an isotropic homogeneous material should separate the two metasurfaces in order to make the total response insensitive to misalignment when the metasurfaces are periodic, and less sensitive otherwise. The thickness of this separating material adds a constant phase to the response of the system. This phase is not included in this section because it is usually not needed, and it can be included in the Jones matrix of the ideal metasurfaces if necessary.

The two ideal metasurfaces are described by the Jones matrices $\hat{T}_1$ and $\hat{T}_2$, where the matrices $\hat{T}_1$ and $\hat{T}_2$ describe the first and second metasurface respectively. The symmetric matrices $\hat{T}_1$ and $\hat{T}_2$ are given by

$$\hat{T}_1 = \begin{pmatrix} a1 & b1 \\ b1 & d1 \end{pmatrix} \qquad \hat{T}_2 = \begin{pmatrix} a2 & b2 \\ b2 & d2 \end{pmatrix}. \tag{4.20}$$

The total response of the system is described by the matrix $\hat{T}_{tot}$, given by

$$\hat{T}_{tot} = \hat{T}_2 \hat{T}_1 = \begin{pmatrix} a_{tot} & b_{tot} \\ c_{tot} & d_{tot} \end{pmatrix}. \tag{4.21}$$

The coefficients a_{tot}, b_{tot}, c_{tot} and d_{tot} are given and, since $\hat{T}_{tot}$ is unitary, they fulfill equations (4.40). The unknowns are a_1, b_1, d_1, a_2, b_2 and d_2 and, if b_{tot} and c_{tot} are different than zero, a_2, b_2 and d_2 are given by

$$\begin{aligned} a_2 &= (z_s s + z_o r) e^{i\phi} \\ b_2 &= r e^{i\phi} \\ d_2 &= -(\bar{z}_s s + \bar{z}_o r) e^{i\phi}, \end{aligned} \tag{4.22}$$

where ϕ can be any real number and r is a real number in the interval $[0, r_{max}]$. a_1, b_1 and d_1 are expressed in terms of a_2, b_2 and d_2:

$$\begin{aligned} a_1 &= c_{tot} \bar{b}_2 + a_{tot} \bar{a}_2 \\ b_1 &= d_{tot} \bar{b}_2 + b_{tot} \bar{a}_2 \\ d_1 &= b_{tot} \bar{b}_2 + d_{tot} \bar{d}_2. \end{aligned} \tag{4.23}$$

Multiple expressions exist for z_s, z_o, s and r_{max} but each of those expressions diverges for some b_{tot} and c_{tot}. Therefore, the following expressions of z_s, z_0 and r_{max} are chosen as

$$\begin{aligned} z_s &= \begin{cases} i(b_{tot}+\bar{c}_{tot}) & \text{if } \mathrm{Re}\left\{\frac{b_{tot}}{\bar{c}_{tot}}\right\} \geq 0 \\ b_{tot}-\bar{c}_{tot} & \text{otherwise} \end{cases} \\ z_o &= \begin{cases} \frac{a_{tot}-d_{tot}}{b_{tot}+c_{tot}} & \text{if } \mathrm{Re}\left\{\frac{b_{tot}}{c_{tot}}\right\} \geq 0 \\ \frac{a_{tot}-d_{tot}}{c_{tot}-b_{tot}} & \text{otherwise} \end{cases} \\ r_{max} &= \begin{cases} \sqrt{\frac{|z_s|^2}{|z_s|^2+|z_o|^2\,\mathrm{Im}\{z_s\}^2}} & \text{if } \mathrm{Re}\left\{\frac{b_{tot}}{c_{tot}}\right\} \geq 0 \\ \sqrt{\frac{|z_s|^2}{|z_s|^2+|z_o|^2\,\mathrm{Re}\{z_s\}^2}} & \text{otherwise} \end{cases}. \end{aligned} \tag{4.24}$$

The parameter s is the root of a quadratic equation and is expressed as

$$\begin{aligned} s &= \frac{-c_b r \pm \sqrt{c_b^2 r^2 + |z_s|^2(1-(|z_o|^2+1)r^2)}}{|z_s|^2} \\ c_b &= \begin{cases} \mathrm{Re}\{z_o\}\,\mathrm{Re}\{z_s\} & \text{if } \mathrm{Re}\left\{\frac{b_{tot}}{c_{tot}}\right\} \geq 0 \\ \mathrm{Im}\{z_o\}\,\mathrm{Im}\{z_s\} & \text{otherwise} \end{cases}. \end{aligned} \tag{4.25}$$

Hence, for every r and ϕ, there are two distinct solutions except at $r = r_{max}$, where s is a double root, and at $r = 0$, which is a degenerate case. In order to optimize a system composed of two metasurfaces, it is usually only r which is changed because ϕ represents simply a constant phase created by the second metasurface, which is canceled by the first metasurface. For the case $r = 0$, taking the solution related to the second root instead of the first root is equivalent as adding π to ϕ. Hence, choosing a single value for s and varying ϕ from 0 to 2π are enough to get all the possible solutions.

For the case $b_{tot} = c_{tot} = 0$ and $a_{tot} = d_{tot}$, the solution is

$$\begin{aligned} a_2 &= \sqrt{1-r^2}\,e^{i\phi_1} \\ b_2 &= r e^{i(\phi_1+\phi_2)/2} \\ d_2 &= -\sqrt{1-r^2}\,e^{i\phi_2}, \end{aligned} \tag{4.26}$$

where ϕ_1 and ϕ_2 can be any real number and r is a real number in the interval $[-1,1]$. a_1, b_1 and d_1 are obtained from equations (4.23). For the case $b_{tot} = c_{tot} = 0$ and $a_{tot} \neq d_{tot}$, the solution is also given by equations (4.26) and (4.23) except that r is zero, meaning that b_1 and b_2 are also zero. The additional degree of freedom when $a_{tot} = d_{tot}$ is due to the invariance of the system to a rotation. Hence, the case $b_{tot} = c_{tot} = 0$ is a trivial case where the total phase shift is given by the sum of the phase shift from both ideal metasurfaces.

Two examples that illustrate the use of the formula presented in this section are given in section 4.3.1, where a polarization rotator and a pseudo-quarter-wave plate are designed. The Jones matrix describing such functions is not symmetric. Hence, a minimum of two ideal metasurfaces are required. The proofs of the different formulas presented in section are given in section 4.7.3.

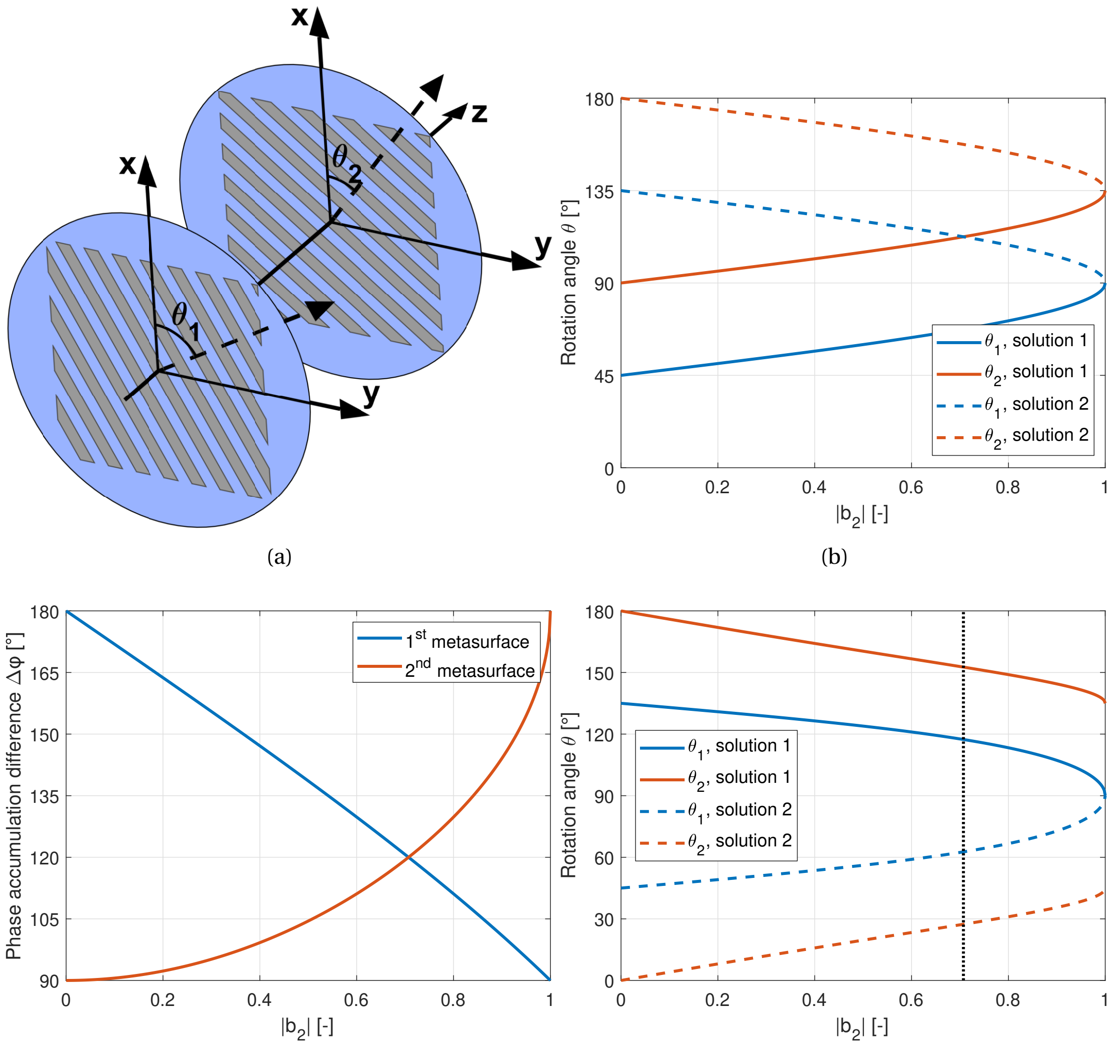
4.3.1 Design of a polarization rotator and a pseudo-quarter-wave plate

The equations presented in section 4.3 find all the pair of ideal metasurfaces that provide a system described by any desired unitary Jones matrix. Such system is shown in fig. 4.6a. If the Jones matrix is not symmetric, it is required to have at least two metasurfaces in order to perform this function. However, it can still be useful that two metasurfaces are used for a system described by a symmetric Jones matrix, either to make the metasurfaces easier to fabricate or to combine two metasurfaces that does not provide the desired function when used separately. For wave plates, which are analogous to metasurfaces, it has been done based on the Poincaré sphere concept [130].

In this section, the technique proposed in section 4.3 is applied for the design of two different elements. The first one is a polarization rotator that rotates the polarization of the incoming light by 90°. It can be considered as a degenerate case since its functionality is invariant to a rotation of the whole system around the propagation direction, meaning that the solution of equations (4.22) and (4.23) describes this rotation invariance instead of different pairs of metasurfaces. The second element is called in this work a pseudo-quarter-wave plate because it transforms a diagonally linear polarized light into a circular polarized light as a quarter-wave plate, but it transforms a x-polarized light into a y-polarized light and vice versa, which is not the case for a quarter-wave plate. The polarization rotator and the pseudo-quarter-wave plate are described by the Jones matrices $\hat{T}_{pr}$ and $\hat{T}_{qw}$ given by

$$\hat{T}_{pr} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad \hat{T}_{qw} = \begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix}. \tag{4.27}$$

For both elements, the design technique is the same. First, the elements of the Jones matrix describing each metasurface are obtained from equations (4.22) and (4.23) for every $|b_2|$ between zero and r_{max}, which is given in (4.24). It gives two sets of solutions since the parameter s, given in (4.25), is a solution of a quadratic equation. Once the coefficients of the Jones matrix are obtained, the difference of the phase accumulations inside the metasurface, called $\Delta\varphi$, is computed from its expression in (4.7). This is an important quantity because the thickness of the metasurface is proportional to $\Delta\varphi$ as shown in equation (4.9). On the contrary, the phase accumulation for each eigen-mode, φ_1 and φ_2, is irrelevant because a constant phase delay in the response of the metasurface is not important in the cases considered in this section. In order to realize a hologram with two metasurfaces, the value of φ_1 and φ_2 has to be considered. Finally, the orientation of each metasurface θ_1 and θ_2, shown in fig. 4.6a,



(a) (b)

(c) (d)

Figure 4.6 – a) Drawing of a system composed of two metasurfaces. The metasurfaces are drawn as silicon lines on a glass substrate, meaning that the propagation constant of the eigen-mode which is excited by an incident light polarized along the silicon lines is larger than the propagation constant of the other eigen-mode. Hence, the dashed arrows represent the fast-axis. b) Relationship between $|b_2|$ and the orientation of the fast axis for both metasurfaces, called θ_1 and θ_2, for the polarization rotator. Solution 1 and solution 2 refer to the two solutions obtained from the parameter s expressed in (4.25). c) Relationship between $|b_2|$ and the phase accumulation difference $\Delta\varphi$ for both metasurfaces for the pseudo-quarter-wave plate. d) Relationship between $|b_2|$ and the orientation of the fast axis for both metasurfaces for the pseudo-quarter-wave plate.

is computed from equation (4.7). θ_1 and θ_2 are chosen as the angle between the x-axis and the fast axis. An incident plane wave polarized along the fast axis excites only the eigen-mode with the lowest propagation constant.

For the polarization rotator, the value of $\Delta\varphi$ is always 180° for both metasurfaces. Hence, the two metasurfaces always have to act as two half-wave plates. This is also true for any polarization rotator. The rotation angle of the two metasurfaces θ_1 and θ_2 is shown in fig. 4.6b and the difference between θ_1 and θ_2 is always 45°, which is the rotation of polarization divided by two. However, θ_1 and θ_2 cover a range of only 90°, which should be 180° since the system is rotation invariant. The remaining 90° is obtained from the property of half-wave plates according to which rotating a half-plate plate by 90° is equivalent to adding a phase delay of 180°. Therefore, rotating both metasurfaces by 90° does not affect the response of the whole system.

For the pseudo-quarter-wave plate, the relationships between $|b_2|$, $\Delta\varphi$, θ_1 and θ_2 are shown in figs. 4.6c and 4.6d. Three configurations are interesting. The first one, which corresponds to $|b_1| = 0$, is a half-wave plate followed by a quarter-wave plate. The fast axis of the half-wave plate is oriented at ±45° from the y-axis and the fast axis of the quarter-wave plate is oriented along the x-axis. The second configuration, which corresponds to $|b_1| = 1$, is the inverse: a quarter-wave plate followed by a half-wave plate. In this case, the half-wave plate is also oriented at ±45° but the half-wave plate is oriented along the y-axis. The most interesting case when metasurfaces are involved is when the two curves representing $\Delta\varphi$ shown in fig. 4.6c cross each other. At this point, which corresponds to $|b_2| = 1/\sqrt{2}$, the maximum thickness of the metasurfaces is minimized. In that case, $\Delta\varphi$ is 120° and the fast axis of the first metasurface, given by the intersection of the black dotted line and the blue curve in fig. 4.6d, is oriented at ±27.4° from the y-axis and the fast axis of the second metasurface is oriented at ±27.4° from the x-axis. Because all the possible pairs of metasurfaces are given by equations (4.22) and (4.23), figs. 4.6c and 4.6d represent all the possible configurations.

4.4 Design of anti-reflective metasurfaces

In sections 4.2 and 4.3, ideal metasurfaces are considered and, as shown in section 4.2.1, the critical assumption is the absence of the reflection at the two interfaces of the metasurface. The other assumption is the existence of two linear eigenpolarizations that excite a single mode of the metasurface, which is true for symmetric cross-sections and a good approximation for asymmetric cross-sections. Therefore, a single mode metasurface can be seen as two independent Fabry-Pérot cavities with low finesse. For silicon-based metasurface on a glass substrate, the reflections at the interfaces of those Fabry-Pérot cavities are usually below 10%. The finesse decreases as material with lower refractive index is used.

The concept of Fabry-Pérot cavity has been described for the first time in 1899 [136] and, since then, it has been extensively studied. The theory on Fabry-Pérot cavities is given in many photonics books [92, 137]. With two highly reflective interfaces, the Fabry-Pérot cavity acts as a resonator. In many lasers, the gain medium is placed inside such resonators in order to enhance the field in that region and produce high power even if the gain per round trip is low. Fabry-Pérot can also be used as a filter for spectroscopy. On the other end, single layer anti-reflective coatings are Fabry-Pérot cavities where the reflection at the interfaces is low.

Metasurfaces can also act as anti-reflective coatings. The common anti-reflective metasurfaces are zeroth-order binary gratings, typically an array of cylinders, composed of the same material as the substrate. Compared to single layer anti-reflective coating, it does not require a material with a specific refractive index and it can be fabricated with a single etch. However, the performance decreases when the first order appears, but, as shown in [127] and in section 4.4.1, the decrease in performance is usually acceptable. Both binary anti-reflective metasurfaces and single layer anti-reflective coatings are broadband since they are Fabry-Pérot cavities with low finesse. It is less broadband than a multi-layer anti-reflective coating, but it is sufficient for many applications. Another group of metasurfaces are 3D structures that act as a smooth transition from the substrate to air [138].

In this section, a simple design technique is provided. This technique is applied in section 4.4.1, where all the possible anti-reflective metasurfaces composed of an array of cylinders for two different materials, glass and silicon, are given.

As mentioned before, single-mode metasurfaces can be seen as two independent Fabry-Pérot cavities and each Fabry-Pérot cavity can be represented by fig. 4.7a. Hence, the overall transmission efficiency T_{tot} is given by

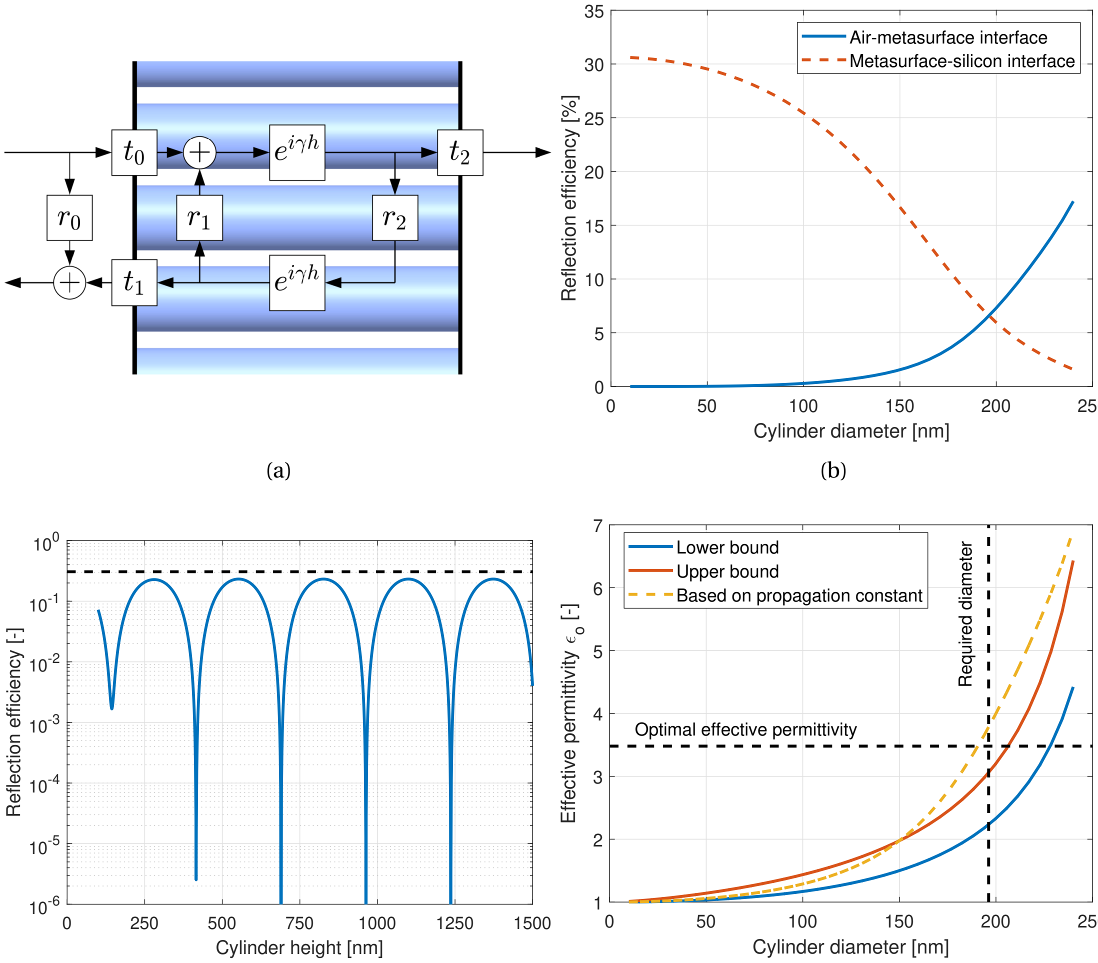
$$T_{tot} = \frac{|t_0|^2 |t_2|^2}{|1 - r_1 r_2 e^{2i\gamma h}|^2}, \tag{4.28}$$

where γ is the propagation constant of the eigen-mode and h is the metasurface thickness. t_0, t_2, r_1 and r_2 are the coupling coefficients shown in section 4.4.1 and they are complex numbers. Since no power is absorbed at the interfaces and assuming that no other modes that can carry power other than those shown in section 4.4.1, are excited, t_0, t_2, r_1 and r_2 are related by the equations

$$|t_0|^2 + |r_1|^2 = 1 \qquad |t_2|^2 + |r_2|^2 = 1. \tag{4.29}$$

T_{tot} is bounded by equations (4.11), where the reflection efficiencies R_1 and R_2 are given by $|r_1|^2$ and $|r_2|^2$ respectively. For anti-reflective coatings and metasurfaces, the important bound is the overall maximum transmission efficiency $T_{tot,max}$, which is

$$T_{tot,max} = \frac{(1 - R_1)(1 - R_2)}{(1 - \sqrt{R_1 R_2})^2} = \frac{(1 - R_1)(1 - R_2)}{(1 - R_1)(1 - R_2) + (\sqrt{R_2} - \sqrt{R_1})^2}. \tag{4.30}$$



(a) (b)

(c) (d)

Figure 4.7 – a) Schema of a Fabry-Pérot cavity, which describes the response of a mono-mode metasurface. The propagation direction of the incident plane wave is from left to right. r_0, t_0, r_1, t_1, r_2 and t_2 are the reflection and transmission coefficients at the interfaces. γ is the propagation constant of the eigen-mode and h is the metasurface thickness. b) Reflection efficiencies $|r_1|^2$ and $|r_2|^2$ at the two interfaces for a metasurface with a lattice constant of 250 nm with different cylinder diameters. c) Overall reflection efficiency for different metasurface thicknesses or cylinder heights for a metasurface with a lattice constant of 250 nm and a cylinder diameter of 196 nm. This is the cylinder diameter where the two curves in fig. 4.7b intersect each other. The horizontal dashed line is the reflection efficiency for a silicon-air interface. d) The Lichtenecker bounds for a metasurface with a lattice constant of 250 nm with different cylinder diameters. The yellow dashed line is the effective permittivity derived from the propagation constant of the eigen-mode propagating in the metasurface, assuming that the effective medium is non-magnetic. The optimal effective permittivity is the permittivity of an anti-reflective coating. The required diameter is the cylinder diameter of an anti-reflective metasurface.

Therefore, the transmission efficiency T_{tot} reaches 100% only if R_1 is equal to R_2. In order to design an anti-reflective metasurface, it is sufficient to find two metasurfaces where R_1 is larger than R_2 for one of the metasurface and the inverse for the other, and, by changing continuously for one metasurface to the other, it exists at least one metasurface that can be used as an anti-reflective metasurface. Then, the optimal thicknesses h_{AR} are found such that the term $r_1 r_2 e^{2i\gamma h}$ in (4.28) is positive real, and they are given by

$$h_{AR} = -\frac{\arg(r_1) + \arg(r_2) + 2\pi m}{2\gamma}, \tag{4.31}$$

where m is a real integer. If the chosen optimal thickness is too thin, the evanescent modes contribute to the overall transmission efficiency T_{tot} and the thickness may need further adjustment. With the Fourier modal method implemented in this work, such adjustment has a negligible computational cost.

When the structures that compose the metasurface is made of the same material as the substrate, a metasurface composed of large cylinders with respect to the unit cell has a low reflection efficiency at its interface with its substrate and a high reflection efficiency with its interface with air, and vice versa for a metasurface composed of small cylinders. Hence, for every unit cell dimensions, it exists a cylinder diameter such that the metasurface can act as an anti-reflective coating. Those metasurfaces are given in section 4.4.1. The same technique can be used to design an anti-reflective metasurface composed of an array of holes.

In fig. 4.7b, we show the reflection efficiencies at the two interfaces of a metasurface made of a square array of silicon cylinders on a silicon substrate. The lattice constant is 250 nm and the wavelength of the incident light is 1064 nm. As expected, the reflection efficiency at the interface between the metasurface and air increases as the diameter of the cylinders increases, and the reflection efficiency at the interface between the metasurface and the silicon substrate decreases. In order to have an anti-reflective metasurface, the cylinder diameter has to be 196 nm, which is where the two reflection efficiencies are equal. In fig. 4.7b, the reflection efficiencies are plotted for every cylinder diameters, but, in practice, the method of bisection is used in order to find the required diameter.

For the optimal cylinder diameter, the reflection efficiency for different cylinder heights is plotted in fig. 4.7c. The position of the dips is accurately given by equation (4.31) except the first dip, where the error on the optimal cylinder height is 1.3%. The reason of this error and also why the minimum reflection efficiency is still above 0.1% is that the metasurface thickness is thin enough for the evanescent waves to play a role on the overall performance. To improve even further the minimum reflection efficiency, the diameter of the cylinders needs to be adjusted. The effect of the evanescent waves can be seen at a lesser extent on the second dip, where the minimum reflection efficiency is higher than for the third dip. The dashed line in fig. 4.7c is the reflection efficiency without anti-reflective metasurface.

In the final part of this section, a first guess on the optimal diameter of the cylinders for an anti-reflective metasurface can be done by using the approximation provided by the effective permittivity theory, which predicts the effective permittivity from the cross-section of the metasurface assuming that the unit cell dimensions are negligible compared to the wavelength. This assumption is, in most cases, not true in the near-infrared regime due to fabrication issues, but it is interesting to see if it can still be applied.

For an anti-reflective coating, the condition on the material parameters such that the reflection at the interfaces are equal, is

$$Z_{AR} = \sqrt{Z_1 Z_2}, \tag{4.32}$$

where Z_{AR}, Z_1 and Z_2 are the wave impedance of, respectively, the anti-reflective coating and the medium below and above. For a uniaxial medium with the extraordinary axis perpendicular to the interfaces, the wave impedance is defined in terms of the ordinary permettivity and permeability, meaning that $Z = \sqrt{\mu_t / \epsilon_t}$, where ϵ_t and μ_t are the ordinary permittivity and permeability respectively.

In the effective permittivity theory, metasurfaces with a symmetric cross-section can be approximated by such uniaxial homogeneous medium and, in [139], the bounds of the ordinary permittivity ϵ_L, called the Lichtenecker bounds, have been provided, which are

$$\int_0^b \frac{dy}{\int_0^a \frac{dx}{\epsilon(x,y)}} \leq \epsilon_L \leq \frac{1}{\int_0^a \frac{dx}{\int_0^b \epsilon(x,y) dy}}, \tag{4.33}$$

where $\epsilon(x,y)$ is the permittivity profile of the metasurface and a and b are the unit cell dimensions along x and y respectively. The interfaces are perpendicular to the z-axis. The advantage of those bounds is that they take the geometry of the unit cell into account. For a metasurface composed of a square array of cylinders, the bounds in (4.33) reduces to

$$\epsilon_s \left(1 - \frac{D}{L} + \frac{1}{2} \int_{-1}^{1} \frac{dx}{\frac{L}{D} + (\frac{\epsilon_s}{\epsilon_c} - 1)\sqrt{1 - x^2}} \right) \leq \epsilon_L \leq \epsilon_s \left(1 - \frac{D}{L} + \frac{1}{2} \int_{-1}^{1} \frac{dx}{\frac{L}{D} + (\frac{\epsilon_c}{\epsilon_s} - 1)\sqrt{1 - x^2}} \right)^{-1}, \tag{4.34}$$

where L is the lattice constant, D is the diameter of the cylinders, ϵ_c is the permittivity inside the cylinders and ϵ_s is the permittivity outside the cylinders.

Using equation (4.32), the permittivity required for an anti-reflective coating between silicon and air is the refractive index of silicon, which is $n_{Si} = 3.48$ in the example given in this section. As shown in 4.7d, the diameter of the cylinders for an anti-reflective metasurface does not cross the permittivity of an anti-reflective coating within the Lichtenecker bounds. However, the upper bound can still be used as a first guest.

4.4.1 Design of glass and silicon anti-reflective metasurfaces

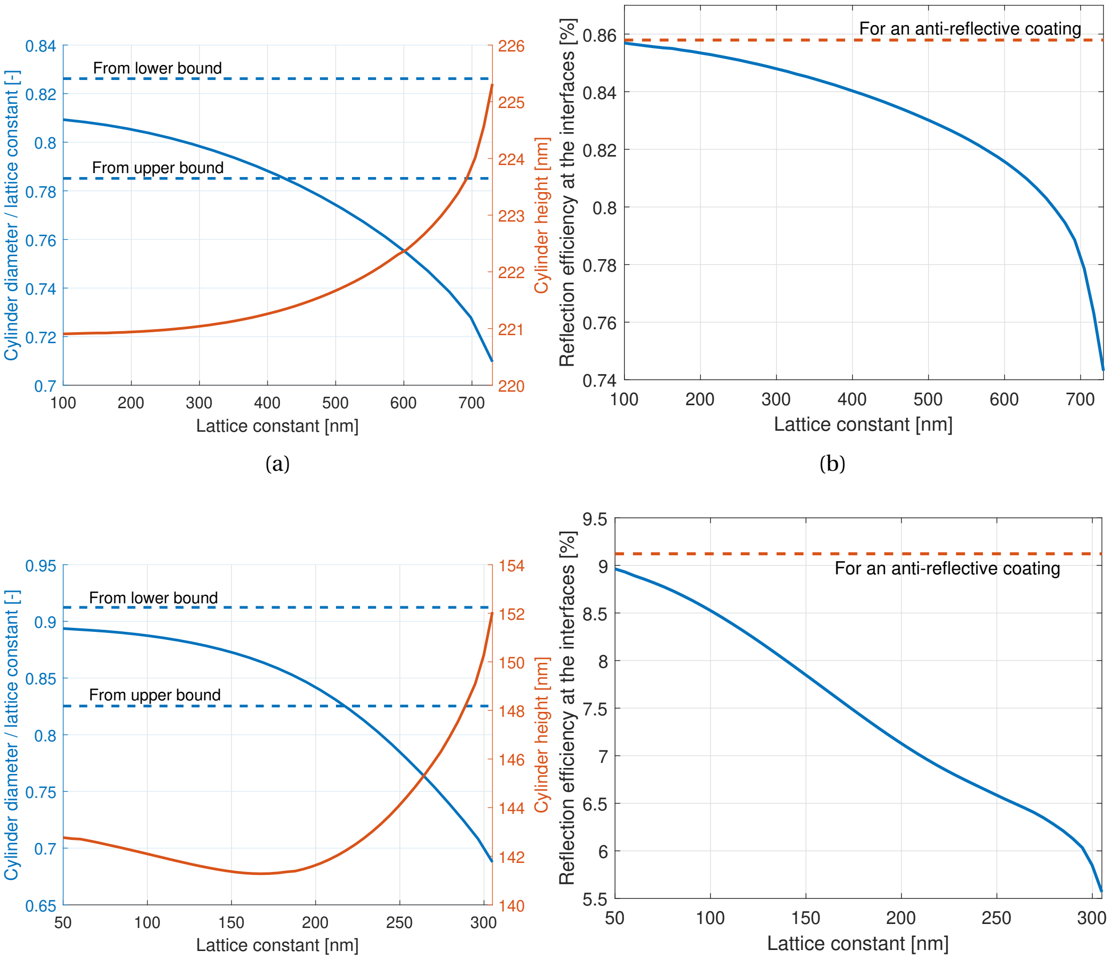
In this section, the design technique presented in section 4.4 is used to obtain all the possible metasurfaces composed of a square array of cylinders for two different materials: glass ($n_{glass} = 1.45$) and silicon ($n_{Si} = 3.48$). The cylinders are made of the same material as the substrate, meaning that such metasurface can be fabricated with a single etch. The metasurfaces are designed for a wavelength of 1064 nm, but, by scaling the dimensions appropriately and assuming that the material is not dispersive, the anti-reflective metasurface for another wavelength is obtained. As mentioned in section 4.4, it is possible to design an anti-reflective metasurface for any given lattice, but the choice of the lattice has an impact on the performance of the anti-reflective metasurface. In the second part of this section, an anti-reflective metasurface made of glass is studied in depth. Such metasurfaces are interesting candidates for high-power application [140] since the substrate withstands high power and the metasurface is made of the same high-quality material. Therefore, quantities such as the field enhancement and the energy flux inside the structures are given and compared with the case of a single layer anti-reflective coating.

As discussed in section 4.4, for any given lattice constant, it exists a metasurface with a specific cylinder diameter and multiple cylinder heights such that this metasurface is anti-reflective. Those dimensions are given in fig. 4.8a for glass anti-reflective metasurfaces and in fig. 4.8c for silicon metasurfaces. The only condition is that the metasurface can be described as two independent and identical Fabry-Pérot cavities, one per polarization. Hence, the metasurface has to be a zeroth-order grating and a single-mode metasurface, meaning that the lattice constant Λ has to be smaller than

$$\Lambda = \frac{\lambda}{n_{sub}}, \tag{4.35}$$

where n_{sub} is the refractive index of the substrate, or of the superstrate if its refractive index is higher. The maximum lattice constant is 734 nm for metasurfaces on a glass substrate and 306 nm for metasurfaces for a silicon metasurfaces. The different ranges for the x-axis in fig. 4.8 are due to this maximum lattice constant. The dimensions of the anti-reflective metasurfaces shown in figs. 4.8a and 4.8c with respect to the lattice constant have similar behaviors: the ratio of the cylinder diameter to the lattice constant decreases and the cylinder height increases as the lattice constant increases, but the increase of the cylinder height can be considered as negligible. Hence, a metasurface is easier to fabricate for large lattice constant.

In figs. 4.8a and 4.8c, the diameter of the cylinders obtained by using the Lichtenecker bounds given in (4.33) is indicated by dashed lines. The Lichtenecker bounds are valid only if the lattice constant is negligible compared to the wavelength and, as expected, the ratio of the cylinder diameter to the lattice constant is within the bounds for a small enough lattice constant.



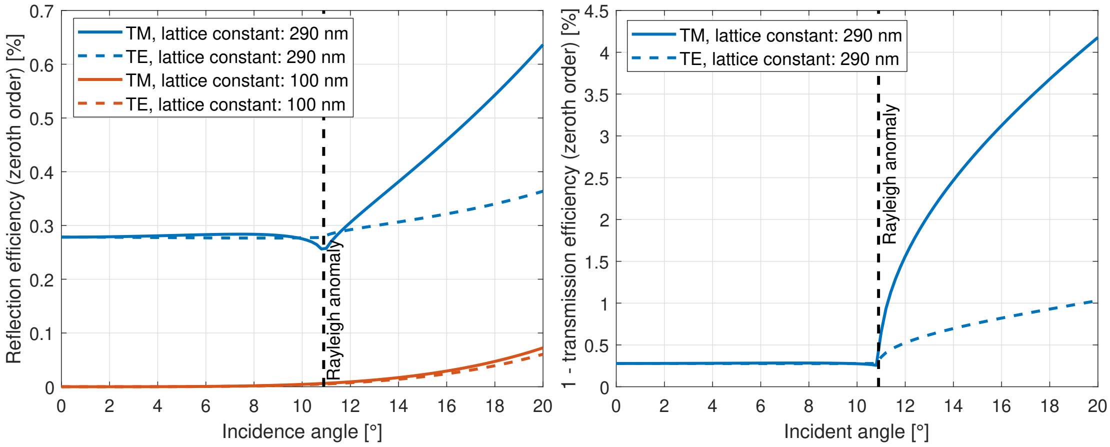
(a) (b)

(c) (d)

Figure 4.8 – a) The cylinder height and the ratio of the cylinder diameter to the lattice constant of an anti-reflective metasurface composed of glass cylinders on a glass substrate in function of the lattice constant. The two dashed line are the ratio of the cylinder diameter to the lattice constant obtained when the effective permittivity is given by the lower and upper Lichtenecker bounds. b) The reflection efficiency at the two interfaces of the metasurface in function of the lattice constant. This reflection efficiency is compared to the reflection efficiency at the interfaces of a single-layer anti-reflective coating for a glass substrate. c) The cylinder height and the ratio of the cylinder diameter to the lattice constant of an anti-reflective metasurface composed of silicon cylinders on a silicon substrate in function of the lattice constant. The two dashed lines are the ratio of the cylinder diameter to the lattice constant obtained when the effective permittivity is given by the lower and upper Lichtenecker bounds. d) The reflection efficiency at the two interfaces of the metasurface in function of the lattice constant. This reflection efficiency is compared to the reflection efficiency at the interfaces of a single-layer anti-reflective coating for a silicon substrate.

Moreover, this ratio seems to converge to the lower bound, which is known to approximate better the effective permittivity for small enough lattice constant [139].

Figures 4.8b and 4.8d show the relationship between the lattice constant and the reflection efficiency at the interfaces. If the theory on effective permittivity is valid, the reflection efficiency at the interfaces of the metasurface should be the same as the one at the interface of the substrate and an anti-reflecting coating and, as expected, it does if the lattice constant is small enough. In general, the reflection efficiency at the interfaces of the metasurface is always smaller than the one at the interfaces of an anti-reflective coating. From the theory on Fabry-Pérot cavity and since the thickness of the metasurface is approximately the same as the thickness of the anti-reflective coating, which is 221 nm for a glass substrate and 143 nm for a silicon substrate, it means that the response of the metasurface should be less sensitive to a change in wavelength and incidence angle. As shown below, it is more complicate due to the presence of evanescent waves.



(a)

(b)

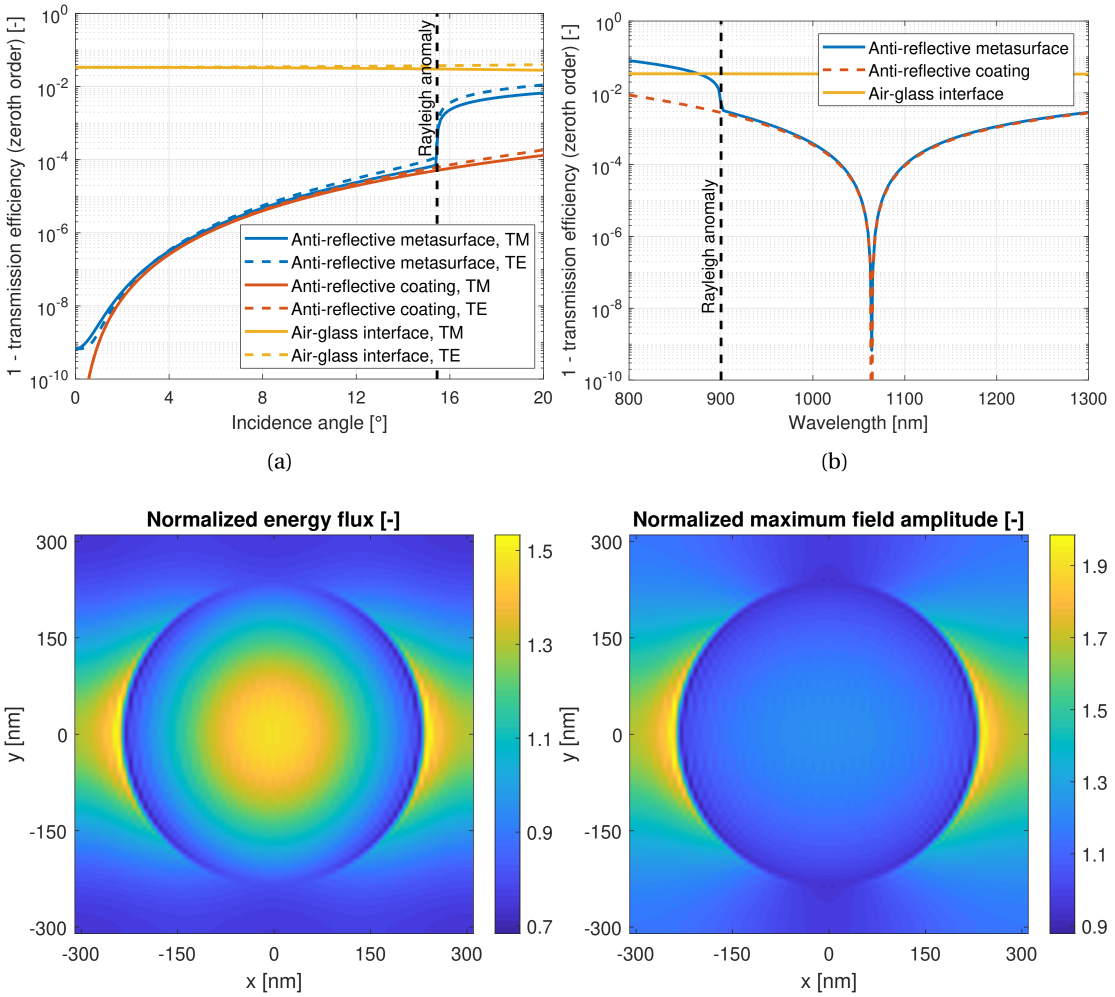
Figure 4.9 – a) The reflection efficiency, which take into account only the zeroth order, of two anti-reflective metasurfaces composed of silicon cylinders on a silicon substrate in function of the incidence angle. The lattice constant of those two metasurfaces is 100 nm and 290 nm. The Rayleigh anomaly corresponds to the appearance of a second propagating order in the silicon substrate. b) The transmission efficiency, which takes into account only the zeroth order, of an anti-reflective metasurface composed of silicon cylinders on a silicon substrate in function of the incidence angle. The lattice constant of the metasurface is 290 nm. The y-axis is one minus the transmission efficiency in order to be compared with fig. 4.9a

To illustrate the effect of the lattice constant on the performance of anti-reflective metasurfaces, two silicon anti-reflective metasurfaces are investigated with very different lattice constants. The first metasurface has a lattice constant of 100 nm, a cylinder diameter of 89 nm and a cylinder height of 142 nm, and the second metasurface has a lattice constant of 290 nm, a cylinder diameter of 212 nm and a cylinder height of 147 nm. First of all, metasurfaces with a small lattice constant are more difficult to fabricate. Typically, the cylinders in the metasurface with the small lattice constant may merge during fabrication due to the small gaps between the cylinders which have an aspect ratio of nearly 13. The merging of two cylinders is not an issue by itself but the merging may not be homogeneous within the metasurface which lead to a source of scattering due to a loss of the periodicity of the metasurface. A solution would be to design a metasurface composed of holes. Such designs are not shown in this work but the design technique is the same as the one proposed in section 4.4.

From the viewpoint of fabrication, it is better to design an anti-reflective metasurface with a large lattice constant, but, as shown in fig. 4.9 where the reflection and transmission efficiencies are plotted for different angles of incidence, drawbacks are present. The first drawback is the presence of the Rayleigh anomaly occurring at an incidence angle of 10.9°, where the first order can propagate in the silicon substrate. While increasing the incidence angle after the Rayleigh anomaly, the performance degrades faster. If the purpose of the anti-reflective metasurface is to reduce reflection, which is typically the case for solar cells [127], the degradation in performance is not so severe as shown in fig. 4.9a. However, if the purpose of the anti-reflective metasurface is to maximize the transmission efficiency, the degradation in performance is more critical as shown in fig. 4.9b. The second drawback of a large lattice constant is that the evanescent modes play a role in the performance of the metasurface, limiting the maximum transmission efficiency to around 99.72% for a lattice constant of 290 nm. The improvement in transmission efficiency is still important compared to the transmission efficiency of a silicon-air interface, which is 69.4%. A solution to counter the effect of the evanescent modes without changing the lattice constant is to choose a larger thickness. From equation (4.31) and as shown in fig. 4.7c, dips in the reflection efficiency occur periodically when varying the cylinder height. For a silicon metasurface, the second dip occurs for a cylinder height of around 420 nm, which is a bit less than three times higher than the cylinder height related to the first dip, losing the advantage of a large lattice constant on the fabrication. For a glass metasurface, the second dip occurs for a cylinder height of around 660 nm.

Anti-reflective metasurfaces are well suited for high-power applications since it is obtained by structuring the substrate, which is made to withstand high power, instead of depositing an anti-reflective material. In the last part of this section, a design of a glass anti-reflective metasurface is proposed along with its performance, maximum energy flux and maximum field amplitude. Those values are compared with the ones from an anti-reflective coating. The chosen lattice constant of the metasurface is 620 nm and, using the design technique proposed in section 4.4, the obtained cylinder diameter and height are respectively 488 nm and 222 nm. This choice of the lattice constant leads to cylinders with a low aspect ratio, less than 0.5, while, as shown in fig. 4.10a, having a very low maximum reflection efficiency of around 10^{-9} and a Rayleigh anomaly occurring at an angle of incidence of around 15.4°. If, for the desired application, the Rayleigh anomaly occurs for a too small angle of incidence, the metasurface has to be designed with a smaller lattice constant.

As shown in figs. 4.10a and 4.10b, the performance of an anti-reflective metasurface is very similar to the performance of an anti-reflective coating. The main differences are the presence



(a) (b)

(c) (d)

Figure 4.10 – a) The transmission efficiency, which takes into account only the zeroth order, of an anti-reflective metasurface composed of glass cylinders on a glass substrate and an anti-reflective coating for a glass substrate in function of the incidence angle. The Rayleigh anomaly corresponds to the appearance of a second propagating order in the glass substrate. Those values are compared with the transmission efficiency of an air-glass interface. b) The transmission efficiency of the same glass anti-reflective metasurface and anti-reflective coating in function of the wavelength. c) Energy flux of the propagating mode. The energy flux is normalized such that its average over the unit cell is one. d) Maximum amplitude of the electric field on a plane parallel to the interface between the glass substrate and the metasurface separated by a distance of 32 nm from this interface. The maximum field amplitude is normalized such that the maximum field amplitude in the glass substrate is one.

of the Rayleigh anomaly and that the reflection efficiency cannot go below 10^{-9} due to the impact of the evanescent modes on the performance.

Two quantities that are important to predict material failure due to a high-power beam are the energy flux and the maximum field amplitude. The energy flux is related to absorption in the material due to defects or impurities and this is usually the main mechanism for material failure when using a continuous wave laser. A high field amplitude can generate photon with smaller wavelength through non-linear effect, which are then absorbed, or it can ionized the atoms, creating defects inside the material. More detail about the different mechanisms which lead to material failure are described in chapter 1 of [141]. To estimate the maximum energy flux, the propagating mode is assumed to be the only mode that carry the power from one interface to the other. This is a reasonable approximation since the evanescent modes carry only 0.06% of the total power. Therefore, the maximum energy flux P_{max} is given by

$$P_{max} = P_{mode,max} \cdot (|a|^2 + |b|^2), \tag{4.36}$$

where $P_{mode,max}$ is the normalized energy flux of the propagating mode at the location inside the cylinders where it reaches its maximum, a is the weight of the forward propagating mode and b is the weight of the backward propagating mode. As shown in fig. 4.10c, the maximum energy flux of the propagating mode occurs at the center of the cylinder and it reaches 1.47. Due to the mode weights a and b and setting the energy flux before the metasurface to one, the maximum energy flux rises to 1.49, meaning that the mode profile is the main contributor to the maximum energy flux. For comparison, the energy flux inside an anti-reflecting coating is 1.02. The maximum field amplitude inside the metasurface is shown in fig. 4.10d and the plane where the field amplitude is computed, is located at 32 nm from the glass-metasurface interface. The field amplitude at the center of the cylinder reaches its maximum at that location. The field amplitude is the maximum field amplitude reached during a time period and is normalized to the field amplitude inside the glass substrate. Hence, the field amplitude E_{max} is given by

$$E_{max} = \frac{1}{\sqrt{2}E_{glass}} \sqrt{|E_x|^2 + |E_y|^2 + |E_z|^2 + |E_x^2 + E_y^2 + E_z^2|}, \tag{4.37}$$

where E_{glass} is the field amplitude inside the glass and E_x, E_y and E_z are the components of the complex electric field. In the metasurface, the field amplitude reaches its maximum of 1.22 at the center of the cylinders and, in an anti-reflective coating, the maximum field amplitude is 1.20.

4.5 Design of a half-wave plate

A half-wave plate is an optical element that changes the polarization angle of a linearly polarized beam. It is usually made of a birefringent material and the thickness is chosen such that the difference of phase accumulations between a beam linearly polarized along the slow axis and a beam linearly polarized along the fast axis while going through the birefringent material, called retardance, is 180°. Metarsurfaces that act as a half-wave plate have been proposed in the literature [51, 128]. In addition, metasurfaces based on the Pancharatnam-Berry phase [31–35, 38] are local half-wave plates with different orientations.

In this section, a design of a metasurface acting as a half-wave plate at a wavelength of 1550 nm is proposed. Since a waveplate with a higher retardance is equivalent to a waveplate with a retardance equal or lower than 180°, such metasurface is the most difficult metasurface to fabricate that mimics a waveplate. Hence, this section gives the dimensions that such metasurface has, along with the design technique.

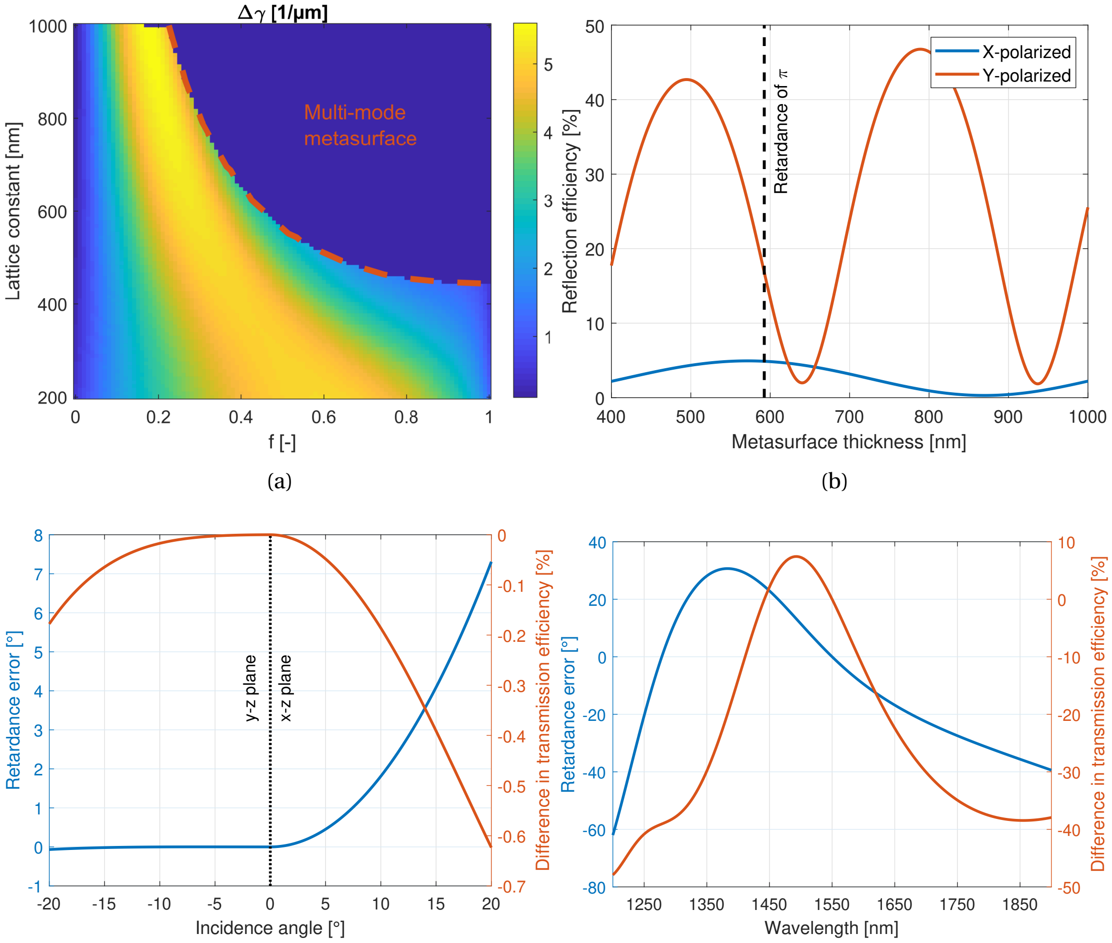
The metasurface is made of parallel lines, which is the simplest structure that mimics a birefringent material and is also simpler to simulate, design and fabricate. The lines are made of silicon and substrate is glass. Glass substrate allows to have a larger lattice constant without having a propagating first order, and the propagation constant difference is larger for the optimal line width, resulting in a thinner metasurface, but this choice has drawbacks as discussed later.

The first step in the design process is to find the metasurface dimensions such that the difference of propagation constants is maximized, allowing to minimized the metasurface thickness. Since the propagation constants do not depend of the thickness of the metasurface, all the possible metasurfaces composed of lines are described by two parameters: the lattice constant and the ratio of the width of the lines to the lattice constant. In this work, a different definition of the ratio, called f, is used, which is given by

$$f = \frac{w - w_{min}}{l - 2w_{min}}, \tag{4.38}$$

where l is the lattice constant, w is the width of the lines and w_{min} is the feature size. This definition of the ratio ensures that, by choosing a value for f between zero and one, the minimum feature size is always below w_{min} for any value of the lattice constant. In this section, w_{min} is 5 nm.

The difference of propagation constants for all possible metasurfaces composed of lines is given in fig. 4.11a. An important feature is that, for each lattice constant, there is a ratio f where the difference of propagation constants reaches a maximum and the maximum value of this difference only weakly changes for the different lattice constants. Another feature is this large zone where the metasurface is multi-mode. A multi-mode metasurface behaves differently than a single-mode metasurface as shown in chapter 5, and the final design should



(a) (b)

(c) (d)

Figure 4.11 – a) Propagation constant difference in function of the ratio f defined in (4.38) and the lattice constant. The red dashed line delimits the region where the metasurface is multi-mode. b) The reflection efficiency for a x-polarized and a y-polarized incident beam in function of the metasurface thickness. The black dashed line is the metasurface thickness required for a phase retardance of 180° when the lattice constant is 700 nm and the line width is 176 nm ($f = 0.246$). c) Retardance error in function of the incidence angle. For the left half of the plot, the plane of incidence is the y-z plane, meaning that the plane is parallel to the lines that compose the metasurface. For the right half of the plot, the plane of incidence is the x-z plane, meaning that the plane is perpendicular to the lines. d) The retardance error and the difference between the transmission efficiency of an x-polarized and a y-polarized input beam in function of the wavelength.

be far enough of this zone in order to avoid a too abrupt degradation of the performance while changing the wavelength or the angle of incidence. For this reason, the chosen lattice constant is 700 nm and the maximum difference of propagation constants occurs for a line width of 176 nm. However, as shown in fig. 4.11b, the transmission efficiency of a x-polarized beam, which is polarized perpendicular to the lines and along the fast axis, is different than the transmission efficiency of a y-polarized beam, which is polarized along the slow axis. In order for a metasurface to act as a half-wave plate, it is important that the transmission efficiencies for both polarizations are equal. An intuitive solution would be to decrease the difference of propagation constants by increasing or decreasing the ratio f such that the retardance of 180° occurs at the metasurface thickness where the blue and red curves in fig. 4.11b cross each other. However, the fringes tend to move faster than the thickness where the retardance of 180° occurs. In the current case, the fringes of the red curve has to move left-ward, meaning that, from the theory on Fabry-Pérot cavities, the propagation constant of the mode excited by a y-polarized beam has to increase which happens when the ratio f increases. Hence, a metasurface with a lattice constant of 700 nm acts as a half-wave plate for a line width of 200 nm and a thickness of 586 nm.

The performance of the metasurface for a variation in the incidence angle or the wavelength is shown in figs. 4.11c and 4.11d. The performance of the metasurface is quite robust for a change in incidence angle, especially when the illumination plane is parallel to the lines. However, for a change in wavelength, the performance decreases rapidly. The difference in transmission efficiency can be reduced by choosing a material with a lower refractive index than silicon since it leads to those high amplitude fringes shown in fig. 4.11b, but, then, the aspect ratio of the structure increases and it is already around three. The overall transmission efficiency is 92.0%.

4.6 Conclusion

Two aspects of single-mode metasurfaces are considered in this chapter and design techniques are provided based on them. The first aspect is the notion that single-mode metasurfaces behave approximately as ideal metasurfaces, which is characterized by three parameters. For a given polarization state of the incident illumination and a desired polarization state of the transmitted fields, we provide the equations that directly give the possible ideal metasurfaces that have such functionalities. Based on this result, the design process for four different types of holograms is proposed.

Ideal metasurfaces are equivalent to ideal waveplates, which means that the Jones matrix that describes them, is symmetric, which limits the functionalities that they can offer. In order to get around this limitation, two ideal metasurfaces are needed and we show that any system described by a Hermitian matrix can be described by two ideal metasurfaces. We also provide the equations that give all the possible combinations of ideal metasurfaces for a given functionality.

The second aspect is the notion that single-mode metasurfaces can be described by two independent Fabry-Pérot cavities. Based on that description, we propose a technique for designing anti-reflective metasurfaces, and we apply it to obtain all the possible anti-reflective metasurfaces consisting of a square array of cylinders, where the material of the cylinders is the same as that of the substrate, and the material is either glass or silicon. The concept of the Fabry-Pérot cavity is also used to design a half-wave plate.

Design techniques based on these two aspects cover most of the applications that can be realized by single-mode metasurfaces. At the same time, they provide an understanding of the intrinsic mechanisms of single-mode metasurfaces.

4.7 Proofs

4.7.1 Proof of the symmetric Jones matrix from a transformation of polarization state

In section 4.2, the Jones matrix, which is symmetric and unitary for ideal metasurfaces and which transforms a polarization state $\vec{p}_1$ into the polarization state $\vec{q}_1$ is given. $\vec{p}_1$ and $\vec{q}_1$ are normalized. The parameter s is defined in (4.2) and, when s is zero, multiple Jones matrices perform the same transformation of polarization states. This section is divided into two parts. The first part is to prove that a unique solution exists if s is different than zero and that all the solutions are expressed in (4.3) when s is zero. The second part is to prove that the expressions in equations (4.2) and (4.3) are correct.

Because $\hat{T}_{tot}$ is a unitary matrix, its elements have to satisfy a set of constraints. For a general 2×2 matrix $\hat{A}$ given by

$$\hat{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{4.39}$$

$\hat{A}$ is unitary if the following equations are satisfied:

$$|a| = |d| \tag{4.40a}$$

$$|a|^2 + |b|^2 = 1 \tag{4.40b}$$

$$a\bar{b} + c\bar{d} = 0. \tag{4.40c}$$

.

For a symmetric unitary matrix such as $\hat{T}_{tot}$, the property (4.40a) is redundant. The properties (4.40) are also used in section 4.7.2 and section 4.7.3.

Along with properties (4.40b) and (4.40c), the equation that needs to be solved is

$$\hat{T}_{tot}\vec{p}_1 = \vec{q}_1, \tag{4.41}$$

which leads to the following set of equations

$$\begin{aligned} a_{tot}p_{x,1} + b_{tot}p_{y,1} &= q_{x,1} \\ b_{tot}p_{x,1} + d_{tot}p_{y,1} &= q_{y,1}, \end{aligned} \tag{4.42}$$

where a_{tot}, b_{tot} and d_{tot} are the unknowns.

For the case $b_{tot} = 0$, the amplitude of a_{tot} and d_{tot} is one due to equation (4.40b), so the amplitude of $p_{x,1}$ and $p_{y,1}$ are equal to the amplitude of $q_{x,1}$ and $q_{y,1}$ respectively due to equations (4.42). Therefore, if $p_{x,1}$ or $p_{y,1}$ is zero, s, which is

$$s = p_{x,1}\bar{q}_{y,1} - p_{y,1}\bar{q}_{x,1}, \tag{4.43}$$

is also zero and, as discussed in section 4.2, the expressions in (4.3) converge to the correct solution. When $p_{x,1}$ and $p_{y,1}$ is different than zero, equations (4.42) admit a single solution, which is given by

$$a_{tot} = \frac{q_{x,1}}{p_{x,1}} \qquad d_{tot} = \frac{q_{y,1}}{p_{y,1}}. \tag{4.44}$$

For the case $b_{tot} \neq 0$, the second equation is multiplied by $\bar{b}_{tot}\bar{p}_{x,1}$ and the property (4.40c) is applied, leading to

$$a_{tot}p_{x,1} = q_{x,1} - b_{tot}p_{y,1} \tag{4.45a}$$

$$|b_{tot}|^2|p_{x,1}|^2 - \bar{a}_{tot}b_{tot}\bar{p}_{x,1}p_{y,1} = \bar{b}_{tot}\bar{p}_{x,1}q_{y,1}. \tag{4.45b}$$

The case $p_{x,1} = 0$ is treated later. By inserting equation (4.45a) into equation (4.45b) and since the norm of the vector $\vec{p}_1$ is one, the following complex equation is obtained:

$$\bar{p}_{x,1}q_{y,1}\bar{b}_{tot} + p_{y,1}\bar{q}_{x,1}b_{tot} = |b_{tot}|^2. \tag{4.46}$$

Since the systems of equations (4.40) and (4.42) are the same after the permutation

$$(a_{tot}, d_{tot}, p_{x,1}, p_{y,1}, q_{x,1}, q_{y,1}) \longleftarrow (d_{tot}, a_{tot}, p_{y,1}, p_{x,1}, q_{y,1}, q_{x,1}), \tag{4.47}$$

equation (4.46) is also obtained after multiplying the adequate equation by $p_{y,1}$ instead of $p_{x,1}$, meaning that equation (4.46) is also valid for the case $p_{x,1} = 0$. For every b_{tot} different than zero and solution of equation (4.46), it exists a single a_{tot} and d_{tot} that satisfy equations (4.42) and (4.40c).

The imaginary part of equation (4.46) is

$$\begin{aligned}&(\bar{p}_{x,1}q_{y,1} - \bar{p}_{y,1}q_{x,1})\bar{b}_{tot} - (p_{x,1}\bar{q}_{y,1} - p_{y,1}\bar{q}_{x,1})b_{tot} = \bar{s}\bar{b}_{tot} - sb_{tot} = 0\\ &s = p_{x,1}\bar{q}_{y,1} - p_{y,1}\bar{q}_{x,1}.\end{aligned} \tag{4.48}$$

Hence, if s is not equal to zero, $b_{tot} = k\bar{s}$ is a solution of equation (4.48) for k real and different than zero, since equation (4.46) is valid only if $b_{tot} \neq 0$.

By replacing b_{tot} in equation (4.46) with the solution of equation (4.48) and after dividing by k, equation (4.46) becomes

$$\begin{aligned}&\bar{p}_{x,1}q_{y,1}s + p_{y,1}\bar{q}_{x,1}\bar{s} = k|s|^2\\ \Rightarrow&|p_{x,1}|^2|q_{y,1}|^2 - |p_{y,1}|^2|q_{x,1}|^2 = |p_{x,1}|^2 - |q_{x,1}|^2 = k|s|^2\\ \Rightarrow&k = \frac{|p_{x,1}|^2 - |q_{x,1}|^2}{|s|^2}.\end{aligned} \tag{4.49}$$

Therefore, for s and b_{tot} different than zero, the expression of b_{tot} in (4.2) is obtained, and the solution of the systems of equations (4.40) and (4.42) is unique.

If s is zero, equation (4.48) is satisfied for any b_{tot} and equation (4.46) becomes

$$\bar{p}_{x,1}q_{y,1}\bar{b}_{tot} + p_{x,1}\bar{q}_{y,1}b_{tot} = |b_{tot}|^2. \tag{4.50}$$

In the complex plane, a circle of radius r and center z_0 is described by

$$\begin{aligned}&|z - z_0|^2 = r^2\\ \Rightarrow&|z|^2 = r^2 - |z_0|^2 + z_0\bar{z} + \bar{z}_0 z.\end{aligned} \tag{4.51}$$

Therefore, the solution of equation (4.50) is a circle of center $\bar{p}_{x,1}q_{y,1}$ touching the origin, meaning that b_{tot} can be described by its expression in (4.3), which is

$$b_{tot} = \bar{p}_{x,1}q_{y,1}(1 + e^{i\phi}). \tag{4.52}$$

When the term $1+e^{i\phi}$ is zero and $p_{x,1}$ and $q_{y,1}$ are different than zero, the expressions in (4.3) converge to the solution (4.44). The case when s, b_{tot} and either $p_{x,1}$ or $q_{y,1}$ are zero has already been treated.

In the previous part of this section, the expressions of b_{tot} in (4.2) and (4.3) are proved when b_{tot} is different than zero. For the case $b_{tot}=0$, the solution is unique when s is different than zero and the expressions in (4.3) converge to the proper solution when s is zero. In the remaining part of this section, it is proved that the expressions in (4.2) and (4.3) satisfy equations (4.42) and (4.40).

Due to the invariance of the system of equations (4.40) and (4.42) to the transformation (4.47), the expressions (4.2) are correct if it satisfies equation (4.40b), equation (4.40c) and one equation in (4.42). Therefore, the following equations have to be satisfied:

$$|a_{tot}|^2+|b_{tot}|^2=1 \tag{4.53a}$$
$$a_{tot}\bar{b}_{tot}+b_{tot}\bar{d}_{tot}=0 \tag{4.53b}$$
$$a_{tot}p_{x,1}+b_{tot}p_{y,1}=q_{x,1}. \tag{4.53c}$$

For the case $s\neq 0$, the expressions given in (4.2) are

$$\begin{aligned} a_{tot} &= \frac{q_{x,1}\bar{q}_{y,1}-\bar{p}_{x,1}p_{y,1}}{s}\\ b_{tot} &= \frac{|p_{x,1}|^2-|q_{x,1}|^2}{s}\\ d_{tot} &= \frac{p_{x,1}\bar{p}_{y,1}-\bar{q}_{x,1}q_{y,1}}{s}\\ s &= p_{x,1}\bar{q}_{y,1}-p_{y,1}\bar{q}_{x,1}. \end{aligned} \tag{4.54}$$

In order to prove that the expressions in (4.54) satisfy equations (4.53), a set of equations is provided. Related to equation (4.53a):

$$\begin{aligned} &|a_{tot}|^2|s|^2=|p_{x,1}|^2|p_{y,1}|^2+|q_{x,1}|^2|q_{y,1}|^2-p_{x,1}\bar{p}_{y,1}q_{x,1}\bar{q}_{y,1}-\bar{p}_{x,1}p_{y,1}\bar{q}_{x,1}q_{y,1}\\ &|a_{tot}|^2|s|^2-|s|^2=|p_{x,1}|^2|p_{y,1}|^2+|q_{x,1}|^2|q_{y,1}|^2-|p_{x,1}|^2|q_{y,1}|^2-|p_{y,1}|^2|q_{x,1}|^2\\ &|b_{tot}|^2|s|^2=(|p_{x,1}|^2-|q_{x,1}|^2)(|q_{y,1}|^2-|p_{y,1}|^2). \end{aligned} \tag{4.55}$$

Related to equation (4.53b):

$$\begin{aligned} &(a_{tot}\bar{b}_{tot}+b_{tot}\bar{d}_{tot})|s|^2=(a_{tot}s+\bar{d}_{tot}\bar{s})(|p_{x,1}|^2-|q_{x,1}|^2)\\ &a_{tot}s+\bar{d}_{tot}\bar{s}=0. \end{aligned} \tag{4.56}$$

Related to equation (4.53c):

$$\begin{aligned}
a_{tot} p_{x,1} s &= q_{x,1} p_{x,1} \bar{q}_{y,1} - |p_{x,1}|^2 p_{y,1} \\
b_{tot} p_{y,1} s &= |p_{x,1}|^2 p_{y,1} - q_{x,1} p_{y,1} \bar{q}_{x,1} \\
a_{tot} p_{x,1} s + b_{tot} p_{y,1} s &= q_{x,1} s.
\end{aligned} \tag{4.57}$$

For the case $s = 0$, the expressions given in (4.2) are

$$\begin{aligned}
a_{tot} &= \frac{q_{x,1}}{p_{x,1}} (1 - |p_{y,1}|^2 (1 + e^{i\phi})) \\
b_{tot} &= \bar{p}_{x,1} q_{y,1} (1 + e^{i\phi}) \\
d_{tot} &= \frac{q_{y,1}}{p_{y,1}} (1 - |p_{x,1}|^2 (1 + e^{i\phi})).
\end{aligned} \tag{4.58}$$

and the amplitude of $p_{x,1}$ and $q_{x,1}$ are equal since

$$|p_{x,1}|^2 |q_{y,1}|^2 = |p_{y,1}|^2 |q_{x,1}|^2 \Rightarrow |p_{x,1}|^2 (1 - |q_{x,1}|^2) = (1 - |p_{x,1}|^2) |q_{x,1}|^2. \tag{4.59}$$

In order to prove that the expressions in (4.58) satisfy equations (4.53), a set of equations is provided. Related to equation (4.53a):

$$\begin{aligned}
&|1 + e^{i\phi}|^2 = 2 + e^{i\phi} + e^{-i\phi} \\
&|a_{tot}|^2 = |1 - |p_{y,1}|^2 (1 + e^{i\phi})|^2 = 1 - |p_{y,1}|^2 (2 + e^{i\phi} + e^{-i\phi} - |p_{y,1}|^2 |1 + e^{i\phi}|^2) \\
&|b_{tot}|^2 = (1 - |p_{y,1}|^2) |p_{y,1}|^2 |1 + e^{i\phi}|^2.
\end{aligned} \tag{4.60}$$

Related to equation (4.53b):

$$\begin{aligned}
&|1 + e^{i\phi}|^2 = 2 + e^{i\phi} + e^{-i\phi} \\
&a_{tot} \bar{b}_{tot} = q_{x,1} (1 - |p_{y,1}|^2 (1 + e^{i\phi})) \bar{q}_{y,1} (1 + e^{-i\phi}) = q_{x,1} \bar{q}_{y,1} (1 + e^{-i\phi} - |p_{y,1}|^2 |1 + e^{i\phi}|^2) \\
&b_{tot} \bar{d}_{tot} = q_{x,1} (1 + e^{i\phi}) \bar{q}_{y,1} (1 - |p_{x,1}|^2 (1 + e^{-i\phi})) = q_{x,1} \bar{q}_{y,1} (1 + e^{i\phi} - |p_{x,1}|^2 |1 + e^{i\phi}|^2).
\end{aligned} \tag{4.61}$$

Related to equation (4.53c):

$$\begin{aligned}
a_{tot} p_{x,1} &= q_{x,1} (1 - |p_{y,1}|^2 (1 + e^{i\phi})) \\
b_{tot} p_{y,1} &= p_{y,1} \bar{p}_{y,1} q_{x,1} (1 + e^{i\phi}).
\end{aligned} \tag{4.62}$$

4.7.2 Proof of the parameters of an ideal single-mode metasurfaces from its Jones matrix

In equations (4.7) and (4.8), the parameters θ, ϕ_1 and ϕ_2 are expressed in terms of the elements of the matrix $\hat{T}_{tot}$. To prove those equations, equation (4.1) is transformed into

$$\hat{R}_{-\theta}\hat{T}_{tot}\hat{R}_{\theta} = \begin{pmatrix} e^{i\varphi_1} & 0 \\ 0 & e^{i\varphi_2} \end{pmatrix}, \tag{4.63}$$

which gives the following set of equations:

$$2b_{tot}\cos(2\theta) + (d_{tot} - a_{tot})\sin(2\theta) = 0 \tag{4.64a}$$

$$a_{tot} + d_{tot} + (a_{tot} - d_{tot})\cos(2\theta) + 2b_{tot}\sin(2\theta) = 2e^{i\varphi_1} \tag{4.64b}$$

$$a_{tot} + d_{tot} - [(a_{tot} - d_{tot})\cos(2\theta) + 2b_{tot}\sin(2\theta)] = 2e^{i\varphi_2}. \tag{4.64c}$$

If b_{tot} is zero, the expressions in (4.8) are directly obtained for $\theta = 0$. Another aspect of equations (4.64) is that, if θ, φ_1 and φ_2 are solutions of equations (4.64), $\theta + \pi/2$ is also a solution and the value φ_1 and φ_2 are swapped.

To get the parameter θ when b_{tot} is different than zero, the equation (4.64a) is multiplied by $\bar{b}_0$ and the property (4.40c) is applied, resulting in the equation

$$|b_{tot}|^2\cos(2\theta) - \mathrm{Re}\{a_{tot}\bar{b}_{tot}\}\sin(2\theta) = 0. \tag{4.65}$$

Then, $\cos(2\theta)$ and $\sin(2\theta)$ are chosen as

$$\begin{aligned} &\cos(2\theta) = \frac{|\mathrm{Re}\{a_{tot}\bar{b}_{tot}\}|}{r_1} \qquad \sin(2\theta) = \frac{s_1|b_{tot}|^2}{r_1} \\ &r_1 = \sqrt{\mathrm{Re}\{a_{tot}\bar{b}_{tot}\}^2 + |b_{tot}|^4} \\ &s_1 = \begin{cases} 1 & \text{if } \mathrm{Re}\{a_{tot}\bar{b}_{tot}\} \geq 0 \\ -1 & \text{otherwise} \end{cases}. \end{aligned} \tag{4.66}$$

.

With this choice, $\cos(2\theta)$ goes to one when b_{tot} goes to zero, meaning that θ converges to zero. Moreover, a numerical error on b_{tot} when b_{tot} is nearly zero has only a negligible impact on θ.

To get the expressions of φ_1 and φ_2, equations (4.64b) and (4.64c) are multiplied by $\bar{b}_{tot}$, the property (4.40c) is applied, $\cos(2\theta)$ and $\sin(2\theta)$ are replaced by their respective expressions shown in (4.66) and the property

$$|\operatorname{Re}\{a_{tot}\bar{b}_{tot}\}| = s_1 \operatorname{Re}\{a_{tot}\bar{b}_{tot}\} \tag{4.67}$$

is applied, giving

$$\begin{aligned} i\operatorname{Im}\{a_{tot}\bar{b}_{tot}\} + \frac{s_1}{r_1}\left[\operatorname{Re}\{a_{tot}\bar{b}_{tot}\}^2 + |b_{tot}|^4\right] &= \bar{b}_{tot}e^{i\varphi_1} \\ i\operatorname{Im}\{a_{tot}\bar{b}_{tot}\} - \frac{s_1}{r_1}\left[\operatorname{Re}\{a_{tot}\bar{b}_{tot}\}^2 + |b_{tot}|^4\right] &= \bar{b}_{tot}e^{i\varphi_2}. \end{aligned} \tag{4.68}$$

Recognizing the coefficient r, equations (4.68) are reduced to

$$i b_{tot}\left[\operatorname{Im}\{a_{tot}\bar{b}_{tot}\} - i s_1 r_1\right] = |b_{tot}|^2 e^{i\varphi_1} \tag{4.69a}$$

$$i b_{tot}\left[\operatorname{Im}\{a_{tot}\bar{b}_{tot}\} + i s_1 r_1\right] = |b_{tot}|^2 e^{i\varphi_2} \tag{4.69b}$$

As b_{tot} goes to zero, φ_1 and φ_2 converge to the phase of respectively a_{tot} and d_{tot}, but this is true only if the property (4.40c) still holds. This may not be true when b_{tot} is nearly zero due to numerical errors, meaning that the phase of d_{tot} cannot be obtained through the phase of a_{tot} and b_{tot}. Therefore, using the property (4.40c), equation (4.69b) is replaced by

$$\begin{aligned} & i b_{tot}\left[\operatorname{Im}\{d_{tot}\bar{b}_{tot}\} + i s_1 r_2\right] = |b_{tot}|^2 e^{i\varphi_2} \\ & r_2 = \sqrt{\operatorname{Re}\{d_{tot}\bar{b}_{tot}\}^2 + |b_{tot}|^4} \end{aligned} \tag{4.70}$$

The expression of φ_1 and φ_2 in (4.7) are directly obtained by getting the phase on both sides of equations (4.69a) and (4.70). Moreover, without using the property (4.40c), equations (4.69a) and (4.70) converge to

$$\begin{aligned} b_{tot}\left[\operatorname{Re}\{a_{tot}\bar{b}_{tot}\} + i\operatorname{Im}\{a_{tot}\bar{b}_{tot}\}\right] &= |b_{tot}|^2 e^{i\varphi_1} \\ b_{tot}\left[\operatorname{Re}\{d_{tot}\bar{b}_{tot}\} + i\operatorname{Im}\{d_{tot}\bar{b}_{tot}\}\right] &= |b_{tot}|^2 e^{i\varphi_2} \end{aligned} \tag{4.71}$$

as b_{tot} goes to zero and the term $|b_{tot}|^2$ becomes negligible compared to $\operatorname{Re}\{a_{tot}\bar{b}_{tot}\}$ and $\operatorname{Re}\{d_{tot}\bar{b}_{tot}\}$. By dividing both sides of equations (4.71) by the term $|b_{tot}|^2$, equation (4.8) is obtained.

4.7.3 Proof of the design of a pair of ideal single-mode metasurfaces

In section 4.3, the solution is given for the system described by

$$\hat{T}_2 \hat{T}_1 = \hat{T}_{tot}, \tag{4.72}$$

where the symmetric unitary matrices $\hat{T}_1$ and $\hat{T}_2$, described in (4.20), are the unknowns and the unitary matrix $\hat{T}_{tot}$, described in (4.21), is given. Since the inverse of an unitary matrix is its complex conjugate, $\hat{T}_1$ is given by

$$\hat{T}_1 = \begin{pmatrix} a_1 & b_1 \\ b_1 & d_1 \end{pmatrix} = \hat{T}_2^H \hat{T}_{tot} = \begin{pmatrix} a_{tot}\bar{a}_2 + c_{tot}\bar{b}_2 & b_{tot}\bar{a}_2 + d_{tot}\bar{b}_2 \\ a_{tot}\bar{b}_2 + c_{tot}\bar{d}_2 & b_{tot}\bar{b}_2 + d_{tot}\bar{d}_2 \end{pmatrix} \tag{4.73}$$

and equations (4.23) are obtained.

Since $\hat{T}_1$ must be symmetric, the off-diagonal elements of the matrix $\hat{T}_1$ are zero, leading to

$$b_{tot}\bar{a}_2 - c_{tot}\bar{d}_2 = (a_{tot} - d_{tot})\bar{b}_2. \tag{4.74}$$

After multiplying (4.74) by b_2, assuming that b_2 is different than zero, and applying the property (4.40c), equation (4.74) becomes

$$c_{tot}a_2\bar{b}_2 + b_{tot}\bar{a}_2 b_2 = (a_{tot} - d_{tot})|b_2|^2. \tag{4.75}$$

For $b_{tot} = c_{tot} = 0$, equation (4.75) degenerates and this case is treated later. For a given $b_2 = re^{i\phi}$, the expression a_2 that satisfies equation (4.75), can be expressed as

$$a_2 = (z_s s' + z_o) r e^{i\phi}, \tag{4.76}$$

where s' can be any real number. It means that z_s and z_o satisfy the equations.

$$c_{tot}z_s + b_{tot}\bar{z}_s = 0 \tag{4.77a}$$

$$c_{tot}z_o + b_{tot}\bar{z}_o = a_{tot} - d_{tot}. \tag{4.77b}$$

Different expressions for z_s and z_o are proposed. For z_s, it is important that the chosen expression is not zero for any b_{tot} and c_{tot} because the coefficient s' is required in order to satisfy (4.40b) later on. The proposed expression of z_s, which is the same as (4.24), is

$$z_s = \begin{cases} i(b_{tot} + \bar{c}_{tot}) & \text{if } \mathrm{Re}\left\{\frac{b_{tot}}{\bar{c}_{tot}}\right\} \geq 0 \\ b_{tot} - \bar{c}_{tot} & \text{otherwise} \end{cases} \tag{4.78}$$

The subdomains are chosen such that z_s is never zero since the term $i(b_{tot}+\bar{c}_{tot})$ is zero when $\mathrm{Re}\{b_{tot}/\bar{c}_{tot}\}=-1$ and the term $b_{tot}-\bar{c}_{tot}$ is zero when $\mathrm{Re}\{b_{tot}/\bar{c}_{tot}\}=1$. As a reminder, b_{tot} and c_{tot} are different than zero. However, any expressions of z_s that are a linear composition with real coefficients of the expressions given in (4.78), are also valid. Due to equations (4.40a) and (4.40c), b_{tot} and c_{tot} have the same norm and the expressions of z_s in (4.78) satisfies equation (4.77a). For z_o, the proposed expression, which is the same as (4.24), is

$$z_o=\begin{cases}\frac{a_{tot}-d_{tot}}{b_{tot}+c_{tot}} & \text{if }\mathrm{Re}\left\{\frac{b_{tot}}{c_{tot}}\right\}\geq 0\\ \frac{a_{tot}-d_{tot}}{c_{tot}-b_{tot}} & \text{otherwise}\end{cases} \tag{4.79}$$

The subdomains are chosen such that z_s does not go toward infinity since the term $b_{tot}+c_{tot}$ is zero when $\mathrm{Re}\{b_{tot}/c_{tot}\}=-1$ and the term $c_{tot}-b_{tot}$ is zero when $\mathrm{Re}\{b_{tot}/c_{tot}\}=1$. z_0 is either purely real or purely imaginary since

$$\begin{aligned}\frac{a_{tot}-d_{tot}}{b_{tot}+c_{tot}}&=\frac{(a_{tot}-d_{tot})(\bar{b}_{tot}+\bar{c}_{tot})}{(b_{tot}+c_{tot})(\bar{b}_{tot}+\bar{c}_{tot})}=\frac{(\bar{a}_{tot}-\bar{d}_{tot})(b_{tot}+c_{tot})}{(b_{tot}+c_{tot})(\bar{b}_{tot}+\bar{c}_{tot})}=\left(\frac{a_{tot}-d_{tot}}{b_{tot}+c_{tot}}\right)^*\\ \frac{a_{tot}-d_{tot}}{c_{tot}-b_{tot}}&=\frac{(a_{tot}-d_{tot})(\bar{c}_{tot}-\bar{b}_{tot})}{(c_{tot}-b_{tot})(\bar{c}_{tot}-\bar{b}_{tot}}=-\frac{(\bar{a}_{tot}-\bar{d}_{tot})(c_{tot}-b_{tot})}{(c_{tot}-b_{tot})(\bar{c}_{tot}-\bar{b}_{tot})}=-\left(\frac{a_{tot}-d_{tot}}{c_{tot}-b_{tot}}\right)^*.\end{aligned} \tag{4.80}$$

In equation (4.80), the property (4.40c) is applied. Due to z_0 being either purely real or purely imaginary, it directly follows that z_0 satisfies equation (4.77b).

The coefficient s' has to be chosen such that a_2, given in (4.76), and b_2 satisfy equation (4.40b), meaning that

$$|z_s s'+z_o|^2 r^2+r^2=1. \tag{4.81}$$

The coefficient s' is real and satisfies the quadratic equation

$$|z_s|^2 r^2 s'^2+2c_b r^2 s'+(|z_o|^2+1)r^2-1=0, \tag{4.82}$$

where c_b is given by

$$c_b=\frac{1}{2}(z_s\bar{z}_o+\bar{z}_s z_o)=\begin{cases}z_o\,\mathrm{Re}\{z_s\} & \text{if }\mathrm{Re}\left\{\frac{b_{tot}}{c_{tot}}\right\}\geq 0\\ -i z_o\,\mathrm{Im}\{z_s\} & \text{otherwise}\end{cases} \tag{4.83}$$

The expression in (4.83) is slightly different than (4.25). The expression in (4.25) has the advantage that it ensures that c_b is purely real as it should be. The solution of the quadratic equation (4.82) is

$$\begin{aligned}s' &= \frac{-c_b r^2 \pm \sqrt{c_b^2 r^4 + |z_s|^2 r^2 (1 - (|z_o|^2+1) r^2)}}{|z_s|^2 r^2} \\ &= \frac{1}{r} \frac{-c_b r \pm \sqrt{c_b^2 r^2 + |z_s|^2 (1 - (|z_o|^2+1) r^2)}}{|z_s|^2} = \frac{s}{r}.\end{aligned} \tag{4.84}$$

In section 4.3, the coefficient s is chosen instead of the coefficient s' because s' diverges as r goes to zero. By using s instead of s', the expression in (4.76) becomes the expression of a_2 in (4.22).

As mentioned before, s and s' must be real, meaning that the term below the square root in equation (4.84), called Δ, has to be zero or positive, which limits the possible value that the coefficient r can have. Since r is the norm of the coefficient b_2, the minimum value of r is zero. Δ is linear with respect to r^2 and, for $r = 0$, Δ is always positive. Finally, Δ has to be negative when r is greater than one since, in this case, equation (4.81) cannot be satisfied, meaning that the assumption that s' is purely real, is wrong. Hence, it exists a maximum value for the coefficient r, called r_{max}, which is in the interval $[0,1]$. r_{max} is obtained by setting Δ to zero:

$$c_b^2 r_{max}^2 + |z_s|^2 - (|z_o|^2 + 1)|z_s|^2 r_{max}^2 = 0 \tag{4.85}$$

Hence, r_{max} is

$$r_{max} = \frac{|z_s|^2}{|z_s|^2 + |z_o|^2 |z_s|^2 - c_b^2}. \tag{4.86}$$

Since c_b^2 is given by

$$c_b^2 = \begin{cases} |z_o|^2 \operatorname{Re}\{z_s\} & \text{if } \operatorname{Re}\left\{\frac{b_{tot}}{c_{tot}}\right\} \geq 0 \\ |z_o|^2 \operatorname{Im}\{z_s\} & \text{otherwise} \end{cases} \tag{4.87}$$

the expression of r_{max} given in (4.24) is obtained.

In order to get d_2, the property (4.40c) is applied:

$$\bar{a}_2 b_2 + d_2 \bar{b}_2 = (\bar{z}_s s + \bar{z}_o r) e^{-i\phi} r e^{i\phi} + d_2 r e^{-i\phi} = 0. \tag{4.88}$$

Hence, if b_2 is different than zero, the expression of d_2 in (4.22) is obtained.

In the previous part of this section, equation (4.75) is derived for b_2 different than zero. However, the obtained solution is still valid in this special case. The case $b_{tot} = c_{tot} = 0$ is treated separately.

If b_2 is zero, a_2 and d_2 have to satisfy equation (4.74) instead of equation (4.75), and the property (4.40c) for the matrix $\hat{T}_2$ is trivially true. Assuming that the expressions in (4.22) are still valid for $r = 0$, they become

$$s = \frac{p}{|z_s|} \qquad a_2 = s z_s e^{i\phi} = p\frac{z_s}{|z_s|}e^{i\phi} \qquad d_2 = -p\frac{\bar{z}_s}{|z_s|}e^{i\phi}, \tag{4.89}$$

where p is ± 1.

Using those expressions, equation (4.74) becomes

$$\frac{p}{|z_s|}\left(b_{tot}\bar{z}_s + c_{tot}z_s\right)e^{-i\phi} = 0, \tag{4.90}$$

which is equivalent to equation (4.77a), meaning that the expressions in (4.22) are still valid when b_2 is zero. As mentioned in section 4.3, the solution with $p = 1$ and $\phi = \phi_0$ for any ϕ_0 real is the same solution with $p = -1$ and $\phi = \phi_0 + \pi$.

For the case $b_{tot} = c_{tot} = 0$ and $a_{tot} = d_{tot}$, equation (4.74) is trivially true, meaning that only the properties (4.40) for the matrix $\hat{T}_2$ need to be satisfied. Due to properties (4.40a) and (4.40b), a_2, b_2 and d_2 are given by

$$a_2 = \sqrt{1-r^2}e^{i\phi_1} \qquad b_2 = re^{i\phi_b} \qquad d_2 = -\sqrt{1-r^2}e^{i\phi_2}, \tag{4.91}$$

and, from property (4.40c), ϕ_b is

$$\phi_b = \frac{\phi_1 + \phi_2}{2} + mpi, \tag{4.92}$$

where m is a real integer. Hence, equation (4.26) is proved.

For the case $b_{tot} = c_{tot} = 0$ and $a_{tot} = d_{tot}$, equation (4.74) is satisfied only if b_2 is zero, leading to the solution given in (4.26) with $r = 0$.

5 Multi-mode metasurface, resonant metasurface and self-coupling mode

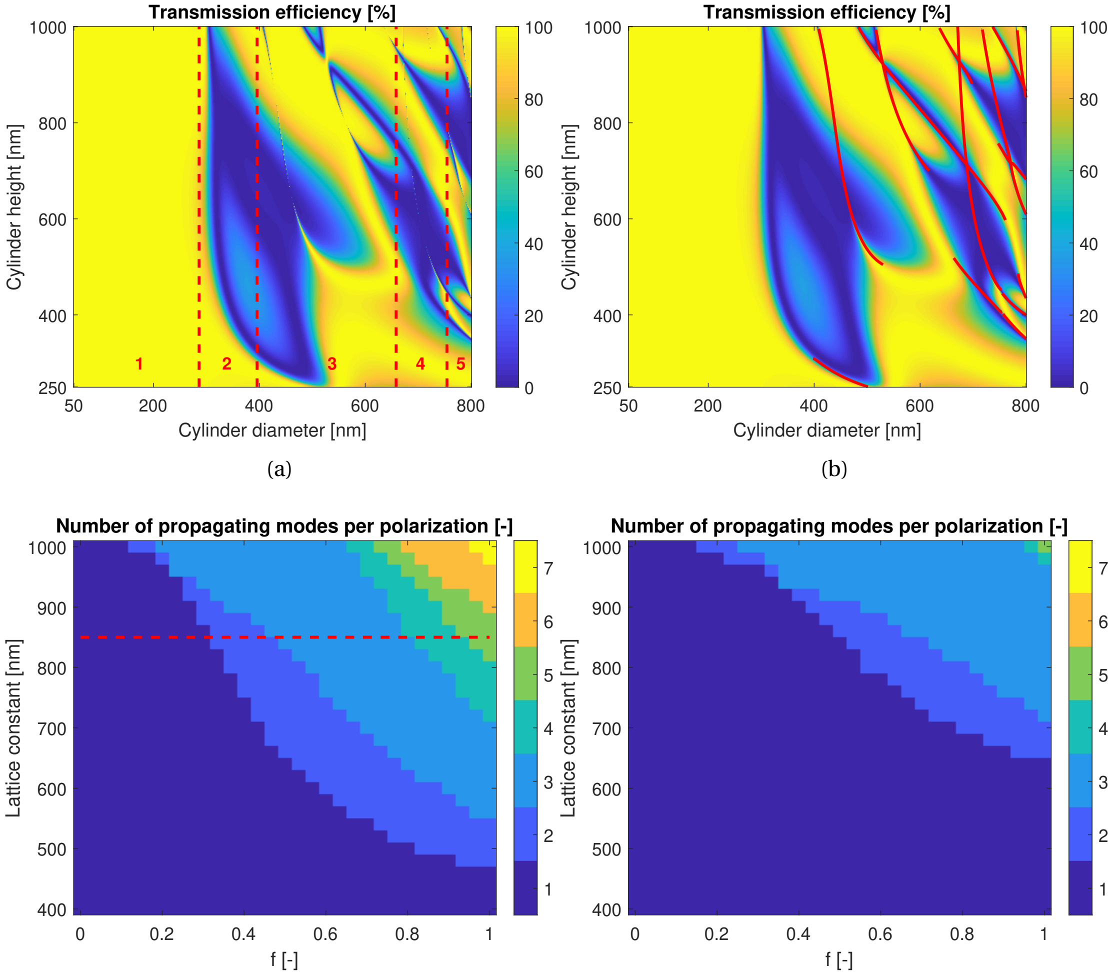
5.1 Introduction

Multi-mode metasurfaces are zeroth-order gratings which have multiple eigen-modes per polarization propagating inside the metasurface. In the single-mode metasurfaces presented in chapter 4, the two propagating eigen-modes, one per polarization, are independent, meaning that single-mode metasurface can be seen as two independent Fabry-Pérot cavities, where the reflection of the eigen-modes at the interfaces is given by a single number. For a multi-mode metasurface, since the propagating eigen-modes couple between each other at both interfaces of the metasurface, it is required to represent the reflection of the propagating eigen-modes at both interfaces as matrices, leading to a more complex system.

Due to this complexity, many interesting phenomena can be observed such as resonances, including Fano resonances [142], and the great diversity of responses makes multi-mode metasurfaces a promising platform. In the literature, they have been used as color filters [46–50], as holograms where the aspect ratio of the structures is much lower than what can be expected from a single-mode metasurface [37, 44, 45], as molecule sensors [53] and as generalized Hartmann-Shack arrays [52]. The main drawbacks are that they are more difficult to design and, for multi-mode metasurfaces composed of cylinders, eigen-modes are not confined within the cylinders as it is the case in single-mode metasurface, which can lead to unexpected results if the cylinders' dimensions vary across the metasurface.

Figure 5.1a is the transmission efficiency of metasurfaces composed of silicon cylinders embedded in glass for a wavelength of 1477 nm. The lattice constant is 850 nm. Depending on the diameter of the cylinders, the metasurface can be either single-mode or multi-mode. In the single-mode region, which is for a diameter up to 286 nm, the transmission efficiency is nearly constant and is close to 100% for the different cylinder heights. In the multi-mode region, the transmission efficiency varies strongly, and, as the number of propagating modes increases, this variation gains in complexity.

The red lines in fig. 5.1b are resonances and high-Q resonances occur only in multi-mode



(a) (b)

(c) (d)

Figure 5.1 – a) Transmission efficiency of a metasurface composed of silicon cylinders embedded in glass for different cylinder diameters and heights. The lattice constant is 850 nm and the wavelength is 1477 nm. The red number at the bottom of the figure is the number of propagating modes per polarization. b) Same as fig. 5.1b except that the resonances are shown by red lines. c) Number of propagating modes per polarization for the same structure as in fig. 5.1b except that the lattice constant ranges from 400 nm to 1000 nm. The red dashed line indicates the set of metasurfaces whose transmission efficiency is shown in fig. 5.1b. d) Same as fig. 5.1c except that the cylinders are made of TiO_2 instead of silicon.

metasurfaces. Since the eigen-modes present in a multi-mode metasurface couple at the interfaces to only one transmitted and one reflected plane wave, it is possible that the weights of the eigen-modes are as such that the contribution of the eigen-modes on the transmitted and reflected plane waves destructively interferes. Hence, the weights of the eigen-modes have to be very high in order to balance the power going out of the metasurface through the transmitted and reflected plane waves and coming in from the incident plane wave. In other words, the optical energy is stored inside the metasurface, meaning that the metasurface is on resonance.

In order to get a multi-mode metasurface, the refractive index of at least one of the material that composes the metasurface has to be larger than the refractive index of the substrate and the superstrate. Hence, metasurfaces composed of silicon on a glass substrate or embedded in glass are typical candidate for a multi-mode metasurface. To illustrate this condition, the number of propagating modes for metasurfaces composed of silicon ($n = 3.48$) cylinders embedded in glass ($n = 1.44$) with different lattice constants and cylinder diameters is shown in fig. 5.1c. If the refractive index of the substrate and the superstrate increases, the maximum lattice constant which is required to get a zeroth order grating, decreases, meaning that there are less options for a multi-mode metasurface. In the extreme case, if the substrate or the superstrate is silicon, the lattice constant has to be smaller than 424 nm in order to have a zeroth-order grating and, from fig. 5.1c, multi-mode metasurface does not exist for such lattice constants. Figure 5.1d is for the same metasurface except that the material that composes the cylinders is TiO_2, which is a lower refractive index material ($n = 2.46$). As expected, the area where multiple propagating modes are present in the metasurface is significantly smaller than in fig. 5.1c.

The parameter f used for the x-axis in figs. 5.1c and 5.1d is related to the ratio between the cylinder diameter d and the lattice constant l and has been chosen in a way that the smallest feature is above d_{min}. Hence, f is given by

$$f = \frac{d - d_{min}}{l - 2d_{min}}, \tag{5.1}$$

where d_{min} is 50 nm for figs. 5.1c and 5.1d.

It is to our knowledge not possible to give design techniques for multi-mode metasurfaces that are comparable to the ones proposed in chapter 4. However, it is possible to clearly differentiate resonant and non-resonant effects through the use of self-coupling modes. A typical non-resonant effect is the large dark blue area in figs. 5.1a and 5.1b. The concept of self-coupled mode, abbreviated to SCM, has already been used to retrieve the quasi-normal modes of a system [124] and to estimate the Q-factor of a multi-mode cavity [123], but it was not named. In this work, the concept of self-coupling mode is developed and used at its full potential for multi-mode metasurfaces. This concept allows a systematic characterization of resonances, facilitates the search for resonances and considerably reduces the number of simulations required to accurately compute the response of a high-Q resonant metasurface as a function of parameters such as the wavelength, the angle of incidence and the metasurface's thickness.

Section 5.2 explains the concept of self-coupled mode and how to use it. The definition and the computation of the self-coupling modes are given in section 5.2.1 and the equations that decompose the transmitted and reflected fields into the contributions of the different self-coupling modes are given in section 5.2.2.

The different advantages of this concept are illustrated by applying it on four different metasurfaces. In section 5.3.1, the two resonances present in the Huygens' metasurface are characterized. The Huygens' metasurface has been introduced by M. Decker [43] and used for the design of a phase-only transmission function in [37, 44]. It is not easy to separate the resonant response from the total response using standard simulation techniques like the Finite-Element Method (FEM) or the Finite Difference Time Domain method (FDTD). The use of the self-coupling modes, due to their relationship with the quasi-normal modes [124], allows to easily extract the resonant response, giving its impact on the transmission and reflection of the metasurface, but also the fields inside the metasurface related to the resonance. Moreover, if the metasurface has multiple resonances intertwined such as in the Huygens' metasurface, the concept of self-coupling mode allows to separate them without any ambiguity.

The second metasurface acts as a narrowband filter. It combines both a non-resonant effect, which is the broadband mirror-like response of the metasurface, and a resonant effect. As shown in section 5.3.2, those two effects can be identified from the simulation of a single metasurface, even if this metasurface is outside the resonance. This example is also used to give an interpolation scheme given in section 5.2.2, which is based on the decomposition of the transmitted and reflected fields. The accuracy of this interpolation scheme is given in the case where the simulated metasurfaces are all outside the resonance.

The third metasurface is an array of AlAs cylinders that might be used as a laser and it is presented in section 5.3.3. This metasurface has been designed such that the resonance, which is sustained by the GaAs quantum wells, emits mainly outside instead than in the substrate. Moreover, the angular spectrum of the resonance has an interesting star-like shape and obtaining such angular spectrum is very computationally expensive with the level of details given in this work without the use of the self-coupling modes.

The last metasurface is composed of silicon cylinders with an obround cross-section on a glass substrate immerged in water. This metasurface can have a very high-Q resonance. This resonance has been designed such that its angular spectrum is strongly asymmetric, meaning that the spatial extent of this resonance is also strongly asymmetric. Section 5.3.4 shows where this asymmetry comes from and also how the concept of self-coupling mode combined to the Fourier modal method given in chapter 3 facilitates greatly the design of such metasurface.

In this work, the focus is on metasurfaces composed of cylinders, but metasurfaces composed of holes can also be used. It should be slightly easier to design a multi-mode metasurface composed of holes since the filling fraction of the high refractive index material can be higher leading to more candidates in the design process. The self-coupling modes can be used equally well in both cases.

5.2 Self-coupling mode

5.2.1 Definition and computation of the self-coupling modes

Binary metasurfaces can be seen as a layer delimited by two interfaces and, from the Fourier modal method presented in chapter 3, the eigen-modes that propagate inside this layer are known. The relationships between the incident, transmitted and reflected plane waves, and the eigen-modes are given by

$$\begin{array}{lll} \hat{T}_3\vec{a}_3 = \vec{t} & \hat{T}_1\vec{p} + \hat{R}_2\vec{b}_2 = \vec{a}_2 & \hat{\Gamma}\vec{a}_2 = \vec{a}_3 \\ \hat{R}_1\vec{p} + \hat{T}_2\vec{b}_2 = \vec{r} & \hat{R}_3\vec{a}_3 = \vec{b}_3 & \hat{\Gamma}\vec{b}_2 = \vec{b}_3, \end{array} \tag{5.2}$$

where $\vec{p}$, $\vec{r}$ and $\vec{t}$ are the weights of the incident, reflected and transmitted plane waves, $\vec{a}_2$ and $\vec{b}_2$ are the weights of, respectively, the forward and backward-propagating eigen-modes just after the first interface, and $\vec{a}_3$ and $\vec{b}_3$ are the weights just before the second interface. $\hat{R}_m$ and $\hat{T}_m$ are the coupling matrices and $\hat{\Gamma}$ is the propagation operator. The equations in (5.2) are illustrated in fig. 5.2a.

If $\vec{p}$, $\vec{r}$ and $\vec{t}$ are related by the matrices $\hat{T}'_1$ and $\hat{R}'_1$ such that

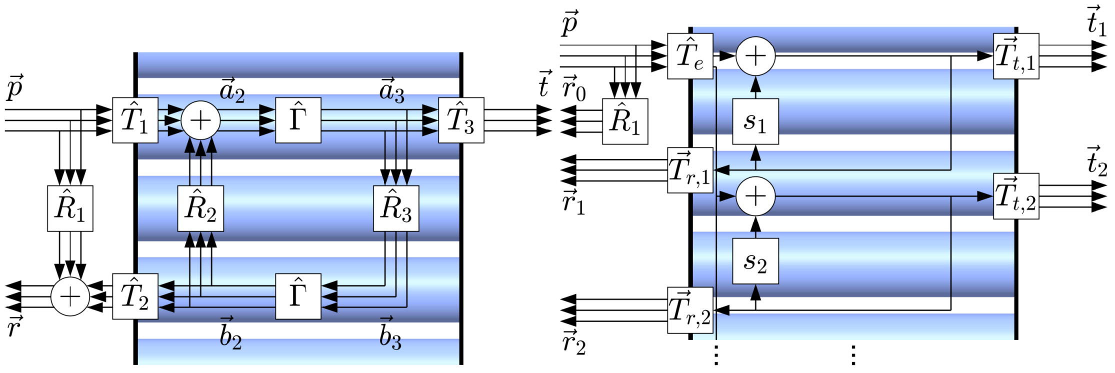
$$\hat{T}'_1\vec{p} = \vec{t} \qquad \hat{R}'_1\vec{p} = \vec{r}, \tag{5.3}$$

the matrices $\hat{T}'_1$ and $\hat{R}'_1$ are given by

$$\begin{aligned} \hat{T}'_1 &= \hat{T}_3\hat{\Gamma}(\hat{I}-\hat{M})^{-1}\hat{T}_1 \\ \hat{R}'_1 &= \hat{R}_1 + \hat{T}_2\hat{\Gamma}\hat{R}_3\hat{\Gamma}(\hat{I}-\hat{M})^{-1}\hat{T}_1 \\ \hat{M} &= \hat{R}_2\hat{\Gamma}\hat{R}_3\hat{\Gamma}, \end{aligned} \tag{5.4}$$

where $\hat{M}$ is known as the roundtrip matrix [124] since it describes the loop inside the metasurface shown in fig. 5.2a. Those equations are derived from equation (3.40) and proved in section 3.5.

If the matrix $\hat{I}-\hat{M}$ is nearly singular, a small variation of the metasurface dimensions leads to a large variation in the matrices $\hat{T}'_1$ and $\hat{R}'_1$, meaning that the metasurface is resonant. Hence, by looking at the solution of the eigen-value equation



(a)

(b)

Figure 5.2 – a) A schema of a metasurface showing how the weights of the eigen-modes are related to each other by the coupling matrices $\hat{R}_m$ and $\hat{T}_m$, and the propagation operator $\hat{\Gamma}$. $\vec{p}$, $\vec{r}$ and $\vec{t}$ contain the weight of the incident, reflected and transmitted plane waves respectively. $\hat{a}_2$ and $\hat{b}_2$ are the weight of the eigen-modes just after the first interface and $\hat{a}_3$ and $\hat{b}_3$ are the weights just before the second interface. b) Same as fig. 5.2a except that the modes inside the metasurface are the self-coupling modes instead of the eigen-modes. $\vec{r}_m$ and $\vec{t}_m$ are the contribution of the self-coupling mode m on respectively the reflected and transmitted plane waves. s_m are the s-value of the self-coupling mode m.

$$\hat{M}\vec{v}_m = s_m \vec{v}_m, \tag{5.5}$$

the metasurface is resonant if one of the eigen-value s_m is close to one.

The physical meaning of equation (5.5) is that a mode composed of the forward-propagating eigen-modes whose weights are given by $\vec{v}_m$, couples only to itself after a roundtrip. Therefore, we name such modes as self-coupling modes. s_m indicates how it couples to itself and it is called the s-value of the self-coupling mode m. If s_m is negative real, the self-coupling mode destructively interferes with itself, meaning that the amplitude of the self-coupling mode is lower than how it is initially excited by the incident plane wave. If s_m is positive real, there is a build up of the fields inside the metasurface and, if s_m reaches one, the metasurface traps perfectly the light inside the metasurface. In general, if a self-coupling mode is excited by a factor 1 from the incident plane wave, the fields of the self-coupling mode are amplified by the factor

$$K = \frac{1}{|1 - s_m|}. \tag{5.6}$$

The convention chosen in this work is that a self-coupling mode is resonant if the absolute value of its s-value is above 0.5, meaning that the field amplification K is 2 when the s-value crosses the real axis. As shown later, when varying the wavelength, the path of s-value in the complex plane is usually a circle.

By describing the modes inside the metasurface with the self-coupling modes, the diagram shown in fig. 5.2b is obtained, where the self-coupling modes are independent from each other. Hence, using the concept of self-coupling mode, the multi-mode metasurface can be seen as multiple Fabry-Pérot cavities instead of a single multi-mode cavity.

As shown in section 5.3, the s-value changes smoothly when varying the metasurface dimensions, the angle of incidence or the wavelength. Hence, the concept of self-coupling mode is not only used to see if a metasurface is resonant, but also if the metasurface is close to a resonance. Another information that can be obtained through the concept of self-coupling mode are the fields related to a resonance and the contribution to this resonance on the transmitted and reflected fields. The fields of a resonance is given by $\vec{v}_m$ and, as shown in section 5.3.1, allow to define the type of the resonance. The contribution of the resonances on the transmitted and reflected plane waves is more important because it can be used to get the Q-factor or the angular spectrum of the resonance. The angular spectrum of a resonance can be used to estimate the spatial extent of the resonance (section 5.3.4). The equations to obtained the contribution of self-coupling modes on the response of metasurfaces are developed in section 5.2.2.

5.2.2 Contribution of the self-coupling modes on the response of a metasurface

The concept of self-coupling mode is useful to analyze a multi-mode metasurface, but it can also be used to get the contribution of each self-coupling mode to the reflected and transmitted plane waves. Based on fig. 5.2b, the contribution of the self-coupling mode m depends of $\vec{T}_{t,m}$, $\vec{T}_{r,m}$, s_m and the m-th line of $\hat{T}_e$. In order to find those different elements, the roundtrip matrix $\hat{M}$ has to be written as

$$\hat{M} = \hat{V}\hat{S}\hat{V}^{-1}, \tag{5.7}$$

where each column of $\hat{V}$ describes a self-coupling mode and $\hat{S}$ is a diagonal matrix containing the s-values. This equation is equivalent to the eigen-value equation (5.5).

By combining equations (5.4) and (5.7), the matrices $\hat{T}_1'$ and $\hat{R}_1'$ can be expressed as

$$\begin{aligned} \hat{T}_1' &= \hat{T}_t(\hat{I}-\hat{S})^{-1}\hat{T}_e \\ \hat{R}_1' &= \hat{R}_1 + \hat{T}_r(\hat{I}-\hat{S})^{-1}\hat{T}_e, \end{aligned} \tag{5.8}$$

where $\hat{T}_e$, $\hat{T}_t$ and $\hat{T}_r$ are given by

$$\begin{aligned}\hat{T}_e &= \hat{V}^{-1}\hat{T}_1\\ \hat{T}_t &= \hat{T}_3\hat{\Gamma}\hat{V}\\ \hat{T}_r &= \hat{T}_2\hat{\Gamma}\hat{R}_3\hat{\Gamma}\hat{V}.\end{aligned} \tag{5.9}$$

Since $\hat{I}-\hat{S}$ is a diagonal matrix, equation (5.8) can be written as a sum such as each term depends on a single s-value:

$$\begin{aligned}\hat{T}_1' &= \sum_{m=1}^{M}\frac{1}{1-s_m}\vec{T}_{t,m}\hat{T}_{e,m:} &&= \sum_{m=1}^{M}\frac{1}{1-s_m}\hat{Q}_{t,m}\\ \hat{R}_1' &= \hat{R}_1+\sum_{m=1}^{M}\frac{1}{1-s_m}\vec{T}_{r,m}\hat{T}_{e,m:} &&= \sum_{m=1}^{M}\frac{1}{1-s_m}\hat{Q}_{r,m}.\end{aligned} \tag{5.10}$$

$\vec{T}_{t,m}$ and $\vec{T}_{r,m}$ are the m-th column of the matrices $\hat{T}_t$ and $\hat{T}_r$ respectively, and $\hat{T}_{e,m:}$ is the m-th line of the matrix $\hat{T}_e$. M is the number of self-coupling modes which is equal to the number of forward-propagating eigen-modes. The contribution of the self-coupling mode m on the transmitted and reflected plane waves is

$$\begin{aligned}\vec{t}_m &= \frac{1}{1-s_m}\hat{Q}_{t,m}\vec{p}\\ \vec{r}_0 &= \hat{R}_1\vec{p}\\ \vec{r}_m &= \frac{1}{1-s_m}\hat{Q}_{r,m}\vec{p},\end{aligned} \tag{5.11}$$

where m ranges from 1 to M.

Because the self-coupling modes are not orthogonal under the Poynting operation, it is less interesting to normalize them and to look at the quantities in the matrix $\hat{T}_e$. The vectors $\vec{T}_{r,m}$ and $\vec{T}_{t,m}$ are used in section 5.3.1 to get the fields outside the metasurface for the two resonances, and they are also used for the design of the metasurface in section 5.3.3, where the resonance emits mainly in one direction. The matrices $\hat{Q}_{t,m}$ and $\hat{Q}_{r,m}$ are 2-by-2 matrices for zeroth-order gratings if the evanescent plane waves are not considered and, with the s-value s_m, they give directly the contribution to the transmitted and reflected plane waves.

The convention used in this chapter is the same as in chapter 3, meaning that the first element in $\vec{p}$, $\vec{r}_m$ and $\vec{t}_m$ is the weight of the TM-polarized plane wave and the second element is the weight of the TE-polarized plane wave. At normal incidence, TM and TE-polarization are equivalent to x and y-polarization respectively.

If the coupling matrices $\hat{T}_m$ and $\hat{R}_m$, and the propagation constant of the eigen-modes vary smoothly while changing the dimensions of the metasurface or the properties of the incident plane wave, the s-value s_m and the elements of the matrices $\hat{Q}_{t,m}$ and $\hat{Q}_{r,m}$ vary also smoothly even if a resonance is present. Therefore, those quantities can be used for the interpolation of the response of a resonant metasurface. This interpolation techniques is presented and applied in section 5.3.2.

5.3 Application of the self-coupling mode on multi-mode metasurfaces

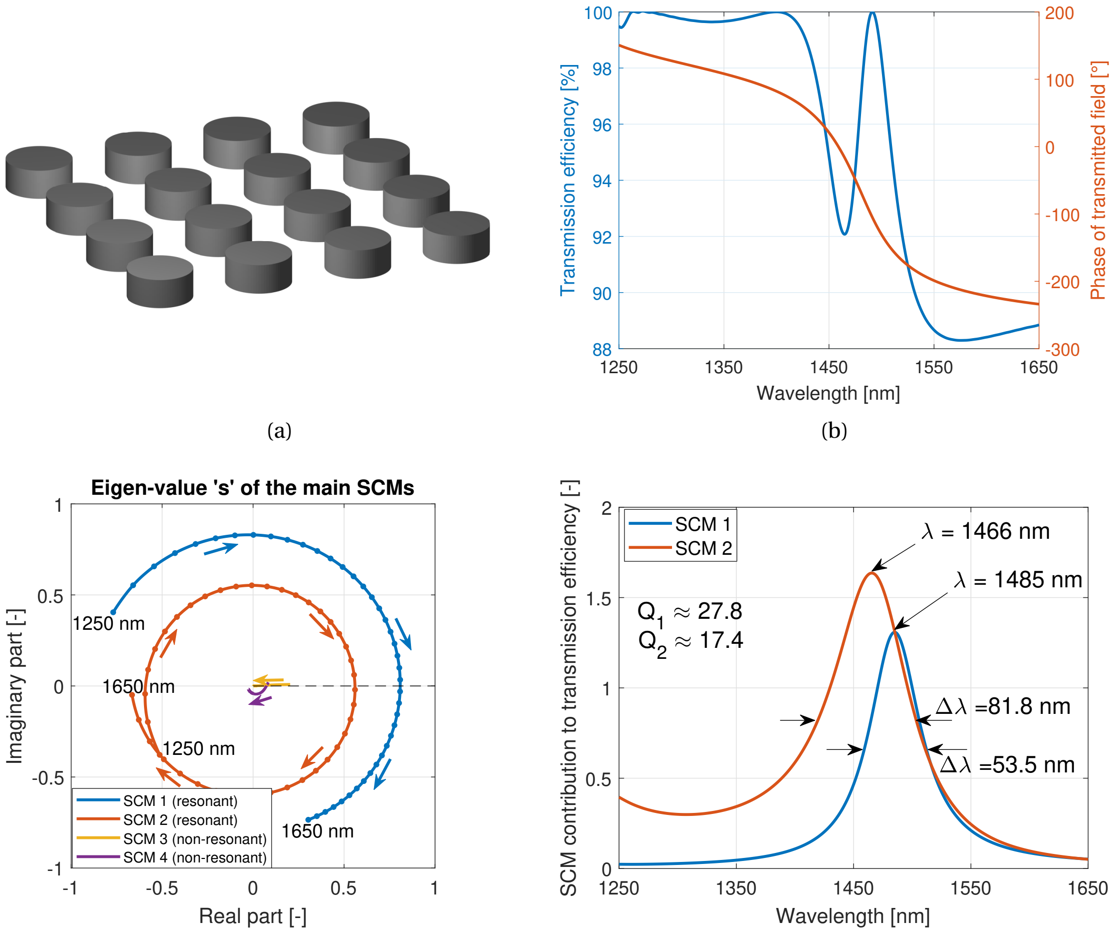
5.3.1 Magnetic and electric dipole resonances in Huygens' metasurface

A Huygens' metasurface is generally an array of silicon cylinders embedded in glass as shown in fig. 5.3a, which is designed such that the electric and magnetic dipole resonances overlap. The effect of those overlapping resonances is that the transmission efficiency can reach nearly 100% and the difference of the phase of the transmitted field for a wavelength before and after the resonances reaches 360°, making the Huygens' metasurface an interesting candidate for phase-only holograms. The Huygens' metasurface has been introduced by M. Decker [43] and used later in [37, 44, 45]. In order to identify and analyze the electric and magnetic dipoles separately, they represent each cylinder by an electric and a magnetic dipole and their polarizabilities are obtained using the coupled discrete dipole approach [143]. Hence, the fields related to the electric and magnetic dipole resonances are the fields produced by the corresponding dipole and the polarizability of those dipoles determines how strong the resonance is.

In this section, the objective is to also identify and analyze both resonances, but the concept of self-coupled mode is used instead. The main difference is that the concept of dipole is not used in order to separate the resonances from the overall response of the metasurface since the concept of self-coupled mode is based only on the eigen-modes present in the metasurface and the coupling matrices at the two interfaces. However, once the fields related to the resonances are obtained, it is possible to describe the fields in terms of dipoles or through other concept.

The Huygens' metasurface analyzed in this section is composed of cylinders with a diameter of 534 nm and a height of 243 nm. The lattice constant is 852 nm and an on-scale drawing is in fig. 5.3a. The transmission efficiency and the phase of the transmitted field in function of the wavelength is given in fig. 5.3b. In this case, the electric and magnetic dipole resonances are not exactly at the same wavelength leading to this small variation in the transmission efficiency. This variation may be enough to roughly estimate the wavelength and width of each resonances, but it will not be accurate.

In order to get an insight into the phenomena occurring inside a multi-mode metasurface, the first quantities that are analyzed are the s-value of the main self-coupling modes. The s-value of the other self-coupling modes are very close to zero and they do not give any valuable information. For the Huygens' metasurface, the s-values in function of the wavelength is



(a)

(b)

(c)

(d)

Figure 5.3 – (a) Scale drawing of the Huygens' metasurface, which is composed of an array of cylinders embedded in glass. The cylinder diameter and height are respectively 534 nm and 243 nm. The lattice constant is 852 nm. (b) Amplitude and phase of the transmitted field of the Huygens' metasurface. (c) Eigen-value s of the main self-coupling modes for different wavelengths. For every 10 nm in wavelength, the value of s is represented by a dot. (d) Contribution of the resonant self-coupling modes to the transmitted intensity. SCM 2 is more excited and has a lower Q-factor than SCM 1. The Q-factor is defined as the wavelength of the resonance divided by its full width at half maximum (FWHM).

given in fig. 5.3c. Four self-coupling modes are of interest. The two resonances can be clearly seen and they are represented by the Self-Coupling Modes (SCM) 1 and 2. SCM 3 and 4 are non-resonant. The s-value of all the self-coupling modes are turning clockwise around the origin of the complex plane. It is an expected behavior because, assuming that the propagation constant of the eigen-modes is positive, the propagation constants decrease as the wavelength increases and, by assuming that the variation of the eigen-mode decomposition of the self-coupling mode is small enough, the phase of the s-values also decreases. The trajectory of SCM 3 is a bit different than the others, but it is an usual behavior when a propagating eigen-mode becomes evanescent.

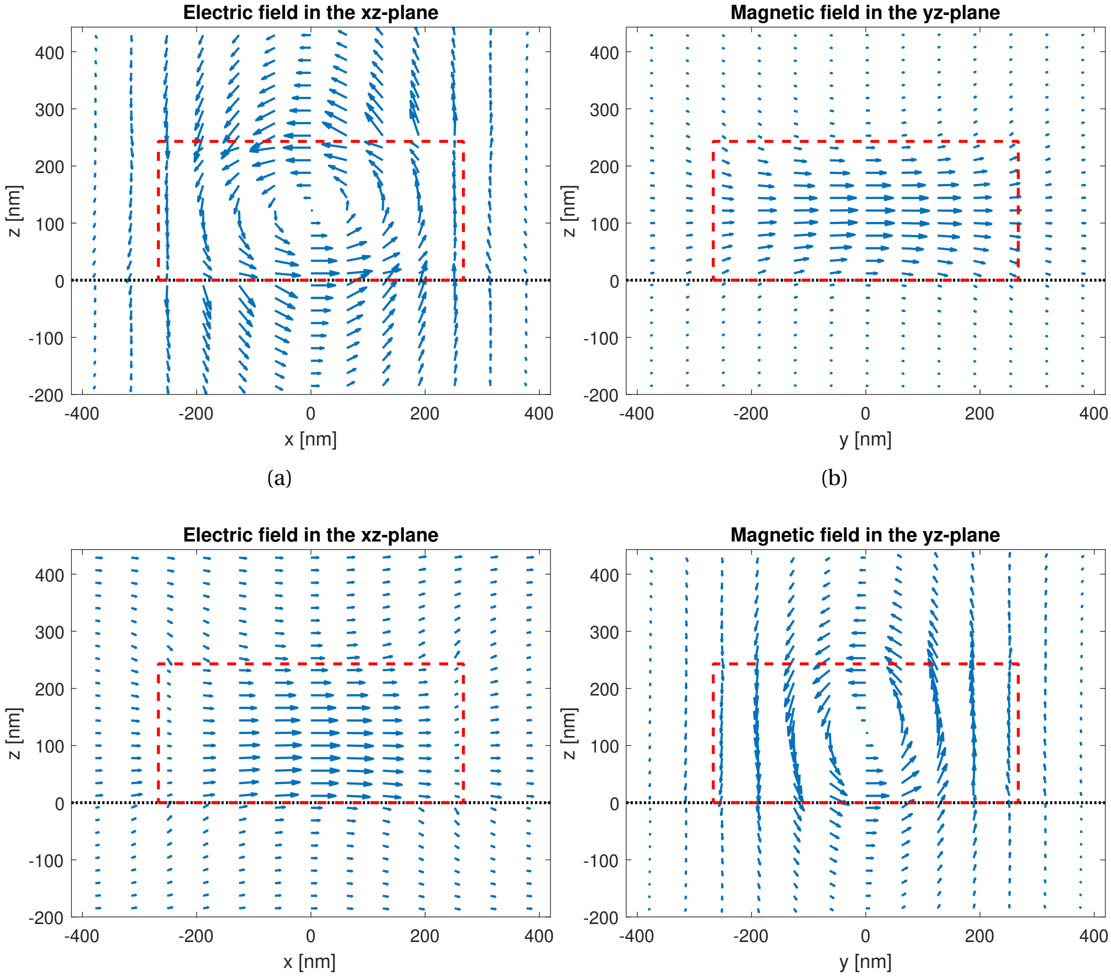
To get the wavelength and the width of the resonance, the contribution of the resonant self-coupling modes on the x-polarized transmitted plane wave is computed based on equation (5.11). Following the convention given in section 5.2.2, $\vec{p}$ is given by $(1,0)^T$. Since the substrate and the superstrate are made of the same material and the x-polarized transmitted plane wave is of interest, the contribution of the self-coupling mode m on the transmission efficiency is given by the amplitude squared of the first element in $\vec{t}_m$. The contribution of the resonant self-coupling modes SCM 1 and SCM 2 is given in fig. 5.3d. From this result, the wavelength, the width and the Q-factor of the resonances can be easily obtained. From fig. 5.3d, the Q-factor of the resonance related to SCM 1 and SCM 2 is around 27.8 and 17.4 respectively, meaning that the resonance related to SCM 1 is sharper.

From fig. 5.3d, it is not possible to associate the resonances to a type of resonance. In order to do that, the fields related to the resonant self-coupling modes need to be computed. The self-coupling modes are described by the vector $\vec{v}_m$, which is the weight of the forward-propagating eigen-modes just after the first interface. Hence, the Fourier coefficients of the electric and magnetic fields, $\widetilde{E}_{in}$ and $\widetilde{H}_{in}$, inside the metasurface are given by

$$\begin{aligned}\widetilde{E}_{in}(z) &= \hat{E}(\hat{\Gamma}(z)+\hat{\Gamma}(h-z)\hat{R}_3\hat{\Gamma}(h))\vec{v}_m\\ \widetilde{H}_{in}(z) &= \hat{H}(\hat{\Gamma}(z)-\hat{\Gamma}(h-z)\hat{R}_3\hat{\Gamma}(h))\vec{v}_m,\end{aligned} \tag{5.12}$$

where the n-th column of the matrices $\hat{E}$ and $\hat{H}$ contains the Fourier coefficients of respectively the electric and magnetic field of the n-th eigen-mode, and the n-th diagonal element of the diagonal matrix $\hat{\Gamma}(z)$ is $e^{i\gamma_n z}$. γ_n is the propagation constant of the n-th eigen-mode and the first and second interfaces are located at respectively $z=0$ and $z=h$, meaning that $\hat{\Gamma}(h)$ is equal to the propagation operator used in section 5.2.

Usually, the fields inside the metasurface are sufficient to determine the type of the resonances, but, for a complete representation, the vectors $\vec{T}_{t,m}$ and $\vec{T}_{r,m}$, that contain the weights of the transmitted and reflected plane waves, are needed, and they are obtained from equation (5.9). In this case, the evanescent plane waves should be considered. The results for SCM 1 and 2 are given in fig. 5.4: From figs. 5.4a and 5.4b, the resonance related to SCM 1 is a magnetic dipole resonance and, from figs. 5.4c and 5.4d, the resonance related to SCM 2 is an electric dipole



(a) (b)

(c) (d)

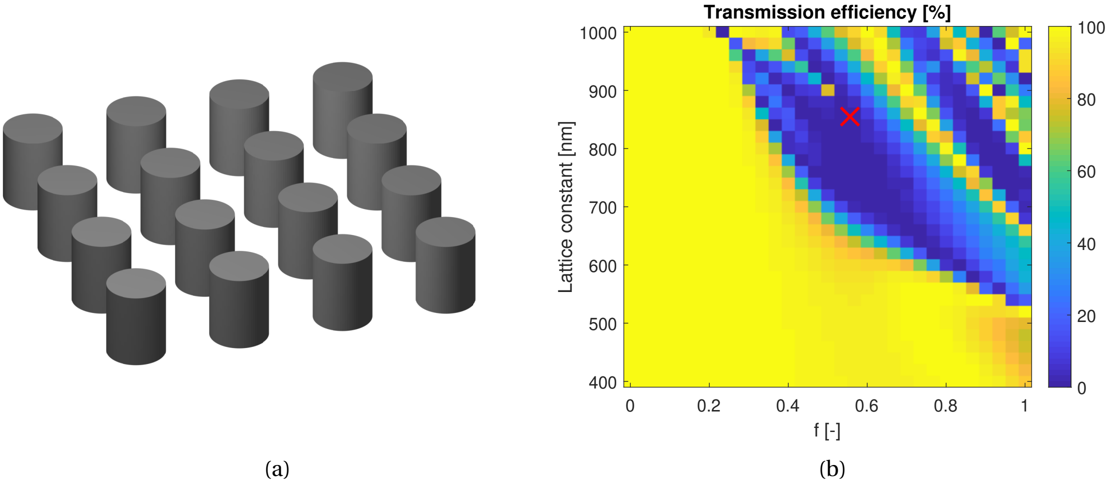
Figure 5.4 – a) Electric field of SCM 1, which is related to the magnetic dipole resonance, in the xz-plane that goes through the center of a cylinder. The red dashed line represents this cylinder. b) Magnetic field of SCM 1 in the yz-plane going through the center of the cylinder. c) Electric field of SCM 2, which is related to the electric dipole resonance, in the same xz-plane as in fig. 5.4a. d) Magnetic field of SCM 2 in the same yz-plane as in fig. 5.4b. In fig. 5.4, the component of the electric or magnetic fields normal to the chosen plane is zero.

resonance.

The fields describing a self-coupling mode cannot satisfy the Maxwell equations everywhere in space except if the s-value is equal to one. In fig. 5.4, the boundary condition at the first interface ($z = 0$) is not fulfilled. If the s-value is equal to one, the self-coupling mode is a quasi-normal mode [124] and a lot of work has be done on this topic [144–147].

5.3.2 Narrowband metasurface and interpolation

In this section, an interpolation scheme is applied to a narrowband metasurface made of an array of silicon cylinders embedded in glass. The cylinder diameter and height are 470 nm and 609 nm respectively, and the lattice constant is 855 nm. A scale drawing is given in fig. 5.5a. As shown in fig. 5.7e, this metasurface acts as a mirror with a reflection of more than 97% for a large wavelength range except at the wavelength between 1470 nm and 1480 nm where a narrow resonance is present. As expected, this mirror-like behavior is also insensitive to a change in the metasurface dimensions as shown in fig. 5.5b, where the metasurface considered in this section is indicated by a red cross. From the analysis of the resonant self-coupling mode, the maximum of the resonance peak is at 1474 nm and the width of this resonance is around 1 nm. Hence, the Q-factor of the resonance is around 1400.



(a)

(b)

Figure 5.5 – a) Scale drawing of the narrowband metasurface, which is composed of an array of cylinders embedded in glass. The cylinder diameter and height are respectively 470 nm and 609 nm, and the lattice constant is 855 nm. b) Transmission efficiency for different cylinder diameters and lattice constants. The cylinder height is 855 nm and the metasurface indicated by a red cross is the metasurface shown in fig. 5.5a. f is related to the ratio between the cylinder diameter and the lattice constant and its expression is given in (5.1) with $d_{min} = 50$ µm.

The s-values of the main self-coupling modes for different wavelengths are given in fig. 5.6a, where the s-values of the resonant self-coupling mode are represented by the blue curve.

Since the s-values of the self-coupling modes turn clockwise around the origin of the complex plane, it is possible to know whether a resonance occurs at a lower or higher wavelength from the simulation of the metasurface at a single wavelength. As shown in fig. 5.6b, the phase of the s-values of the resonant self-coupling mode is approximately linear with respect to the wavelength. Therefore, from the simulation of the metasurface for two different wavelengths, the wavelength of the resonance can be estimated.

The mirror-like behavior of the metasurface is due to the destructive interference of the contribution of the two non-resonant self-coupling modes to the transmitted plane wave and, since those two self-coupling modes are non-resonant, this mirror-like behavior is broadband. The explanation of the broadband mirror-like behavior and the presence of the resonance in the neighborhood can be made from the analysis of the self-coupling modes of a metasurface at a single wavelength.

Using the concept of self-coupling mode, it is possible to interpolate the response of a resonant metasurface even if the metasurface is simulated only outside the resonance. In this section, the narrowband metasurface is simulated every 10 nm in wavelength from 1440 nm to 1520 nm. As shown in figs. 5.7b and 5.7c, it is not possible to interpolate directly the transmitted field amplitude or the transmission efficiency. However, there is an anomaly in fig. 5.7b at a wavelength between 1470 nm and 1480 nm.

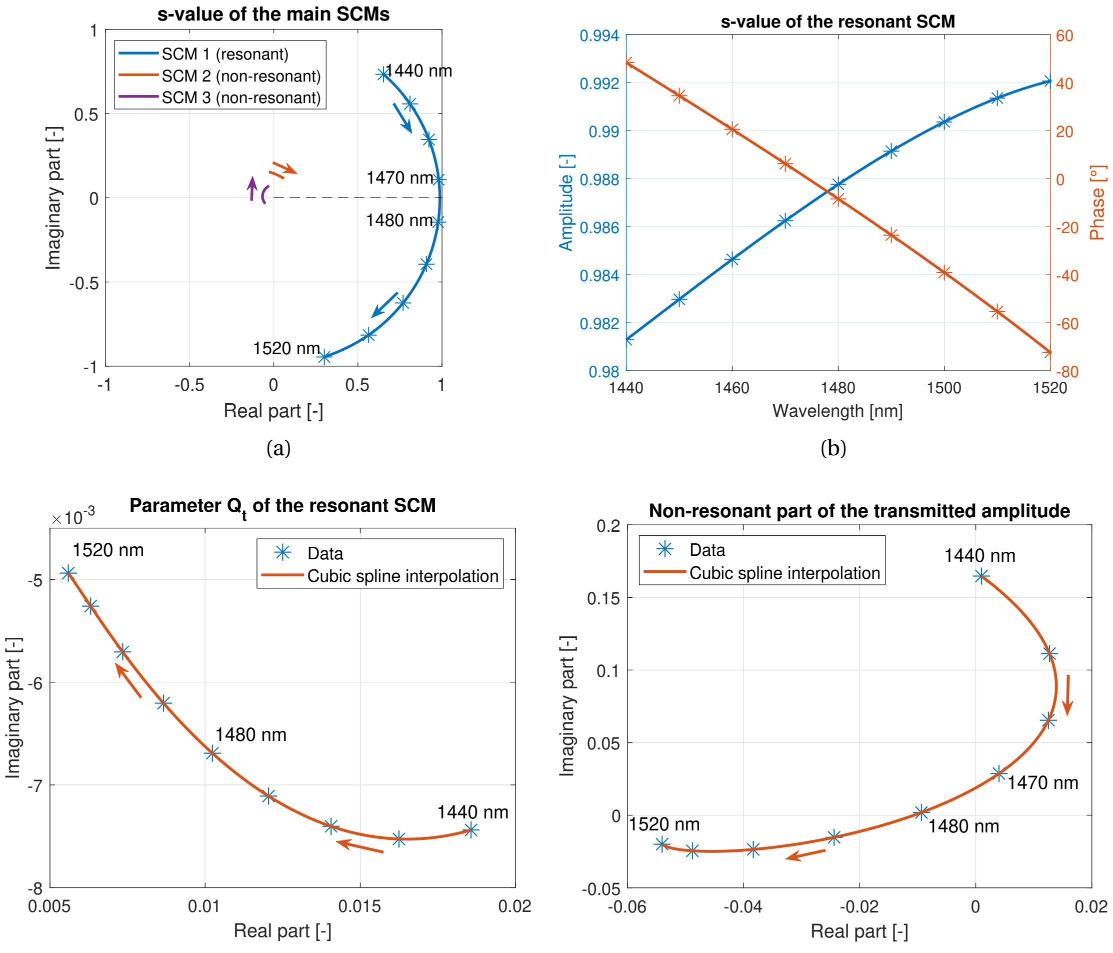
From fig. 5.6a, the self-coupling mode SCM 1 is strongly resonant, meaning that its contribution to the transmitted plane wave varies greatly as its s-value goes through the real axis. As a reminder, the contribution of the self-coupling mode m on the transmitted plane wave is given by (5.11)

$$\vec{t}_m = \frac{1}{1 - s_m} \hat{Q}_{t,m} \vec{p}. \tag{5.13}$$

Equation (5.13) can be simplified because, due to the symmetry of the metasurface considered in this section, the metasurface is polarization independent and the polarization state of the transmitted plane wave is the same as that of the incident plane wave. For simplicity, $\vec{p}$ is chosen to be $(1,0)^T$. Therefore, only the first component of $\vec{t}$ is taken into account and only one element in $\hat{Q}_{t,m}$ is needed.

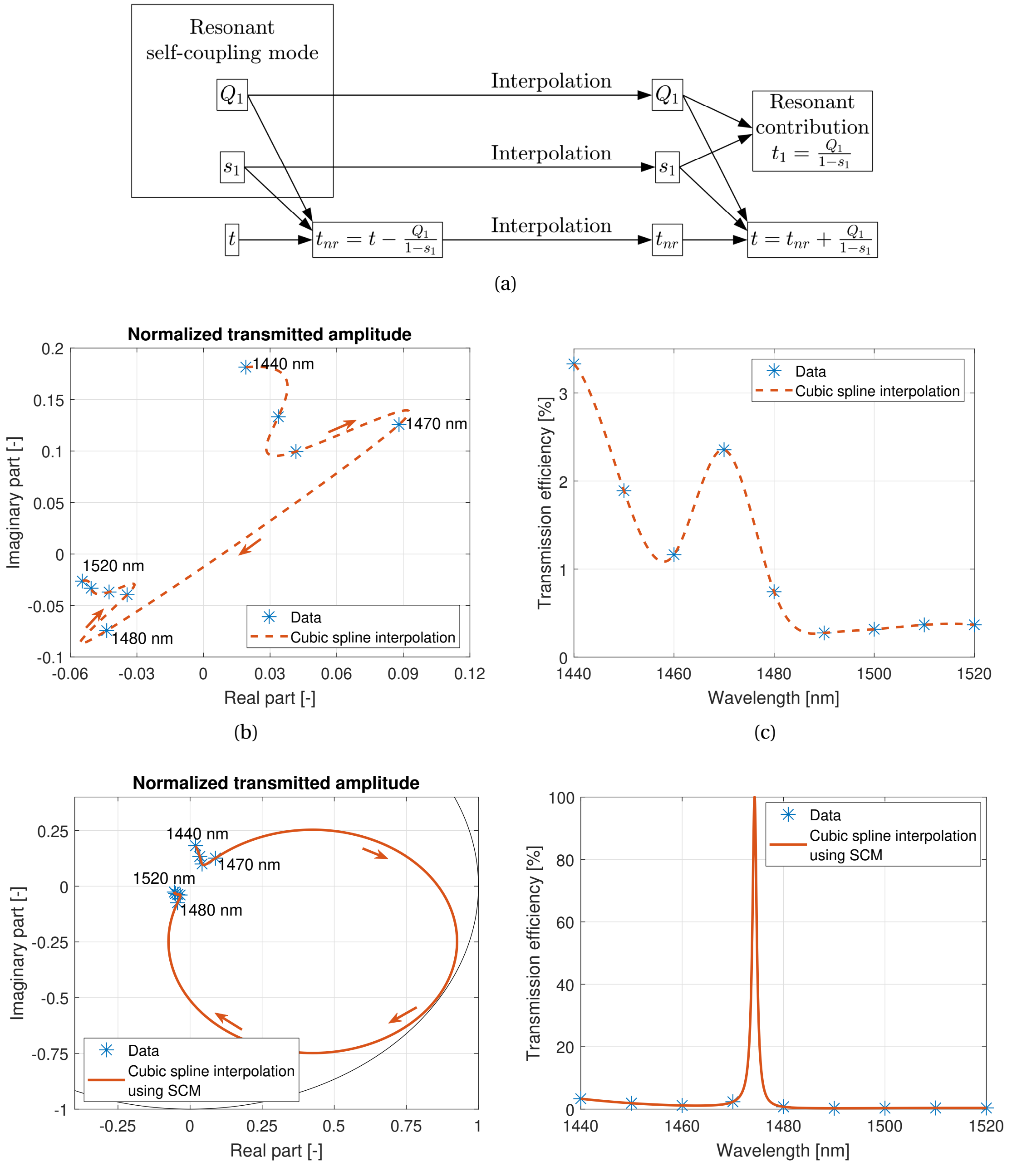
Since the contribution of the resonant self-coupling mode leads to strong variations in the metasurface's response, its contribution is removed from the total response and the remaining transmitted field t_{nr} is a smooth function as shown in fig. 5.6d. In other words, the total transmitted field t is given by

$$t = t_{nr} + \frac{Q_{t,1}}{1 - s_1}, \tag{5.14}$$



(a) (b) (c) (d)

Figure 5.6 – a) s-value of the main self-coupling modes in function of the wavelength. For every 10 nm in wavelength, the s-value of the resonant self-coupling mode SCM 1 is marked by a star. The dashed line is the positive real axis. The red and blue curves are the cubic spline interpolation of the data points represented by stars of the corresponding color. b) Amplitude and phase of the s-value of the resonant self-coupling mode. c-d) The parameter Q_t of the resonant self-coupling mode and the non-resonant part of the transmitted amplitude in function of the wavelength. The red curves are the cubic spline interpolation of the data points represented by stars.



(a)

(b)

(c)

(d)

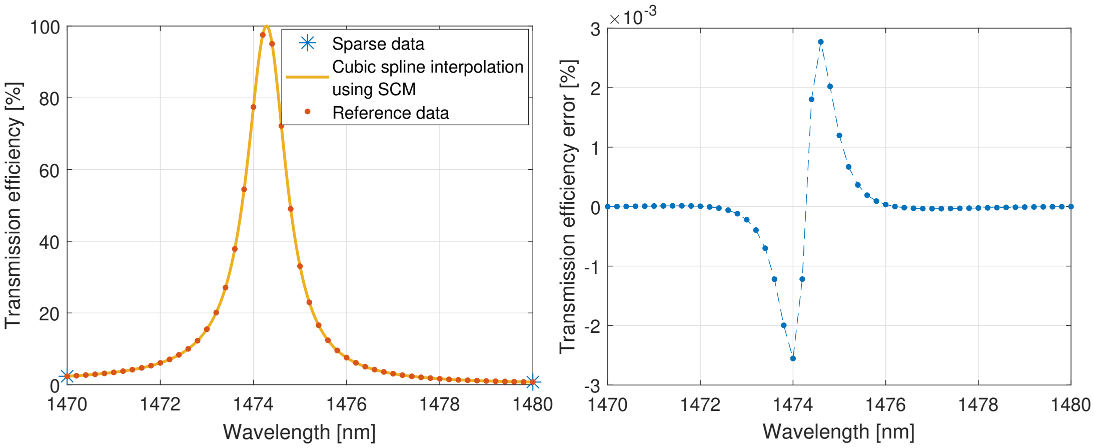
(e)

Figure 5.7 – a) Diagram summarizing the interpolation scheme. Q_1, s_1 and t are obtained from the simulation of the metasurface and the three smooth functions, Q_1, s_1 and t_{nr}, are interpolated. Then, the interpolated transmitted amplitude is obtained along with the resonant contribution. b-c) Naive cubic spline interpolation of the transmitted field and the transmission efficiency from the data points represented by stars. The transmitted amplitude is normalized such that the amplitude squared is the transmission efficiency. d-e) Interpolation of the transmitted field and the transmission efficiency using the concept of self-coupling mode. The black line in fig. 5.7d is the amplitude corresponding to 100% transmission efficiency.

where t_{nr} is a smooth function.

As shown in figs. 5.6b and 5.6c, $Q_{t,1}$ and s_1 are also smooth, so, by interpolating those three smooth functions, it is possible to accurately interpolate the response of the metasurface. The interpolation scheme is summarized in fig. 5.7a.

In fig. 5.8a, the interpolated response using the technique presented in this section is plotted and the accuracy of this interpolation is given in fig. 5.8b. In this case, the error on the transmission efficiency is lower than 0.003%, even if the metasurface is simulated outside the resonance. About the method used for the interpolation of t_{nr}, $Q_{t,1}$ and s_1, a large improvement has been observed by using the spline cubic interpolation method instead of a piece-wise cubic interpolation which gives an error of up to 1.5% on the transmission efficiency.



(a)

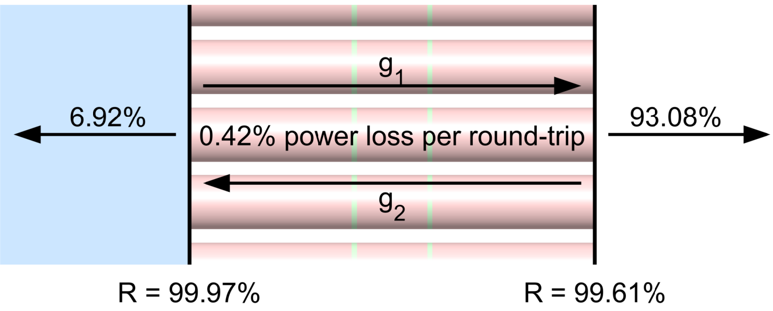
(b)

Figure 5.8 – a) Comparison between the interpolation of the transmission efficiency based on the concept of self-coupling mode, and the reference data, which is the transmission efficiency from the simulation of metasurfaces. b) Absolute error of the interpolated transmission efficiency shown in fig. 5.8a.

The interpolation scheme given in this section can be used when multiple resonances are present by adding more terms in (5.14). A more complicate example is given in section 5.3.3, where the contribution of two resonant self-coupling modes on the transmitted field for both resonances is obtained.

5.3.3 GaAlAs metasurface for laser application and computation of the angular spectrum of a resonance for a symmetric metasurface

The metasurface considered in this section is composed of an array of aluminum-arsenide (AlAs) cylinders on a glass substrate whose dimensions are given in fig. 5.10a. It is designed to act as a metasurface-based ultra-thin laser emitting at 870.6 nm, which corresponds to the band-gap of gallium-arsenide (GaAs), and works in a similar way as a vertical-cavity surface-emitting laser (VCSEL). The main difference between the VCSEL and the metasurface-based laser is that the Bragg grating of the VCSEL is replaced by the structuring of the AlAs layer. In this section, the concept of self-coupling mode is used to describe parameters that are specific for the laser, and the angular spectrum of the resonance is provided. Due to the symmetry of the AlAs metasurface, the number of resonances at normal incidence is even and, in the considered metasurface, two resonances, one per polarization, are present. Because of these two resonances, the interpolation scheme presented in section 5.3.2 has to be adapted.



(a)

Figure 5.9 – a) The AlAs metasurface seen as a cavity bounded by two mirrors. The resonant self-coupling mode, which is composed of forward and backward-propagating eigen-modes, is represented by the two arrows in the cavity. g_1 and g_2 is the gain applied to respectively the forward and backward-propagating eigen-modes to the GaAs quantum wells represented by the light green lines. The two arrows outside the cavity represent the plane waves going away from the cavity. The values above these arrows is the ratio of the power emitted by the cavity in the corresponding direction to the total emitted power.

As mentioned earlier, the metasurface-based laser and the VCSEL have the same working principle. The optical gain is obtained from the GaAs quantum wells, which are placed inside a cavity, and the laser emits perpendicularly to the active region. The cavity is usually made of $Ga_xAl_{(1-x)}As$, but, in this section, the chosen material is AlAs for simplicity. There are several differences, which lead to advantages and disadvantages. In terms of fabrication, the advantage of the metasurface-based laser is that it does not have a Bragg grating, but the substrate has to be a low refractive index material such as glass, meaning that the AlAs layer cannot be grown by epitaxy on the substrate. Moreover, the AlAs has to be etched in order to obtain a metasurface and the performance of the metasurface-based laser is strongly dependent of the quality of the etching process. The difficulties related to the etching process are less present for the VCSEL.

In terms of the physics, the fields inside a metasurface-base laser are larger for the same Q-factor and output power because the effective length of the cavity is smaller. It has the advantage that the gain per round-trip is lower for the same output power, but the absorption of the materials inside the cavity is more problematic. Another specificity of the metasurface-based laser is that the physics is more complex because the gain region does not cover the whole xy-plane, meaning that the different modes may not be amplified the same way. Because the gain per round-trip is small, it should not impact significantly the performance of the laser if it is not taken into account during the design process.

Different set of parameters are considered during the design process. The first two sets are the metasurface dimensions and the s-value of the resonant self-coupling mode, which gives how the light is trapped inside the metasurface. The parameter that is specific for such application is the direction in which the resonance emits. For a metasurface-based laser, it is important that the resonance emits mainly outside the structure instead of in the substrate, even if it is possible to add a metal layer in order to reflect back a part of the light to the metasurface. Using the concept of self-coupling mode, the direction of emission can be obtained from the column corresponding to the resonant self-coupling mode of the matrix $\hat{T}_r$ and $\hat{T}_t$ expressed in (5.9). Since the metasurface considered in this section is a zeroth-order grating, only two plane waves are considered in the substrate and superstrate. Therefore, $\vec{t}_r$ and $\vec{t}_t$, which is the column of interest in $\hat{T}_r$ and $\hat{T}_t$, are 2-elements vectors that describe the plane waves emitted by the resonance. From those vectors, the ratio of the power flow going in the two directions can be computed assuming that there is no gain inside the cavity.

A self-coupling mode can be seen as two sets of eigen-modes, one propagating forward and the other propagating backward, and they create a standing wave inside the metasurface. By assuming that the gain is represented by a real number that scales the weight of the eigen-modes equally, the system can be seen as shown in fig. 5.9, where g_1 and g_2 are the gains due to the GaAs quantum wells on respectively the forward and backward propagating modes. From equation (5.9) and by taking into account the gain, the weights of the emitted plane waves are expressed as

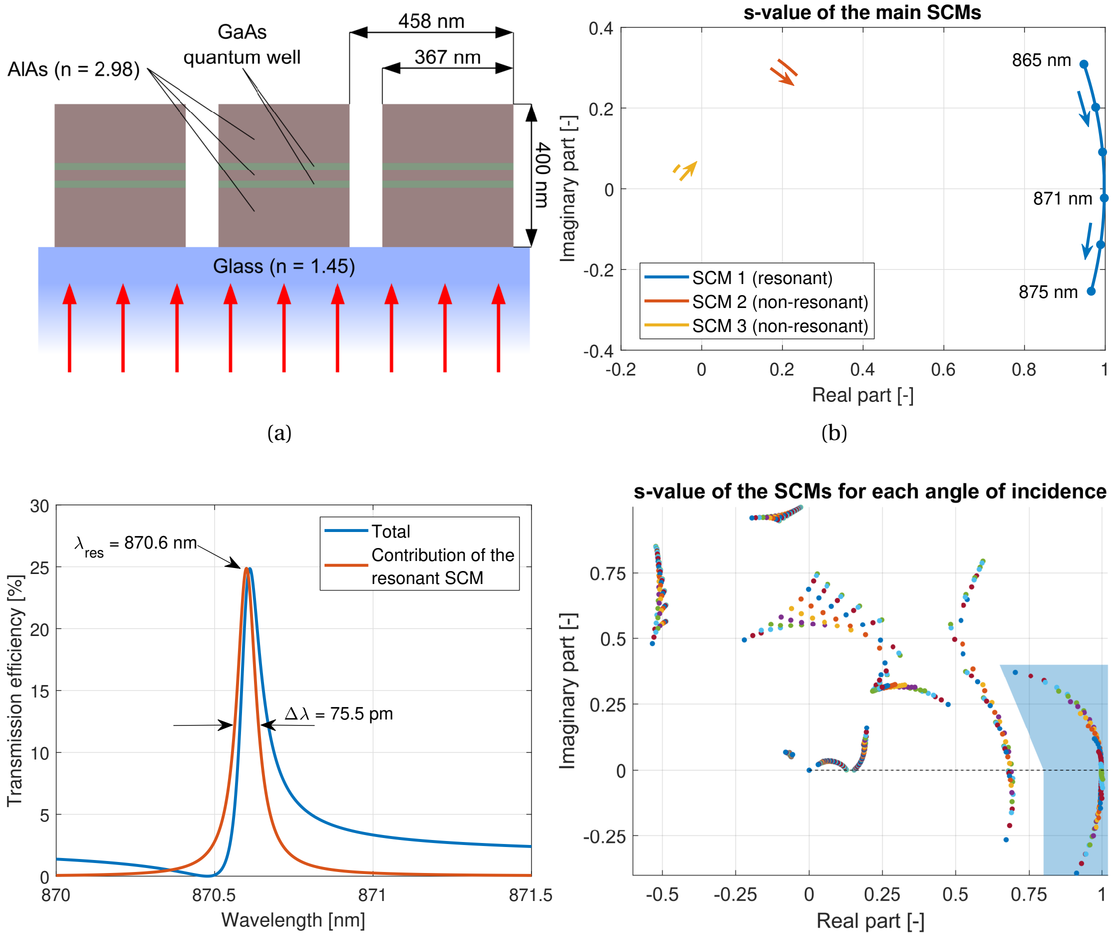
$$\begin{aligned} \vec{t}_r &= \sqrt{g_1 g_2}\,\hat{T}_2 \hat{\Gamma} \hat{R}_3 \hat{\Gamma}\, \hat{v}_r \\ \vec{t}_t &= \sqrt{g_1}\,\hat{T}_3 \hat{\Gamma}\, \vec{v}_r, \end{aligned} \tag{5.15}$$

where $\vec{v}_r$ is the weights of the forward-propagating eigen-modes that compose the resonant self-coupling mode just after the first interface. Since the self-coupling modes are not normalized and only the ratio of the power flows is meaningful, not taking the gain into account is equivalent to the case where the gain applies only on the forward-propagating eigen-modes, meaning that $g_2 = 1$. If the gains g_1 and g_2 are assumed equal, the value of g_1 needs to be computed.

The system is lasing when the gain is equivalent to the loss, meaning that the s-value of the resonant self-coupling mode including the gain is one. If s_r is the s-value of the resonant self-coupling mode without gain, it means that

$$g_1 g_2 s_r^2 = 1. \tag{5.16}$$

Since g_1 and g_2 are assumed equal, g_1 and g_2 are equal to $|s_r|^{-1}$ and the ratio of the power



(a) (b)

(c) (d)

Figure 5.10 – a) Scale drawing of the AlAs metasurface with its dimensions. The red arrows represent the illumination used to characterize the resonance. b) *s*-value of the main self-coupling modes in function of the wavelength. There is a dot every 2 nm. c) Transmission efficiency and the contribution of the resonant self-coupling mode on the efficiency in function of the wavelength. The wavelength and width of the resonance is estimated from the contribution of the resonant self-coupling mode. d) *s*-value of the self-coupling modes in function of the angle of incidence. The *s*-values in the light blue region are the *s*-values of the resonant self-coupling mode of interest.

flows to the total emitting power is obtained from the power flow due to the weight of the plane waves described by $\vec{t}_r$ or $\vec{t}_t$. Using the convention given in section 3.2.1, the power flow emitted on both sides of the metasurface is proportional to

$$\begin{aligned} P_r &= n_{sub}^3 |t_{r,TM}|^2 + n_{sub}|t_{r,TE}|^2 \\ P_t &= n_{sup}^3 |t_{t,TM}|^2 + n_{sup}|t_{t,TE}|^2, \end{aligned} \tag{5.17}$$

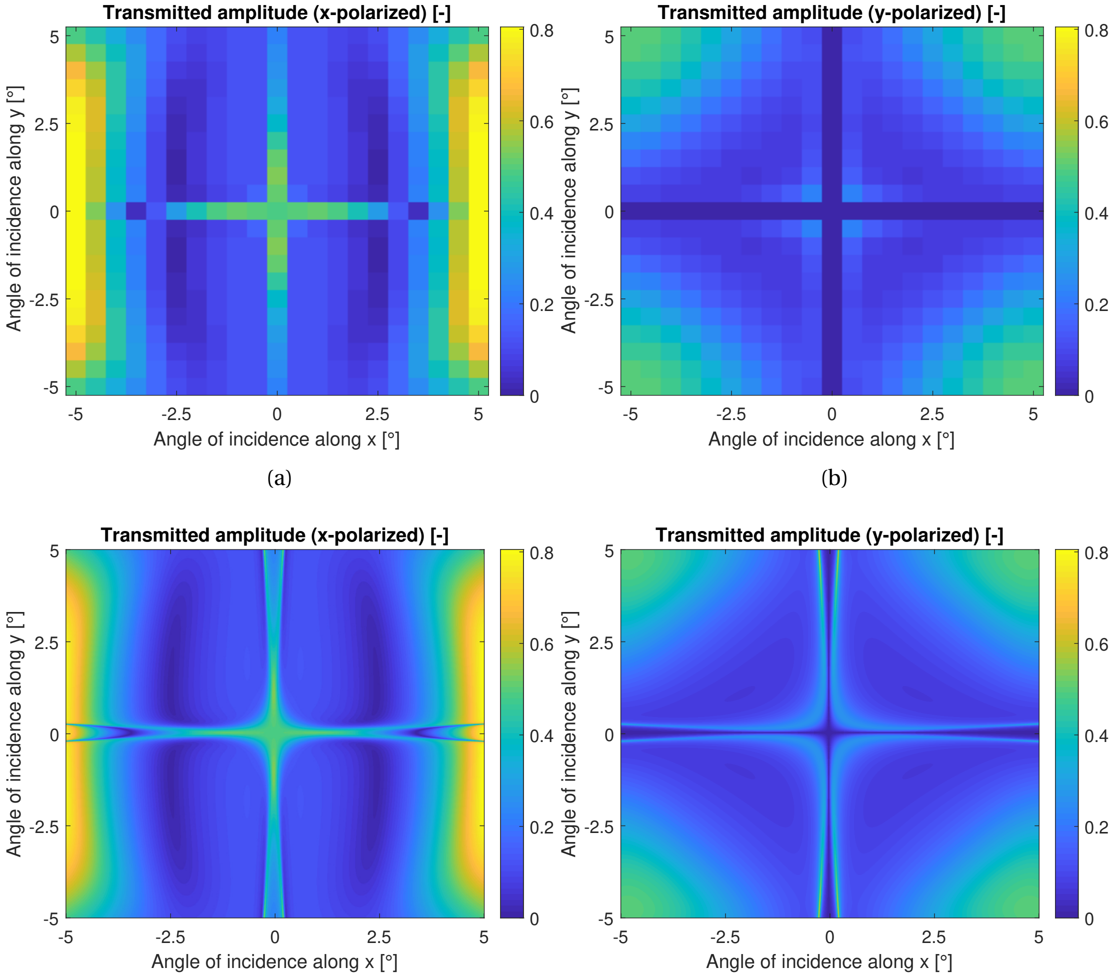
where n_{sub} is the refractive index of the substrate, which is glass, and n_{sup} is the refractive index of the superstrate, which is air. For the AlAs metasurface, those ratios are given in fig. 5.9. The main results are that 93.1% of the emitted power goes outside the structure and, since g_1 and g_2 are known, the power loss per round-trip, which is compensated by the gain, is 0.42%.

The resonance is characterized by illuminating the metasurface from the substrate as shown in fig. 5.10a. From the s-value of the main self-coupling modes shown in fig. 5.10a and due to the symmetry of the unit cell, one resonant self-coupling mode per polarization is present. The transmission efficiency and the contribution of the resonant self-coupling mode are given in fig. 5.10c and, by looking at the contribution, the resonance occurs at 870.6 nm and the full-width half-maximum of the resonance is 75.5 pm, giving a Q-factor of around 11 500.

It is more difficult to get the angular spectrum of the resonance than the spectral response because it is a function that depends on two variables, meaning that the angular spectrum is more computationally expensive to get. It is therefore particularly advantageous to use an efficient interpolation scheme. In fig. 5.10c, the spectral characteristics of the resonance are obtained from the contribution of the resonance on the transmitted plane waves. The same can be done for the angular characteristics of the resonance by looking at its contribution on the transmitted plane waves for different angles of incidence.

For the AlAs metasurface, two resonant self-coupling modes with the same s-value due to the symmetry are present at normal incidence, but, for the different angles of incidence, the s-value of the resonant self-coupling modes splits. The s-value of the self-coupling modes present in the metasurface is shown in fig. 5.10d and the s-value of the resonant self-coupling modes for different angles of incidence is the cluster in the blue area. The particularity of this cluster is that the imaginary part of the s-values can be both positive and negative. In this case, it can be shown that it exists a set of angles of incidence such that the s-value stays on the real axis, meaning that the angle of incidence can be changed while staying on resonance. When the metasurface resonates at normal incidence, the result is the star-like pattern shown in figs. 5.11 and 5.12.

Since two resonant self-coupling modes are present, the transmitted plane waves for each angle of incidence are described by the sum of the contribution of the two resonant self-coupling modes, and a non-resonant term $\vec{t}_{nr}$, which is a smooth function. Hence, from equation (5.11), the transmitted amplitude $\vec{t}_{tot}$ is given by



(a) (b)

(c) (d)

Figure 5.11 – a-b) Amplitude of the transmitted plane waves in function of the angle of incidence for x and y-polarizations. Each pixel corresponds to a simulation. c-d) Interpolation based on the concept of self-coupling mode of the amplitude shown in figs. 5.11a and 5.11b

$$\vec{t}_{tot} = \vec{t}_{nr} + \frac{1}{1-s_1}\hat{Q}_{t,1}\vec{p} + \frac{1}{1-s_2}\hat{Q}_{t,2}\vec{p}. \tag{5.18}$$

If the Fourier modal method presented in chapter 3 is used, the plane waves are TM and TE-polarized and this choice of polarization has a singularity at normal incidence, which may lead to anomaly in the angular spectrum that cannot be interpolated properly. Hence, it is safer to decompose the fields in the substrate and superstrate into x and y-polarized plane waves. In this section, the illumination is chosen to be x-polarized with $\vec{p}_{xy} = (1,0)^T$, but, due to the symmetry of the metasurface, the angular spectrum is the same for x and y-polarized illumination.

The matrices that transform the weights of the TM and TE-polarized plane waves into the weights of the x and y-polarized plane waves are given by equation (3.30) and, using those matrices and assuming that the illumination is x-polarized, equation (5.18) becomes

$$\vec{t}_{tot,XY} = \vec{t}_{nr,XY} + \frac{1}{1-s_1}\vec{Q}'_{t,1} + \frac{1}{1-s_2}\vec{Q}'_{t,2}, \tag{5.19}$$

where $\vec{t}_{tot,XY}$ describes the transmitted field in terms of x and y-polarized plane waves, and $\vec{t}_{nr,XY}$ and $\hat{Q}'_{t,m}$ are given by

$$\begin{aligned} \vec{t}_{nr,XY} &= \hat{P}\begin{pmatrix} k_{z,sup}s_x & \mu k_0 s_y \\ k_{z,sup}s_y & -\mu k_0 s_x \end{pmatrix}\vec{t}_{nr} \\ \vec{Q}'_{t,m} &= \hat{P}\begin{pmatrix} k_{z,sup}s_x & \mu k_0 s_y \\ k_{z,sup}s_y & -\mu k_0 s_x \end{pmatrix}\hat{Q}_{t,m}\frac{1}{\mu k_0 k_{z,sub}}\begin{pmatrix} \mu k_0 s_x & \mu k_0 s_y \\ k_{z,sub}s_y & -k_{z,sub}s_x \end{pmatrix}\vec{p}_{XY}. \end{aligned} \tag{5.20}$$

$k_{z,sub}$ and $k_{z,sup}$ are the z-component of the k-vector in respectively the substrate and the superstrate. $\hat{P}$ is a diagonal matrix that normalizes the weights of the plane waves such that the efficiency is obtained by taking the absolute square of those weights if either the x or the y-polarized plane wave is excited. For $\vec{p}_{xy} = (1,0)^T$ and assuming that the substrate and the superstrate are lossless, $\hat{P}$ is given by

$$\begin{aligned} \hat{P} &= \begin{pmatrix} \sqrt{\frac{\mathrm{Re}\{[\psi_{X,sup}|\psi_{X,sup}]\}}{\mathrm{Re}\{[\psi_{X,sub}|\psi_{X,sub}]\}}} & 0 \\ 0 & \sqrt{\frac{\mathrm{Re}\{[\psi_{Y,sup}|\psi_{Y,sup}]\}}{\mathrm{Re}\{[\psi_{X,sub}|\psi_{X,sub}]\}}} \end{pmatrix} \\ &= \begin{pmatrix} \sqrt{\frac{(k_x^2+k_{z,sup}^2)k_{z,sub}}{(k_x^2+k_{z,sub}^2)k_{z,sup}}} & 0 \\ 0 & \sqrt{\frac{(k_y^2+k_{z,sup}^2)k_{z,sub}}{(k_x^2+k_{z,sub}^2)k_{z,sup}}} \end{pmatrix}, \end{aligned} \tag{5.21}$$

where $\psi_{X,sup}$ and $\psi_{Y,sup}$ are respectively the x and y-polarized plane wave in the superstrate,

$\psi_{X,sub}$ is the x-polarized plane wave in the substrate and $\mathrm{Re}\{[\psi|\psi]\}$ is the power flow along the z direction due to the mode ψ.

From section 3.2.1, $[\psi_X|\psi_X]$ and $[\psi_Y|\psi_Y]$ are

$$[\psi_X|\psi_X] = \frac{k_x^2 + k_z^2}{\mu k_0 k_z} \qquad [\psi_Y|\psi_Y] = \frac{k_y^2 + k_z^2}{\mu k_0 k_z}. \tag{5.22}$$

As a reminder, x and y-polarized plane waves are not orthogonal in term of the power flow.

Using the equation in the form given in (5.19), seven interpolations are required and, for each incident angle, the determination of the two resonant self-coupling modes, named SCM 1 and SCM 2, has to be done such that s_1, s_2, $\vec{Q}_{t,1}$ and $\vec{Q}_{t,2}$ are smooth functions, which is a difficult task. A practical solution to reduce the complexity of this problem is to write equation (5.19) in the form

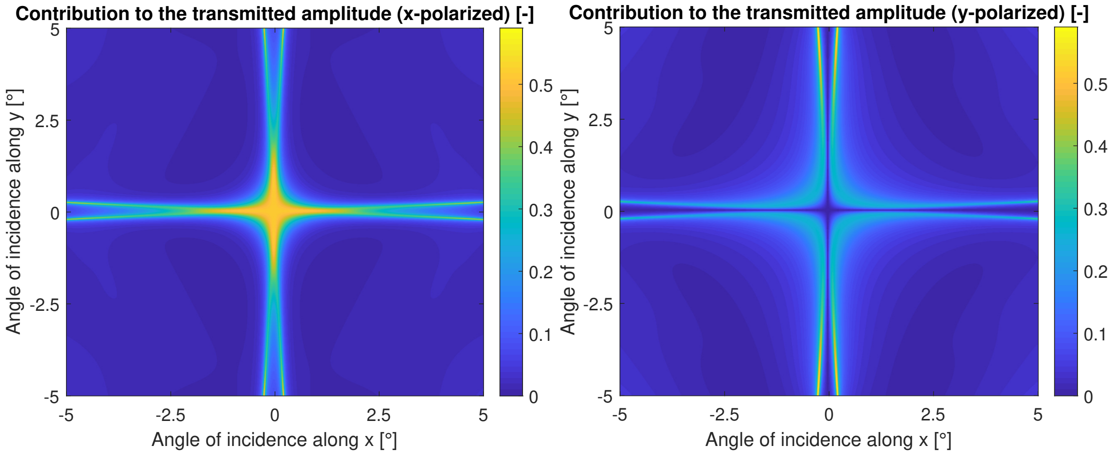
$$\vec{t}_{tot,XY} = \vec{t}_{nr,XY} + \frac{1}{1-s_{res}}\vec{Q}'_{t,res} \tag{5.23}$$

with

$$\begin{aligned} s_{res} &= s_1 + s_2 - s_1 s_2 \\ \vec{Q}'_{t,res} &= (1-s_2)\vec{Q}'_{t,1} + (1-s_1)\vec{Q}'_{t,2}. \end{aligned} \tag{5.24}$$

Hence, the two resonant self-coupling modes are treated as it is a single one, reducing the number of required interpolations and the need to carefully determine which self-coupling mode is SCM 1 and 2.

Before interpolation, the amplitude of the transmitted plane waves, which is obtained from $\vec{t}_{tot,XY}$, are given in figs. 5.11a and 5.11b, where each pixel corresponds to a simulation. After interpolating $\vec{t}_{nr,XY}$, s_{res} and $\vec{Q}'_{t,res}$, the absolute value of the x and y-components of the resulting $\vec{t}_{tot,XY}$ is shown in figs. 5.11c and 5.11d and the contribution of the resonant self-coupling modes is given in figs. 5.12a and 5.12b. Due to the high frequency features, it would be computationally intensive to get such results with most methods with the exception of methods based on the quasi-normal modes due to their similarities with the self-coupling modes. For the metasurface-based laser, the s-value of the resonant self-coupling modes are more relevant because the metasurface emits only if the presence of the gain changes one of those s-values to one. If the presence of the gain does not affect the phase of the s-value, the first condition for lasing is that the phase of one of the s-values is real. If the phase of the s-value is not zero, the phase indicates if the metasurface may emits at a different wavelength or angle. The second condition is that the quantum wells can provide the necessary gain to compensate the loss. As a first approximation, the maximum gain is proportional to the



(a)

(b)

Figure 5.12 – a-b) Contribution of the resonant self-coupling modes on the amplitude of the transmitted plane waves after interpolation for different angles of incidence and for x and y-polarization.

number of photons in the metasurface. Hence, based on fig. 5.9, the second condition is that one of the s-value is close enough to one such that g_1 and g_2 is smaller or equal to the maximum possible gain g_{max}, which depends of the number of quantum wells and their properties.

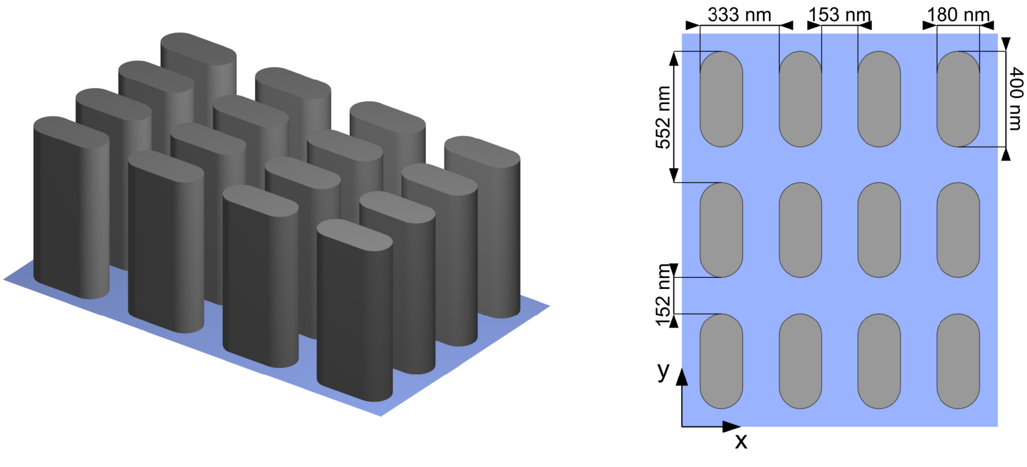
At a wavelength of 1120 nm, one of the s-value of the resonant self-coupling modes at the locations of the lines present in fig. 5.12a is purely real and close to one, meaning that the angular spectrum of the light emitted by the metasurface is composed of four lines as in fig. 5.12a. Such feature may not be wanted for a laser, but it is worth knowing that it can occur.

As a general comment, a metasurface-based laser composed of holes is better than the ones based on cylinders, as proposed in this section, because it is then possible to create an electrical circuit with the quantum wells inside a p-n junction.

5.3.4 Design of high-Q metasurface for sensing

In this section, a resonant metasurface composed of silicon cylinders with obround cross-section on a glass substrate in water is presented. A 3D drawing of the metasurface is given in fig. 5.13a and its cross-section with the different dimensions are given in fig. 5.13b. The metasurface thickness is 807 nm. The first part of this section shows how chapter 3 combined with the concept of self-coupling modes can help in the design of a resonant metasurface. The second part is the estimation of the spatial extent of a resonance and on the effect of a change in the metasurface thickness on the angular spectrum of the resonance.

The metasurface considered in this section operates under y-polarized illumination and it is designed so that the angular spectrum of the resonance is highly asymmetric in order to be used as the sensor based on surface plasmon resonance presented in [148], where the dimensions of the structure vary in one direction, shifting the wavelength at which the resonance occurs. Moreover, the Q-factor is maximized while taking into account the absorption of silicon and water. The wavelength is chosen to be 1120 nm because the absorption of silicon and water is similar and, while increasing the wavelength, the absorption of water increases and the absorption of silicon decreases. At that wavelength, the refractive index of water is 1.33 and its extinction coefficient is $7.09 \cdot 10^{-6}$. For silicon, its refractive index is 3.56 and its extinction coefficient is $1.70 \cdot 10^{-5}$.

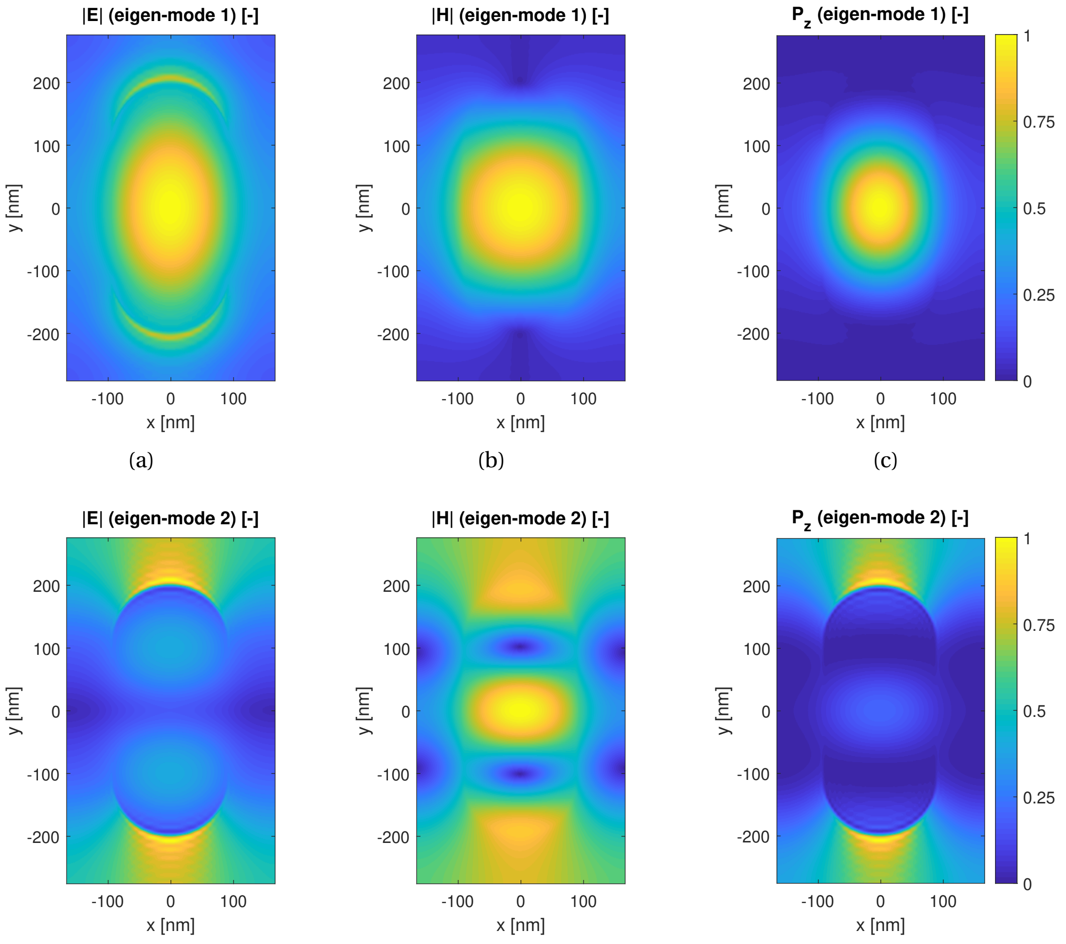


(a) (b)

Figure 5.13 – a) Scale drawing of the metasurface considered in this section. The cylinders are made of silicon and they are surrounded by water. The substrate is made of glass. b) Top view of the metasurface with the dimensions of the unit cell and the obround cross-section of the cylinders.

In order to guarantee that the resonance has an asymmetric angular spectrum, which means that the spatial extent of the resonance is also asymmetric, the condition that only two propagating eigen-modes are excited in the metasurface is imposed. The amplitude of the electric and magnetic field, and the power flow along the z direction of the first eigen-mode are shown in figs. 5.14a to 5.14c. This eigen-mode is confined inside the cylinders and it is the fundamental eigen-mode since the eigen-mode present in single-mode metasurfaces has the same characteristics. The fields of the second eigen-mode are mostly located within the gap at the bottom and top of the cylinders as shown in figs. 5.14d to 5.14f. Since the fields are weak in the gap at the left and right of the cylinders, the light inside the metasurface propagates more easily along the y-axis than along the x-axis, leading to the asymmetry in the angular spectrum of the resonance. The fields of a third eigen-mode are located partly within the gaps at the left and right of the cylinders, meaning that the asymmetry is expected to be weaker and, if this asymmetry is taken into account during the design process, it requires the simulation of the structure at different angles of incidence.

The first drawback of having only two propagating eigen-modes inside the metasurface is that there are less resonances in the design space, so it is more difficult to find metasurfaces with the desired resonance. The second drawback is that the resonance leads to a dip in transmission since the broad-band mirror-like effect, which requires two non-resonant self-coupling modes as shown in section 5.3.2, cannot exist.



(a) (b) (c)

(d) (e) (f)

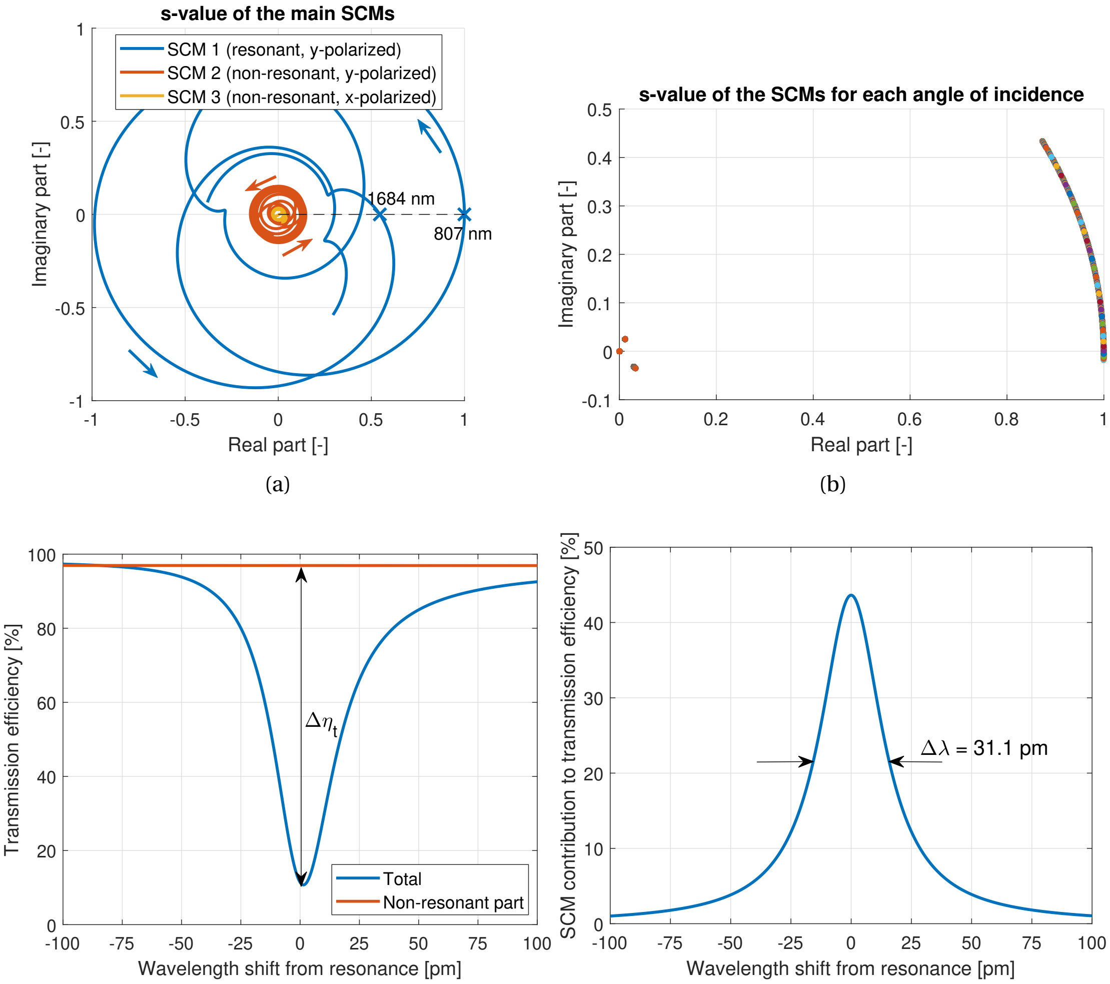
Figure 5.14 – a-c) Amplitude of the electric field, amplitude of the magnetic field and the power flow along z of the first eigen-mode over an unit cell. They are normalized such that the maximum is one. d-f) Same quantities as in figs. 5.14a to 5.14c for the second eigen-mode.

The design of a high-Q resonant metasurface is challenging because the objective is to find a very sharp feature, meaning that the response of a metasurface is not sufficient in order to know if the structure is on resonance, how sharp is the resonance and how the resonance affects the response. Using the concept of self-coupling mode, the simulation of a metasurface answers most of those concerns, but it requires a second simulation to estimate where the resonance is. The Fourier modal method presented in chapter 3 allows to get the response of the metasurface for any metasurface thickness at a negligible computational cost. Hence, it is possible to get the s-value of the self-coupling modes in the complex plane for any metasurface thickness, which gives fig. 5.15a, and, from those s-values, a set of resonant metasurfaces.

In fig. 5.15a, there are two resonant metasurfaces for a thickness between 200 nm and 2000 nm. The first resonance occurs at a thickness of around 807 nm, which corresponds the thickness of the metasurface considered in this section, and has a high Q-factor since the s-value is very close to one. The second resonance has a low Q-factor and occurs for a very thick metasurface. The benefits of using this techniques to get the resonance is two-fold. First, the s-value turns mostly counterclockwise around the origin while increasing the metasurface thickness. Hence, it is easier to identify anomalies. The reason behind this direction of rotation is that, since the propagation constant of the eigen-modes is positive, the phase accumulation during a round-trip increases while increasing the metasurface thickness. A similar explanation is given earlier for the direction of rotation while varying the wavelength. The second benefit is that the number of propagating modes inside the metasurface stays constant, meaning that the dynamic of the system does not change due to the apparition of an additional propagating mode.

In order to get the metasurface considered in this section, four dimensions need to be optimized: two for the shape of the cylinders and two for the dimensions of the unit cell. The metasurface thickness is used in order to be on resonance, so it is not counted as a dimension to optimize. The first part of the design process is to simulate a random set of metasurfaces, trying to focus on the parameter space where metasurfaces with two propagating eigen-modes are present. From the simulated metasurfaces, all the resonances are listed using the techniques described earlier and, for each resonance, the parameters of interest that can be obtain without the simulation of additional metasurfaces, are computed. For the metasurface considered in this section, the parameters of interest are the aspect ratios of the cylinders and the gaps, the s-value of the resonant self-coupling modes and the difference between the transmission efficiency with and without the contribution of the resonant self-coupling mode. This difference is given by $\Delta\eta_t$ and it is shown in fig. 5.15c. The Q-factor cannot be estimated from the simulation of a single metasurface. However, the s-value of the resonant self-coupling modes gives an indication on the amplification of the fields inside the metasurface and it is strongly related to the Q-factor. Based on the knowledge of the Fabry-Pérot cavity, the Q-factor depends also on the cavity length.

The second part is to choose a suitable resonance from the list and to use the metasurface that have this resonance as a starting point for a local optimization. The local optimization algorithm used in this work is the *fminsearch* function in MatLab, which is based on the Nelder-Mead simplex method [149, 150]. At each iteration, the resonance closest to the resonance of the previous iteration is considered and a merit function that takes into account the parameters of interest is computed. The merit function f which is used in this section and that needs to be minimized is



(a) (b)

(c) (d)

Figure 5.15 – a) s-value of the main self-coupling modes in function of the metasurface thickness. The black dashed line is the positive real axis. b) s-value of the self-coupling modes for different angles of incidence. c) Transmission efficiency and the non-resonant part of the transmission efficiency in function of the wavelength. $\Delta\eta_t$ is the difference between those two curves at the wavelength of the resonance. d) Contribution of the resonant self-coupling mode to the transmission efficiency. The width of the resonance is 31.1 pm and the Q-factor is around 36 000.

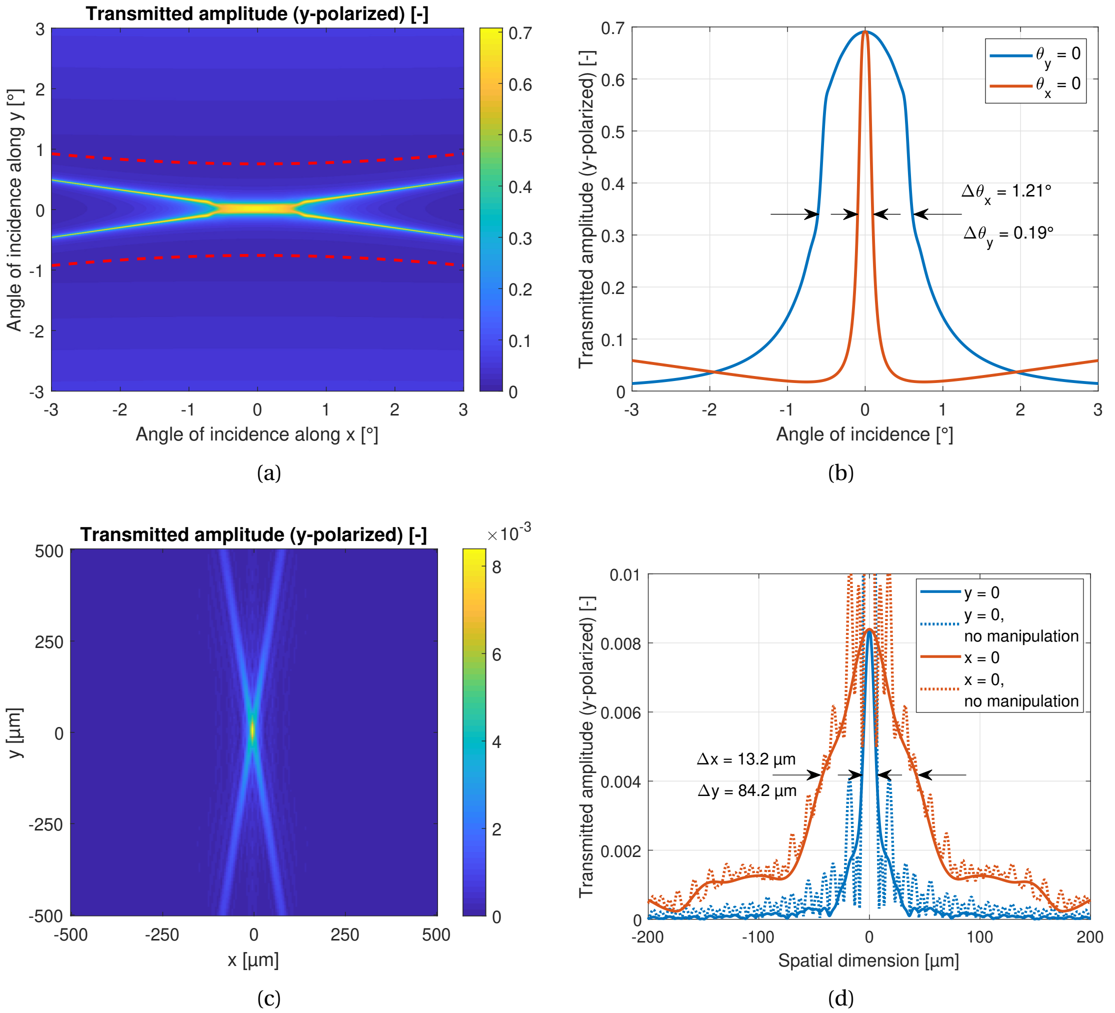
$$f = C_1(1 - s_r) + C_2(A_c^2 + A_g^2) + C_3 e^{\frac{\eta_t}{\eta_{t,ref}}} + C_4(h - h_{target})^2, \tag{5.25}$$

where s_r is the s-value of the resonant self-coupling modes, A_c and A_g are the largest aspect ratio of respectively the cylinder and the gap, η_t is the transmission efficiency, $\eta_{t,ref}$ is the transmission efficiency which is considered as sufficient, h is the metasurface thickness and h_{target} is the desired metasurface thickness.

The different exponent used in (5.25) determines how critical the parameters are expected to be. The efficiencies η_t and $\eta_{t,ref}$ should be replaced by $\Delta\eta_t$ and $\Delta\eta_{t,ref}$ with a change in the sign before the constant C_3, but, for metasurfaces with two propagating eigen-modes, the non-resonant part of the transmission efficiency, which is the red curve in fig. 5.15c, is always close to 100% and the resonance is a dip in the transmission efficiency. The last term in (5.25) is used only when the thickness of the silicon layer deposited on a glass substrate is known and the other dimensions of the metasurface are adapted in order to have the resonance at a wavelength closer to the desired one, which is 1120 nm in this case. The values of C_1, C_2 and C_3 used for the optimization of the metasurface presented in this section are 10^4, 10^{-2} and 0.5 respectively, but they have been adjusted depending on the result after an optimization. In order to have a resonance for a specific metasurface thickness, the constant C_4 is gradually increased between the optimizations until the obtained thickness is sufficiently close to the desired thickness.

The response of the metasurface in function of the wavelength is plotted in fig. 5.15c and the Q-factor is estimated from the contribution of the resonant self-coupling modes, which is given in fig. 5.15d. The Q-factor is around 36 000. The angular spectrum of the resonance is obtained as shown in section 5.3.3 except that only a single resonant self-coupling mode is present. By looking at the s-values of the resonant self-coupling mode in function of the angle of incidence given in fig. 5.15b, the s-values are both above and below the real axis even if the metasurface is resonant at normal incidence. Therefore, lines should appear in the angular spectrum as in section 5.3.3, which is confirmed by fig. 5.17a. Without taking into account those lines, the angular width of the resonance, given in fig. 5.16b is $\Delta\theta_x = 1.21°$ and $\Delta\theta_y = 0.19°$. The angular width is important to consider because, if the divergence of the source is larger than the angular width of the resonance, the dip due to the resonance shown in fig. 5.15c becomes more shallow since a part of the illumination does not couple with the resonance.

The spatial extent of a resonance allows an estimation of how delocalized the resonance is. In a multi-mode metasurface, there is a strong coupling between the cylinders, meaning that a defect or, for the case of sensing, the presence of a particle inside the metasurface has an impact on the response of the metasurface over a large area. In order to estimate the spatial extent of the resonance, a simple approach is to focus a y-polarized light on the metasurface and to compute the angular spectrum of the transmitted field, which is shown in fig. 5.17a. The transmitted field in the spatial domain is then obtained through the Fourier transform.



(a)

(b)

(c)

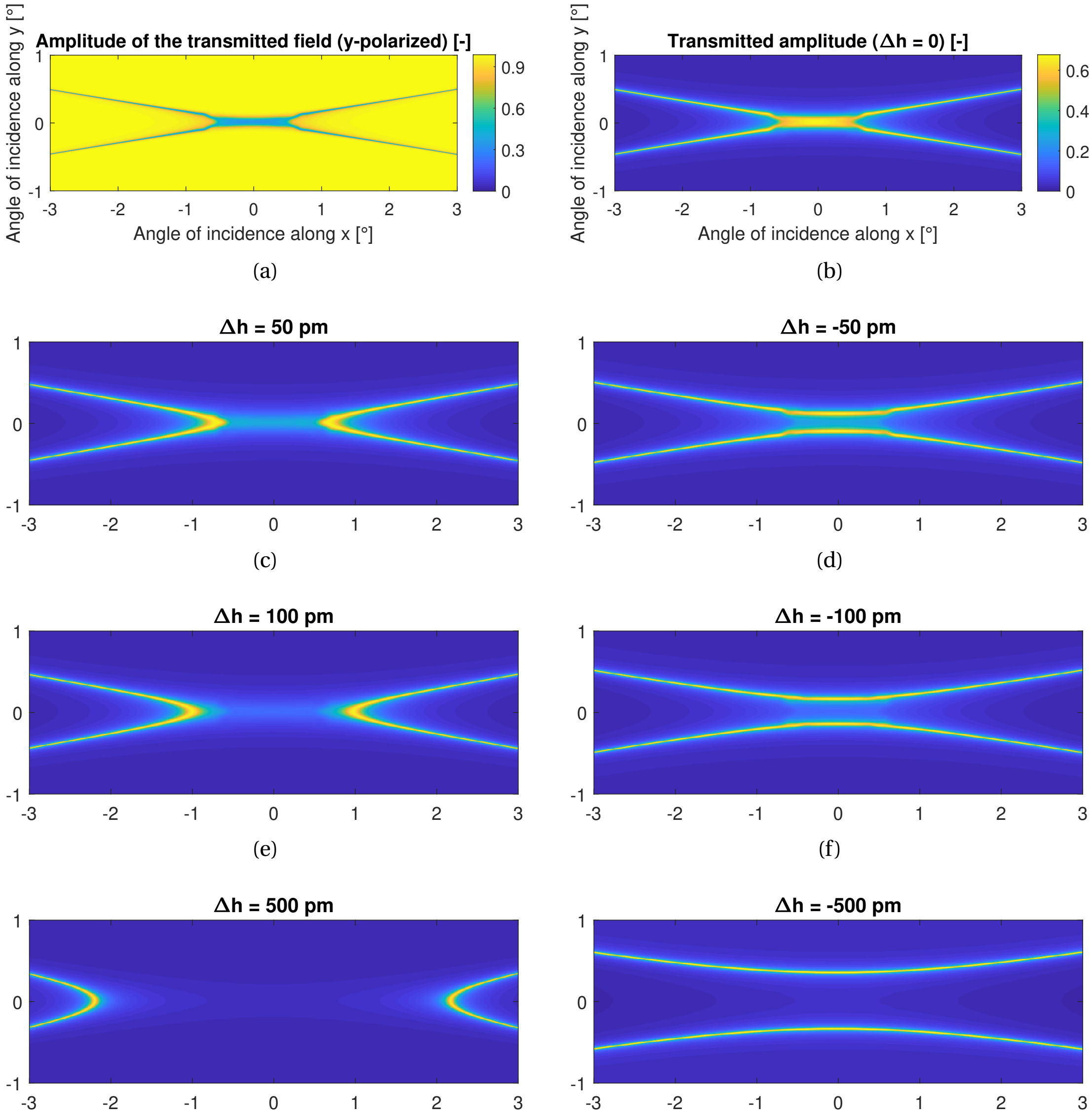
(d)

Figure 5.16 – a) Angular spectrum of the resonance. It is set to zero outside the region delimited by the two red dashed lines for the computation of the spatial extent of the resonance shown in fig. 5.16c. b) Angular spectrum at $\theta_y = 0$ (blue curve) and $\theta_x = 0$ (red curve). θ_x and θ_y are the angle of incidence along x and y respectively. c) Spatial extent of the resonances. The two lines are the footprint of a leaky waveguide mode. d) Spatial extent at $y = 0$ (blue curve) and $x = 0$ (red curve). The dotted curves are obtained by taking the Fourier transform of the angular spectrum of the resonance without setting the angular spectrum to zero outside the region delimited by the red dashed line in fig. 5.16a.

The problems are that the non-resonant part of the transmitted field is insensitive to small defects because the fields inside the metasurface are not enhanced, and it leads to a large peak in the spatial domain, which has nothing to do with the resonance. In order to capture better the spatial extent of the resonance, the Fourier transform is applied on the angular spectrum of the resonance as shown in fig. 5.16a. However, the angular spectrum of the resonance is not an integrable function and, even if the evanescent transmitted plane waves are not considered, it is not reasonable to compute the angular spectrum for all the angles of incidence. By simply taking into account the angular spectrum for an angle of incidence between $-3°$ and $3°$, it is first an arbitrary choice and its Fourier transform oscillates heavily as shown by the dotted curves in fig. 5.16d. A better solution is to define a contour around the resonance where the amplitude of the angular spectrum is as small as possible, and to set the angular spectrum outside this contour to zero before applying the Fourier transform. The presence of the lines in fig. 5.16a complicates this manipulation, but, as shown in fig. 5.16b, for $\theta_y = 0$, where θ_y is the angle of incidence along y, the amplitude of the angular spectrum for $\theta_x = \pm 3°$ is close to zero. Hence, the angular spectrum is set to zero when $|\theta_x|$ is larger than $3°$. By plotting the amplitude of the angular spectrum in function of θ_y for each θ_x, the resonance is between two minimums, creating a valley represented by the two red dashed lines in fig. 5.16a. Therefore, the angular spectrum outside the region delimited by the two red dashed lines is set to zero before applying the Fourier transform. The result is shown in fig. 5.16c and, as shown in fig. 5.16d, the oscillation disappears. The estimated spatial width of the resonance along x and y is respectively $\Delta x = 13.2$ µm and $\Delta y = 84.2$ µm. As predicted by the analysis of the fields of the propagating eigen-modes (fig. 5.14), the resonance is asymmetric and Δy is significantly larger than Δx.

The lines present in fig. 5.16a indicate the presence of leaky waveguide modes as shown in fig. 5.16c. When the metasurface presented in this section is used for sensing, the metasurface is on resonance, meaning that, in the ideal case, the transmission efficiency is around 11%, which is the transmission efficiency at the dip in fig. 5.15c, and the field amplitude inside the metasurface is very high. If a particle is present in the metasurface, this particle will affect the resonance and, because of those leaky waveguide mode, it is expected that this particle creates a X-shape in the transmitted field. A disadvantage of those leaky waveguide modes is that the resonance is more affected by the limited size of the metasurface than if those leaky waveguide modes are not present.

Due to the tolerances on the metasurface dimensions, the wavelength of the resonance shifts and, if the source emits at a well-defined wavelength, it may not be possible to use the metasurface at resonance. However, if lines are present in the angular spectrum of the resonance, it is possible to find an angle incidence such that the metasurface resonates at the condition that the first order does not propagate. This statement is deduced from fig. 5.17, where the angular spectrum of the resonance is given for different metasurface thicknesses. For a sub-atomic difference Δh in the metasurface thickness, the two lines becomes two parabolas and, as $|\Delta h|$ increases, the distance between the two parabolas also increases. It is expected that the metasurface is out of resonance for a difference Δh which has the same



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 5.17 – a) Transmission efficiency in function of the angle of incidence. b) Contribution of the resonant self-coupling mode to the transmission efficiency. c-h) Contribution of the resonant self-coupling mode to the transmission efficiency for different metasurface thicknesses.

order of magnitude as the resonance width ($\Delta\lambda = 31$ pm).

In practice, the Q-factor of fabricated metasurfaces is significantly lower due to imperfections and tolerances on the dimensions and the refractive indices. However, if the metasurface remains periodic, the degradation of the Q-factor should be limited because the light can escape the metasurface by only two channels, the reflected and transmitted zeroth order, which is the condition for the presence of a high-Q resonance. Even for a lower Q-factor, the particularity of the resonance presented in this section should remain.

5.4 Conclusion

Multi-mode metasurfaces have non-intuitive and complex responses, including high-Q resonances. In order to analyse these resonances and to facilitate the design of resonant metasurfaces, the concept of the self-coupling mode is developed. We show that this concept has two important applications: the systematic characterization of a resonance and the interpolation of the response of a metasurface with high-Q resonances. These applications are illustrated in this chapter by four different examples.

The characterization of the resonance using the concept of the self-coupling mode is applied to the Huygens' metasurface where magnetic and electric dipole resonances are present. We show from this example that this concept allows to single out both resonances, to obtain their spectral responses and to get the fields related to those two resonances. From those fields, we can associate the resonant self-coupling mode to the corresponding dipole resonance.

In order to obtain the response of a resonant metasurface as a function of any parameters of the system with a minimum of simulations, we propose an interpolation scheme, which allows an accurate interpolation of the metasurface response even if it is simulated only outside the resonance. We apply this interpolation scheme to a narrowband metasurface, where the transmission efficiency outside and at resonance is a few percent and nearly 100%, respectively. By simulating the narrowband metasurface only outside the resonance, the error in the response is less than 0.003%.

This interpolation scheme is also applied to an AlAs metasurface and an asymmetric metasurface to obtain the angular spectrum of the resonance. In the case of the AlAs metasurface, the interpolation scheme has to be adapted because two resonances are present at normal incidence due to the symmetry, and they are difficult to separate. The resulting angular spectrum is a star-like pattern for the AlAs metasurface and a cross for the asymmetric metasurface. We associate the branches of those patterns with leaky waveguide modes.

From the angular spectrum of a resonance, it is possible to estimate the spatial extent of the resonance, which is the spread of the resonance in the metasurface. We use the capability of the self-coupling mode concept to identify resonances and to interpolate the resonant's response in order to fully characterize the asymmetric resonance presented in section 5.3.4.

From those four examples, we show the self-coupling mode is a powerful concept for the analysis and design of multi-mode metasurfaces.

6 Adjoint method

6.1 Introduction

The objective of the design process is to find a structure that has the best performance under some constraints. In order to reach this objective, a merit function that characterizes the performance is defined and an optimization method is used to find the minimum or maximum of the merit function. Optimization methods can be divided into two groups. The first group performs a global optimization and it includes evolutionary algorithms [151] and Bayesian optimizations [152, 153]. The second group finds the nearest local optimum and the most common methods are given in [67]. Another optimization method which is mainly used for the design of diffractive optical elements is the Iterative Fourier Transform Algorithm (IFTA) [91].

If the number of parameters is large and the optical structure has to be simulated with a rigorous method, most of those optimization methods require too many simulations. Methods that converge fast enough to the optimum are based on the gradient of the merit function such as the gradient descent and the quasi-newton method. Two efficient methods can be used to compute the derivative of the merit function. The first one is the algorithmic differentiation [86, 87], also called automatic differentiation. This method is based on the fact that any computer program is a sequence of operations and, using the chain rule, the derivative of the merit function can be obtained. The second method is the adjoint method [63], which is the topic of this chapter. It allows to get the functional derivative of the merit function with respect to the permettivity and permeability by computing the field generated by two sources, requiring therefore two simulations. The disadvantage of the adjoint method over the algorithmic differentiation is that it gives an approximation of the derivative as the algorithmic differentiation gives the derivative at working precision. On the other side, if the Fourier modal method is used for the simulation of the structure, the adjoint method is expected to be significantly faster than the algorithmic differentiation due to the complexity of the Fourier modal method.

In section 6.4, the equations for the adjoint method are provided for a periodic diffractive element, which is simulated with the Fourier modal method, and for any merit function composed of the efficiency of the diffraction orders. The derivation of those equations are given in section 6.7.2, emphasizing the important steps in the derivation since they need to be derived again if another type of merit function is used. It is also shown in section 6.4 that, at normal incidence and by using the Fourier modal method in chapter 3, a single simulation is required to get the functional derivative of the merit function.

The key elements of the adjoint method are given in two separate sections. The first one, given in section 6.2, is the plane waves generated by a plane source. The adjoint method involves an adjoint source which is described in terms of magnetization and polarization densities in a plane, and the Fourier modal method requires the weight of the incident plane waves instead. The notion of plane source is equivalent to the notion of Generalized Sheet Transition Conditions (GSTC) [154], which describes a discontinuity in the field by a polarization, magnetization and current densities. The GSTC can be used to represent a metasurface by surface susceptibilities [155, 156], by impedance and admittance dyadics [157], or by Hertz potentials [158]. However, many of those works assume that the z-component of the polarization, magnetization and current densities is zero [155, 157, 158], which is not the case in section 6.2. The equations provided in this work are simpler than the ones in [156].

The second key element is the foundation of the adjoint method. As explain in section 6.3, an infinitesimal change in the permettivity and permeability is equivalent to a source, and the impact of those possible sources on the merit function can be obtain by the fields generated by the adjoint source due to the properties of the Green tensor in reciprocal media.

In section 6.5, the adjoint method is applied to a 5x7 beam splitter, which has been introduced in [159]. From this example, the derivative computed using the adjoint method is compared with the numerical derivative. Then, the beam splitter is optimized using the quasi-Newton method, showing that the merit function converges to an optimum even if there is a difference between the derivative from the adjoint method and the numerical derivative. A standard unconstrained optimization is used in section 6.5, meaning that the obtained permittivity is a continuous function. In order to go further and get a diffractive optical element that can be fabricated, the boundaries between the media need to be optimized instead of the permittivty. A typical approach is to introduce a penalty function [160] in order to push the permittivity to one of the two considered materials. Another approach is to compute the effect of an infinitesimal deformation of the boundaries on the figure of merit, called the shape derivative, which can be obtain in a similar way as the functional derivative. The shape derivative requires a description of the boundaries, which can be done through parametric functions or the level set function. The shape derivative for both descriptions is given in [63].

6.2 Plane waves generated by a plane source

In this section, the equations to get the plane waves generated by a plane source and the source that generates a known set of plane waves are provided. Those equations are necessary for the derivation of the adjoint method, but they can also be used in simulations using methods like the Finite Difference Time Domain method (FDTD) in order to design a source, or for the design of the susceptibilities that describe the desired metasurface [155].

The plane waves generated by a plane source can be found by solving the Maxwell equation for a homogeneous medium:

$$\begin{aligned} \nabla \times \vec{E} &= i k_0 (\mu \vec{H} + \vec{M}) \\ \nabla \times \vec{H} &= -i k_0 (\epsilon \vec{E} + \vec{P}), \end{aligned} \tag{6.1}$$

where the magnetization $\vec{M}$ and the polarization density $\vec{P}$ describe a plane source at $z = 0$, which are given by

$$\vec{M} = \begin{pmatrix} M_x \\ M_y \\ M_z \end{pmatrix} e^{i(k_x x + k_y y)} \delta(z) \qquad \vec{P} = \begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} e^{i(k_x x + k_y y)} \delta(z). \tag{6.2}$$

$\delta(z)$ is the Dirac function. Through the Fourier transform, any plane source with varying amplitude and phase along the xy-plane can be described as a sum of the plane sources expressed in (6.2) with different k_x and k_y.

As proved in section 6.7.1, the fields generated by a plane source have the form

$$\begin{aligned} \vec{E} &= A_1 u(z) \vec{E}_{TM} + A_2 (1 - u(z)) \vec{E}_{TM}^- + B_1 u(z) \vec{E}_{TE} + B_2 (1 - u(z)) \vec{E}_{TE}^- + C_1 \vec{n}_z e^{i(k_x x + k_y y)} \delta(z) \\ \vec{H} &= A_1 u(z) \vec{H}_{TM} + A_2 (1 - u(z)) \vec{H}_{TM}^- + B_1 u(z) \vec{H}_{TE} + B_2 (1 - u(z)) \vec{H}_{TE}^- + C_2 \vec{n}_z e^{i(k_x x + k_y y)} \delta(z). \end{aligned} \tag{6.3}$$

where $u(z)$ is the unit step function, $\vec{n}_z$ is the unit vector parallel to the z-axis and $\vec{E}_{TM}$ and $\vec{E}_{TE}$ are the fields of respectively the TM and TE-polarized plane waves given in section 3.2.1. As a reminder, the TM-polarized plane waves are described as

$$\vec{E}_{TM} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -k_{\parallel} \end{pmatrix} e^{i(k_x x + k_y y + k_z z)}, \qquad \vec{H}_{TM} = \begin{pmatrix} -\epsilon k_0 s_y \\ \epsilon k_0 s_x \\ 0 \end{pmatrix} e^{i(k_x x + k_y y + k_z z)} \tag{6.4}$$

and the TE-polarized plane waves are described as

$$\vec{E}_{TE} = \begin{pmatrix} \mu k_0 s_y \\ -\mu k_0 s_x \\ 0 \end{pmatrix} e^{i(k_x x + k_y y + k_z z)}, \qquad \vec{H}_{TE} = \begin{pmatrix} k_z s_x \\ k_z s_y \\ -k_{\parallel} \end{pmatrix} e^{i(k_x x + k_y y + k_z z)} \tag{6.5}$$

where k_z is obtained from the dispersion relation

$$k_z^2 + k_\parallel^2 = \epsilon\mu k_0^2 \tag{6.6}$$

and $k_\parallel$, s_x and s_y are given by

$$k_\parallel = \sqrt{k_x^2 + k_y^2}, \qquad \begin{cases} s_x = 1,\ s_y = 0 & \text{if } k_\parallel = 0 \\ s_x = k_x/k_\parallel,\ s_y = k_y/k_\parallel & \text{otherwise} \end{cases} \tag{6.7}$$

In order to find the plane waves generated by a plane source instead of the plane waves that are fully absorbed by a plane sink, the sign of k_z is chosen such that the power flow related to the forward-propagating plane waves described in (6.4) and (6.5) is toward the positive z-direction and the imaginary part of k_z is positive. If ϵ and μ describe an active medium, the choice of the sign of k_z is more problematic [122].

The operator $(\cdot)^-$ used in (6.3) is defined in section 2.3 and transforms a forward-propagating plane wave to a backward-propagating plane wave in the following way:

$$\begin{aligned} \vec{E} &= \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} e^{i(k_x x + k_y y + k_z z)} \quad \Rightarrow \quad \vec{E}^- = \begin{pmatrix} E_x \\ E_y \\ -E_z \end{pmatrix} e^{i(k_x x + k_y y - k_z z)} \\ \vec{H} &= \begin{pmatrix} H_x \\ H_y \\ H_z \end{pmatrix} e^{i(k_x x + k_y y + k_z z)} \quad \Rightarrow \quad \vec{H}^- = \begin{pmatrix} -H_x \\ -H_y \\ H_z \end{pmatrix} e^{i(k_x x + k_y y - k_z z)}. \end{aligned} \tag{6.8}$$

The coefficients A_1, A_2, B_1, B_2, C_1 and C_2 in equation (6.3) are given by

$$\begin{aligned} A_1 &= \frac{i}{2\epsilon k_z}(k_z(s_x P_x + s_y P_y) - k_\parallel P_z + \epsilon k_0(s_x M_y - s_y M_x)) \\ A_2 &= \frac{i}{2\epsilon k_z}(k_z(s_x P_x + s_y P_y) + k_\parallel P_z - \epsilon k_0(s_x M_y - s_y M_x)) \\ B_1 &= \frac{i}{2\mu k_z}(k_z(s_x M_x + s_y M_y) - k_\parallel M_z - \mu k_0(s_x P_y - s_y P_x)) \\ B_2 &= -\frac{i}{2\mu k_z}(k_z(s_x M_x + s_y M_y) + k_\parallel M_z + \mu k_0(s_x P_y - s_y P_x)) \\ C_1 &= -\frac{P_z}{\epsilon} \\ C_2 &= -\frac{M_z}{\mu}. \end{aligned} \tag{6.9}$$

If one wants to find the source that generates a known set of plane waves, the solution is given by

$$
\begin{aligned}
M_x &= \frac{1}{k_0}(ik_z s_y(A_1 - A_2) - i\mu k_0 s_x(B_1 - B_2) + k_y C_1) \\
M_y &= -\frac{1}{k_0}(ik_z s_x(A_1 - A_2) + i\mu k_0 s_y(B_1 - B_2) + k_x C_1) \\
M_z &= -\mu C_2 \\
P_x &= -\frac{1}{k_0}(i\epsilon k_0 s_x(A_1 + A_2) + ik_z s_y(B_1 + B_2) + k_y C_2) \\
P_y &= -\frac{1}{k_0}(i\epsilon k_0 s_y(A_1 + A_2) - ik_z s_x(B_1 + B_2) - k_x C_2) \\
P_z &= -\epsilon C_1
\end{aligned}
\tag{6.10}
$$

The coefficients of the bound modes C_1 and C_2 can be chosen arbitrarily.

6.3 Variation of the fields due to a change in material parameters and reciprocity

The adjoint method is based on the concept of reciprocity and on the equivalence between an infinitesimal difference in the permittivity and permeability, and a source described by a polarization and magnetization density $\vec{P}$ and $\vec{M}$. The fields generated by this source are the variation of the electric and magnetic fields, $\delta\vec{E}(\vec{x})$ and $\delta\vec{H}(\vec{x})$, due to a variation of the material parameters. If the medium is reciprocal, there is a link between the fields at the position $\vec{x}'$ due to a source at position $\vec{x}$ and the fields at the position $\vec{x}$ due to a source at position $\vec{x}'$. The adjoint method uses this link to get the effect of the variation of the material parameters, which is equivalent to a source, on the figure of merit by computing the fields at the position where the material parameters vary due to a virtual source called the adjoint source. In this section, it is shown why an infinitesimal difference in the material parameters is equivalent to a source. Then, the properties of the Green tensor of a reciprocal medium, which are required in the derivation of the adjoint method, are given.

Initially, the electric field $\vec{E}(\vec{x})$ and magnetic field $\vec{H}(\vec{x})$ satisfy the Maxwell equation

$$
\begin{aligned}
\nabla \times \vec{E}(\vec{x}) &= ik_0\mu(\vec{x})\vec{H}(\vec{x}) \\
\nabla \times \vec{H}(\vec{x}) &= -ik_0\epsilon(\vec{x})\vec{E}(\vec{x}).
\end{aligned}
\tag{6.11}
$$

For an infinitesimal variation of the permittivity and permeability given by $\delta\epsilon(\vec{x})$ and $\delta\mu(\vec{x})$, the Maxwell equation (6.11) becomes

$$\begin{aligned}\nabla\times(\vec{E}(\vec{x})+\delta\vec{E}(\vec{x})) &= ik_0(\mu(\vec{x})+\delta\mu(\vec{x}))(\vec{H}(\vec{x})+\delta\vec{H}(\vec{x}))\\ \nabla\times(\vec{H}(\vec{x})+\delta\vec{H}(\vec{x})) &= -ik_0(\epsilon(\vec{x})+\delta\epsilon(\vec{x}))(\vec{E}(\vec{x})+\delta\vec{E}(\vec{x})),\end{aligned} \tag{6.12}$$

where $\delta\vec{E}(\vec{x})$ and $\delta\vec{H}(\vec{x})$ are the infinitesimal change of respectively the electric and magnetic fields due to $\delta\epsilon(\vec{x})$ and $\delta\mu(\vec{x})$.

Assuming that the terms $\delta\mu(\vec{x})\delta\vec{H}(\vec{x})$ and $\delta\epsilon(\vec{x})\delta\vec{E}(\vec{x})$ are negligible and recognizing equation (6.11) in equation (6.12), equation (6.12) becomes

$$\begin{aligned}\nabla\times\delta\vec{E}(\vec{x}) &= ik_0(\mu(\vec{x})\delta\vec{H}(\vec{x})+\delta\mu(\vec{x})\vec{H}(\vec{x}))\\ \nabla\times\delta\vec{H}(\vec{x}) &= -ik_0(\epsilon(\vec{x})\delta\vec{E}(\vec{x})+\delta\epsilon(\vec{x})\vec{E}(\vec{x})).\end{aligned} \tag{6.13}$$

The infinitesimal change of the fields can be seen as the solution of the Maxwell equation with a polarization density $\vec{P}$ and a magnetization density $\vec{M}$, which constitute a source, and are given by:

$$\vec{P}=\delta\epsilon(\vec{x})\vec{E}(\vec{x}) \qquad \vec{M}=\delta\mu(\vec{x})\vec{H}(\vec{x}). \tag{6.14}$$

For shape optimization, meaning that the boundary between two homogeneous media is optimized instead of the permettivity and permeability, the terms $\delta\mu(\vec{x})\delta\vec{H}(\vec{x})$ and $\delta\epsilon(\vec{x})\delta\vec{E}(\vec{x})$ in (6.12) cannot be neglected. This case is not considered in this work, but it is treated in section 5.1 in [63].

By introducing the Green tensors, $\delta\vec{E}(\vec{x})$ and $\delta\vec{H}(\vec{x})$ are given by:

$$\begin{pmatrix}\delta\vec{E}(\vec{x}')\\ \delta\vec{H}(\vec{x}')\end{pmatrix}=\iiint\begin{pmatrix}\hat{G}_{EP}(\vec{x}',\vec{x}) & \hat{G}_{EM}(\vec{x}',\vec{x})\\ \hat{G}_{HP}(\vec{x}',\vec{x}) & \hat{G}_{HM}(\vec{x}',\vec{x}))\end{pmatrix}\begin{pmatrix}\delta\epsilon(\vec{x})\vec{E}(\vec{x})\\ \delta\mu(\vec{x})\vec{H}(\vec{x})\end{pmatrix}d\vec{x} \tag{6.15}$$

In reciprocal medium, the Green tensor has the following properties:

$$\begin{aligned}\hat{G}_{EP}(\vec{x}',\vec{x}) &= \hat{G}^T_{EP}(\vec{x},\vec{x}')\\ \hat{G}_{HM}(\vec{x}',\vec{x}) &= \hat{G}^T_{HM}(\vec{x},\vec{x}')\\ \hat{G}_{HP}(\vec{x}',\vec{x}) &= -\hat{G}^T_{EM}(\vec{x},\vec{x}').\end{aligned} \tag{6.16}$$

Those properties are proved in Appendix A of [63].

6.4 Adjoint method for periodic diffractive optical elements

The adjoint method gives the functional derivatives based on the fields generated by two different sources, the primary source and an adjoint source. Hence, a first equation describes the adjoint source and a second equation gives the functional derivatives. Those two equations depend of the definition of the merit function. In this section, the adjoint method is applied to a system composed of a periodic diffractive optical element with a real-valued figure of merit F based on the efficiency of the diffraction orders, denoted η_m, and the adjoint source is expressed in such way that it is suitable for the Fourier modal method (chapter 3). The subscript m refers to m-th order, which is a plane wave characterized by the components of the k-vector $k_{x,m}$ and $k_{y,m}$, and a polarization, which can be either TM or TE.

Since the adjoint method uses two sources, two simulations, one for each source, are usually required: the direct simulation and the adjoint simulation. Figure 6.1a is a schema representing the direct simulation, where the primary source generates the incident plane wave that interacts with the diffractive optical element described by the permittivity $\epsilon(\vec{x})$ and the permeability $\mu(\vec{x})$. The figure of merit is obtained from the weight of the transmitted plane waves t_m at the output plane. The adjoint simulation is schematized in fig. 6.1b, where the adjoint source is defined at the output plane by the magnetization and polarization densities $\vec{M}(\vec{x}')$ and $\vec{P}(\vec{x}')$. This adjoint source emits a set of plane waves whose weight is q_m and generates the adjoint fields $\vec{E}_{adj}(\vec{x})$ and $\vec{H}_{adj}(\vec{x})$. By convention, $\vec{x}$ is a position in the region where the permettivity and permeability are optimized, and $\vec{x}'$ is a position in the output plane. In fig. 6.1, the subscripts of the weights t_m and q_m are not in the same order because, due to the convention used in this work, the only difference between the k-vector of a forward and backward-propagating plane wave related to the order m is the change in the sign of k_z.

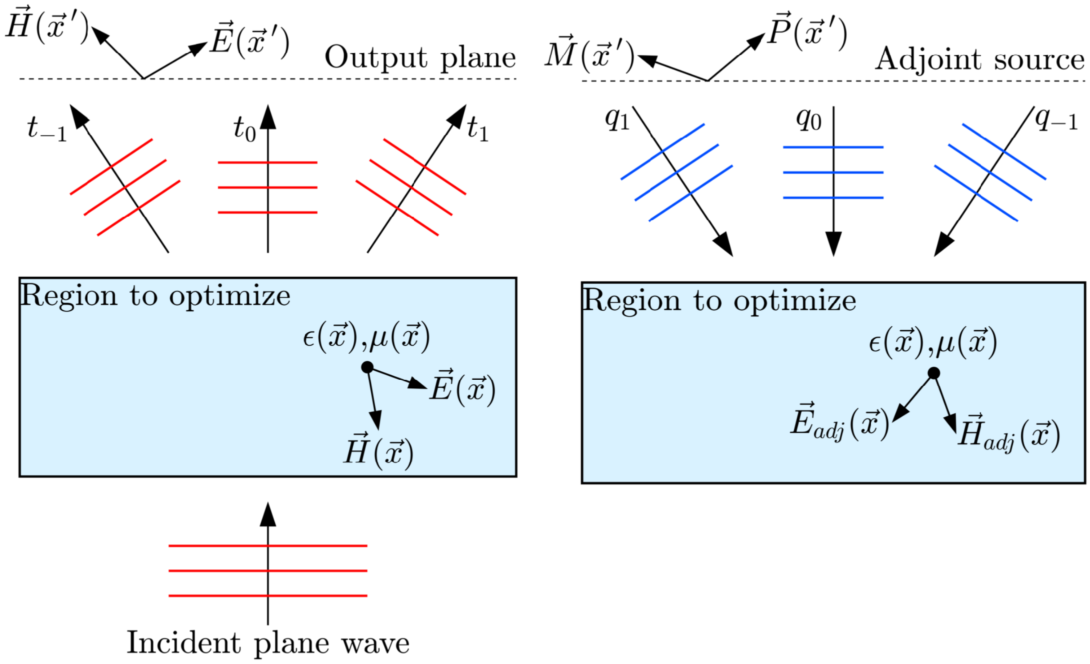
In order to use the adjoint method, the merit function and therefore the diffraction efficiencies η_m need to be expressed in terms of the fields. It can be done in multiple ways and three of them based on the integration of the fields in the output plane are given:

$$\eta_m = \frac{\Phi_m}{\Phi_{in}} \left| \frac{1}{p_m \mu k_0^2 |\Lambda|} \iint_\Lambda \vec{E}(\vec{x}') \cdot \vec{E}_m^*(\vec{x}') d\vec{x}' \right|^2 \tag{6.17a}$$

$$\eta_m = \frac{\Phi_m}{\Phi_{in}} \left| \frac{1}{2 p_m k_0 k_{z,m} |\Lambda|} \iint_\Lambda (\vec{E}(\vec{x}') \times \vec{H}_p^-(\vec{x}') - \vec{E}_p^-(\vec{x}') \times \vec{H}(\vec{x}')) \cdot \vec{n}_z d\vec{x}' \right|^2 \tag{6.17b}$$

$$\eta_m = \frac{\Phi_m}{\Phi_{in}} \left| \frac{1}{2 p_m k_0 k_{z,m} |\Lambda|} \iint_\Lambda (\vec{E}(\vec{x}') \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \vec{H}(\vec{x}')) \cdot \vec{n}_z d\vec{x}' \right|^2, \tag{6.17c}$$

where $\vec{E}(\vec{x}')$ and $\vec{H}(\vec{x}')$ are the fields generated by the primary source, Λ is the unit cell, $|\Lambda|$ is the unit cell area, p_m is ϵ if the diffraction order m is TE-polarized and μ otherwise, $\vec{E}_m(\vec{x}')$ and $\vec{H}_m(\vec{x}')$ are the fields of the order m given by (6.4) or (6.5) depending of the polarization, and Φ_m/Φ_{in} is the ratio of the power flow of the plane wave described by $\vec{E}_m(\vec{x}')$ and $\vec{H}_m(\vec{x}')$ to the



(a) (b)

Figure 6.1 – a) Direct simulation where the diffractive optical element, which is the region to optimize, is illuminated by a plane wave. The figure of merit is computed from the fields at the output plane, which are composed of the plane waves related to the diffraction orders. In the adjoint method, the required quantities are the weight t_m of the transmitted plane waves at the output plane, and the fields $\vec{E}(\vec{x})$ and $\vec{H}(\vec{x})$ inside the region to optimize. b) Adjoint simulation where the adjoint source defined at the output plane by the magnetization and polarization densities $\vec{M}(\vec{x}')$ and $\vec{P}(\vec{x}')$, emits a set of plane waves with the weights q_m propagating toward the region to optimize. In the adjoint method, the required quantities are the fields $\vec{E}_{adj}(\vec{x})$ and $\vec{H}_{adj}(\vec{x})$ in the region to optimize.

power flow of the incident light. In equation (6.17b), for a diffraction order m described by $k_{x,m}$ and $k_{y,m}$, the order p is described by $-k_{x,m}$ and $-k_{y,m}$ with the same polarization. The three different expressions of the diffraction efficiency η_m in (6.17) are only valid for lossless medium and if the order m is propagating.

The integrals in (6.17) act as filter, whose output is the weight of the plane wave described by $\vec{E}_m(\vec{x}')$ and $\vec{H}_m(\vec{x}')$ that composes the fields $\vec{E}(\vec{x}')$ and $\vec{H}(\vec{x}')$. Hence, the expressions in (6.17) are equivalent to

$$\eta_m = \frac{\Phi_m}{\Phi_{in}} |t_m|^2. \tag{6.18}$$

Since equations (6.17) express the same quantity, the adjoint method gives the same expression of the functional derivatives even if their derivation is different. The expression (6.17a) is the simplest one, but it fails if the fields at the output plane generated by a source in the region to optimize, are composed of forward and backward-propagating waves, which is rarely the case in practice. The expression (6.17b) is the one used in [89] and the expression (6.17c) is based on the Poynting operation (chapter 2). For derivation of the adjoint given in section 6.7.2, the expression (6.17c) is chosen.

As proved in section 6.7.2, the weight of the plane wave related to the diffraction order p emitted by the adjoint source are given by

$$q_p = c_m \frac{\delta F}{\delta \eta_m} t_m^* \qquad c_m = \begin{cases} 1 & \text{for } k_{\parallel,m} = 0 \\ -1 & \text{otherwise} \end{cases}, \tag{6.19}$$

where F is the figure of merit and $k_{\parallel,m}$ is given by $\sqrt{k_{x,m}^2 + k_{y,m}^2}$.

The Bloch phases of the direct and adjoint simulation have different sign. Therefore, if the illumination for the direct simulation is at normal incidence, meaning that the Bloch phase is zero, and the Fourier modal method in chapter 3 is used, the S-matrix obtained from the direct simulation that describes the diffractive optical element, can also be used for the adjoint simulation. Hence, a single simulation is enough to get the functional derivatives of the figure of merit.

The functional derivatives of the figure of merit depend of the fields generated by the primary source, $\vec{E}(\vec{x})$ and $\vec{H}(\vec{x})$, and the fields generated by the adjoint source, $\vec{E}_{adj}(\vec{x})$ and $\vec{H}_{adj}(\vec{x})$. They are given by

$$\begin{aligned}\frac{\delta F}{\delta\epsilon(\vec{x})} &= -\frac{k_0}{|\Lambda|\Phi_{in}}\operatorname{Im}\left\{\vec{E}_{adj}(\vec{x})\cdot\vec{E}(\vec{x})\right\}\\ \frac{\delta F}{\delta\mu(\vec{x})} &= \frac{k_0}{|\Lambda|\Phi_{in}}\operatorname{Im}\left\{\vec{H}_{adj}(\vec{x})\cdot\vec{H}(\vec{x})\right\}\end{aligned}\tag{6.20}$$

For a variation of the permittivity and permeability given by $\delta\epsilon(\vec{x})$ and $\delta\mu(\vec{x})$, the variation of the figure of merit δF is

$$\delta F = \iiint \frac{\delta F}{\delta\epsilon(\vec{x})}\delta\epsilon(\vec{x})d\vec{x} + \iiint \frac{\delta F}{\delta\mu(\vec{x})}\delta\mu(\vec{x})d\vec{x}\tag{6.21}$$

In the Fourier modal method, the diffractive optical element is divided into layers and each layer is described by a permittivity profile, assuming here for simplicity that the permeability is always one. Moreover, the permittivity profile is usually divided into pixels. In other words, the layer is divided into cuboids where the permittivity is assumed constant inside and the fields depend only on z. In this case, the variation of the merit function due to a change of the permittivity of the cuboid n, denoted $\delta\epsilon_n$, is

$$\delta F = -\frac{|A_p|k_0\delta\epsilon_n}{|\Lambda|\Phi_{in}}\int_{z_1}^{z_2}\operatorname{Im}\left\{\vec{E}_{adj,n}(z)\cdot\vec{E}_n(z)\right\}dz,\tag{6.22}$$

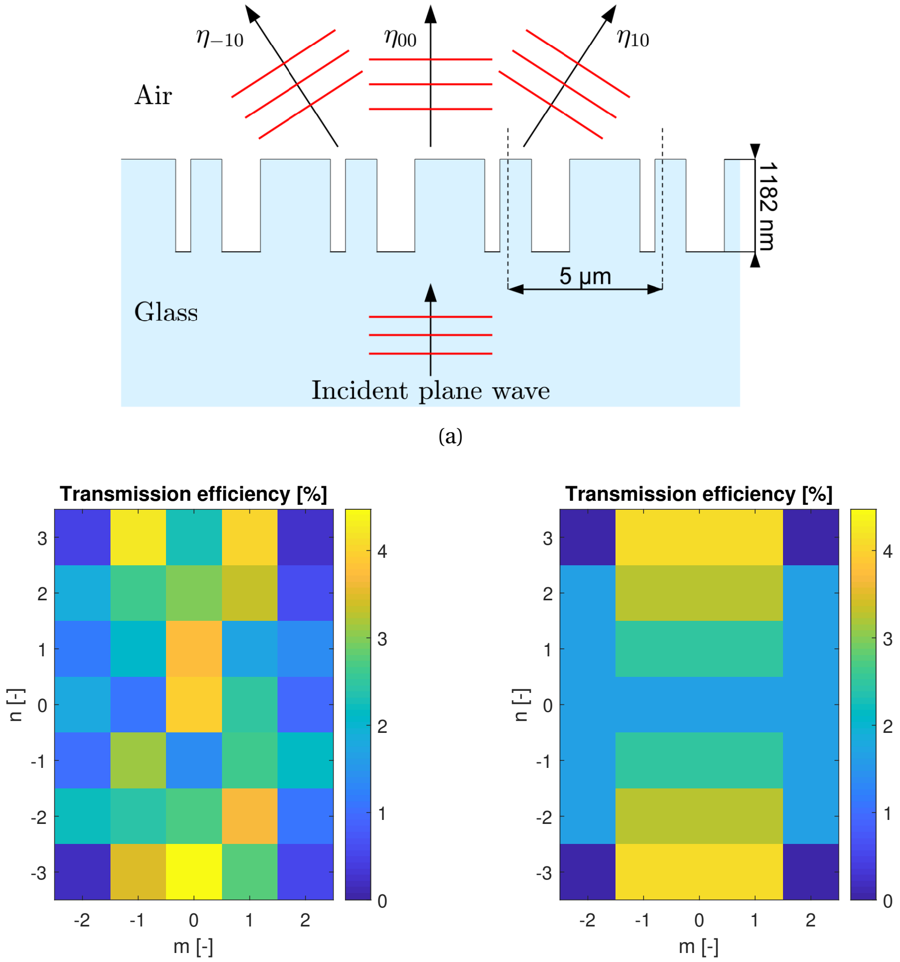
where $|A_p|$ is the area of the pixel, z_1 and z_2 are the position where the interfaces that delimite the layer are, and $\vec{E}_{adj,n}(z)$ and $\vec{E}_n(z)$ are the electric fields in the cuboid n.

The derivation of equations (6.17c), (6.19), and (6.20) is given in section 6.4. For another merit function, section 6.4 also provides the necessary steps in order to get the required equations for the computation of the functional derivative.

6.5 Application of the adjoint method for the design of a beam splitter

In this section, the adjoint method is applied on a 5x7 beam splitter, which has been introduced in [159]. It is a binary diffractive optical element composed of glass ($n = 1.45$) with a square lattice working at a wavelength of 940 nm under y-polarized illumination at normal incidence. The lattice constant and the thickness of the beam splitter are fixed to 5 μm and 1182 nm respectively. A schema of the beam splitter is given in fig. 6.2a. Due to the lattice constant, the maximum diffraction angles are 22.1° along x and 34.3° along y. The thickness has been chosen in order to have a difference in the phase after the beam splitter of around π under the thin-element approximation.

The initial beam splitter, which is used as the starting point for the optimization, is given in



(a)

(b)

(c)

Figure 6.2 – a) Schema of the beam splitter, which is illuminated from the glass substrate at normal incidence. The lattice constant of the beam splitter is 5 µm and the thickness is 1182 nm. η_{mn} is the efficiency of the diffraction order (m,n). b) Diffraction efficiencies of the beam splitter before optimization. Their values are given in table 6.5a. c) Desired diffraction efficiencies. Except at the corner ($(m,n) = (\pm2,\pm3)$), the difference between the desired efficiencies and the diffraction efficiencies of the optimized beam splitter is negligible. Hence, their values are shown in 6.5b.

fig. 6.3d and it has been obtained using the Iterative Fourier Transform Algorithm (IFTA) [91]. The desired output, which is determined by the diffraction efficiencies, is shown in fig. 6.2c and the output of the initial beam splitter is shown in fig. 6.2b and table 6.5a.

Two metrics are used to characterize the performance of the beam splitter. The first one is the sum of the errors squared of the efficiencies, and is given by

$$F = \sum_{mn} (\eta_{mn} - \eta_{d,mn})^2, \tag{6.23}$$

where η_{mn} and $\eta_{d,mn}$ is respectively the obtained and desired efficiency of the diffraction order (m,n). The metric F is the figure of merit. The angle θ_x and θ_y of the diffraction order (m,n) is

$$\theta_x = \operatorname{asin}^{-1}\left(\frac{L}{\lambda}\right) \qquad \theta_y = \operatorname{asin}^{-1}\left(\frac{L}{\lambda}\right), \tag{6.24}$$

where L is the lattice constant and λ is the wavelength.

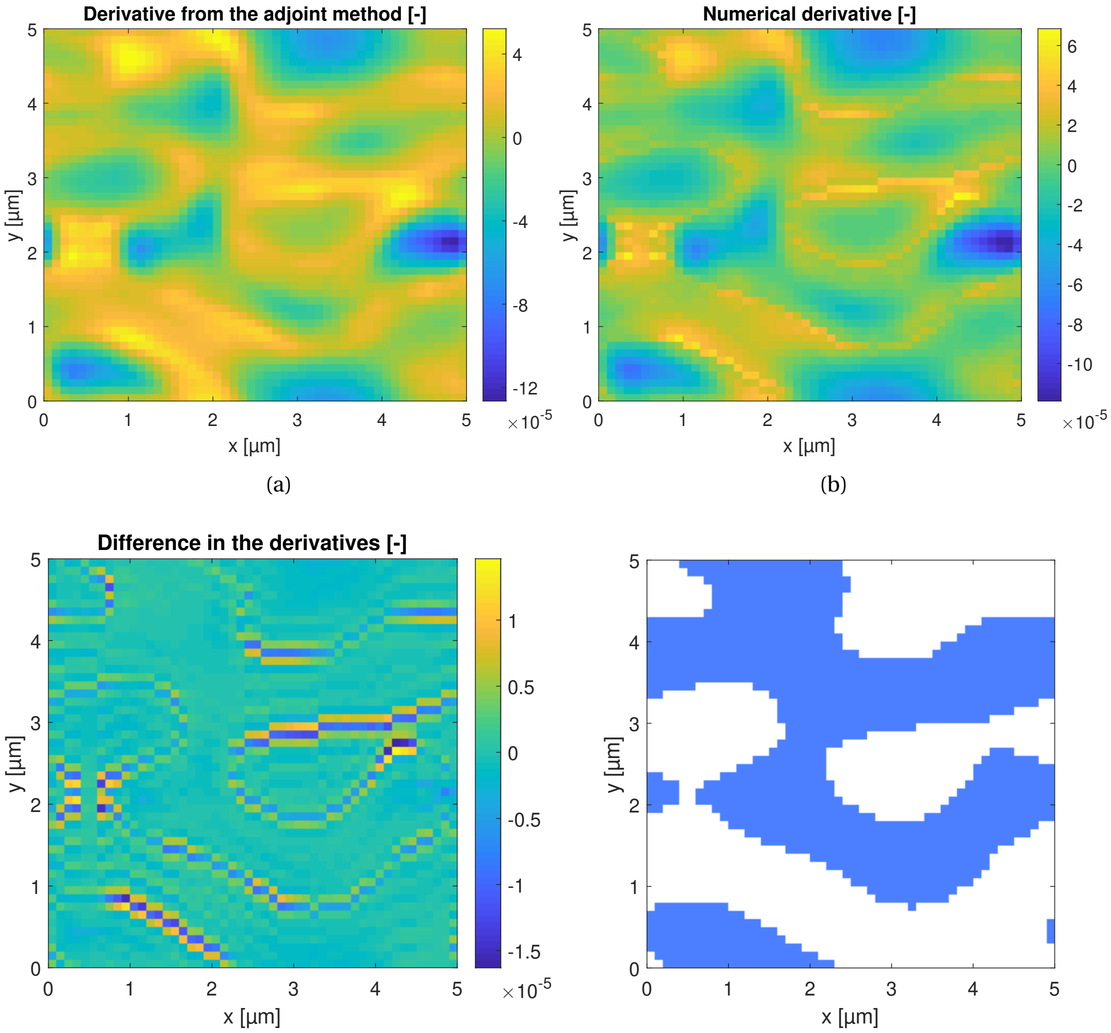
The second metric is the uniformity error and is defined as

$$UE = \frac{\max\limits_{m,n}(\eta_{mn}/\eta_{d,mn}) - \min\limits_{m,n}(\eta_{mn}/\eta_{d,mn})}{\max\limits_{m,n}(\eta_{mn}/\eta_{d,mn}) + \min\limits_{m,n}(\eta_{mn}/\eta_{d,mn})}. \tag{6.25}$$

The orders (m,n) taken into account in equations (6.23) and (6.25) are the orders whose efficiency $\eta_{d,mn}$ is shown in fig. 6.2c and different than zero. The sum of the desired diffraction efficiencies $\eta_{d,mn}$ is 80%.

The adjoint method gives the estimation of the derivative of the figure of merit, which is an estimation of the effect of the variation of the permittivity of a cuboid, whose cross-section is a pixel present fig. 6.3d, on the merit function. Since the permittivity profile is defined by a matrix, the figure of merit is derivated with respect of the permittivity of the cuboid that composes the beam splitter. In order to obtain the derivative from the adjoint method, equations (6.19) and (6.22) are used. The integral in (6.22) can be solved numerically, but, since the fields of the eigen-modes obtained from the Fourier modal method are given for any position z, it is possible to solve this integral analytically.

The estimation of the derivative obtained from the adjoint method is given in fig. 6.3a. This estimation is compared with the numerical derivative shown in fig. 6.3b, which is obtained by computing the figure of merit after varying slightly the permittivity of a cuboid. By comparing figs. 6.3a and 6.3b, both derivatives have the same order of magnitude and the same features with the difference that the features in the numerical derivative are sharper. From fig. 6.3c, which is the difference between the derivative from the adjoint method and the numerical

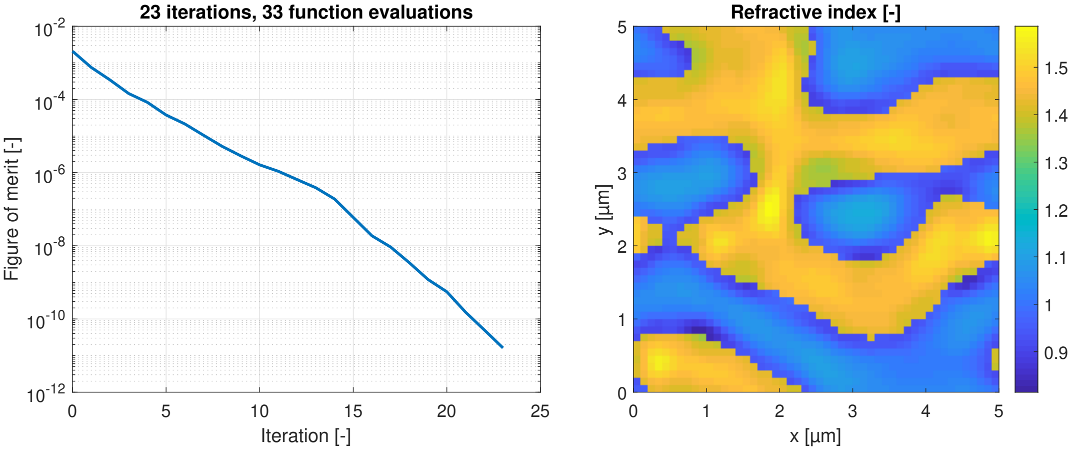


(a) (b)

(c) (d)

Figure 6.3 – a) Estimation of the derivative of the figure of merit with respect to the permittivity obtained from the adjoint method. b) Numerical derivative of the figure of merit. c) Difference between figs. 6.3a and 6.3b. d) Permettivity profile of the beam splitter before optimization. The blue region is glass ($n = 1.45$) and the white region is air.

derivative, the largest differences are at the boundary between glass and air.



(a)

(b)

Figure 6.4 – a) Figure of merit for the different iterations. The number of iterations is lower than the number of function evaluations because an iteration includes a line search and the line search may require more than one function evaluation. The function evaluation includes the computation of the figure merit and its derivative. b) Refractive index profile of the optimized beam splitter.

In order to optimize the beam splitter, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, which is a variation of the quasi-newton method, combined with a line search method based on polynomial interpolation (section 2.6 and 3.2 in [161]) is used. As shown in fig. 6.4a, the optimization process converges rapidly to a small value, even if the derivative of the merit function given by the adjoint method is an estimation. Before the optimization, the transmission efficiencies are given in table 6.5a and, from those values, the figure of merit (6.23) and the uniformity error (6.25) are $2.08 \cdot 10^{-3}$ and 73.9% respectively. After optimization, the refractive index profile is shown in fig. 6.4b and the transmission efficiencies, given in table 6.5b, are virtually the same as the target diffraction efficiencies. From the resulting efficiencies, the figure of merit (6.23) and the uniformity error (6.25) are $1.62 \cdot 10^{-11}$ and 0.009% respectively.

The optimized beam splitter cannot be made because its permittivity profile (fig. 6.4b) is continuous and some regions have a refractive index below one due to the lack of constraints in the optimization. However, the results obtained in this section show that the derivative of the figure of merit given by the adjoint method is close enough to the actual derivative and, therefore, the adjoint method can be used in conjunction with a gradient-based optimization algorithm in order to optimize an optical structure. Further work on this topic will be provided by D. C. Kim [159].

n \ m	-2	-1	0	1	2
-3	0.17%	3.45%	4.47%	2.78%	0.55%
-2	2.22%	2.42%	2.67%	3.67%	1.10%
-1	1.02%	3.14%	1.39%	2.59%	2.12%
0	1.77%	1.11%	3.98%	2.47%	0.96%
1	1.13%	2.08%	3.75%	1.74%	1.39%
2	1.86%	2.60%	2.96%	3.31%	0.60%
3	0.44%	4.25%	2.29%	4.00%	0.23%

(a)

n \ m	-2	-1	0	1	2
-3	0.08%	4.08%	4.08%	4.08%	0.24%
-2	1.63%	3.27%	3.27%	3.27%	1.63%
-1	1.63%	2.45%	2.45%	2.45%	1.63%
0	1.63%	1.63%	1.63%	1.63%	1.63%
1	1.63%	2.45%	2.45%	2.45%	1.63%
2	1.63%	3.27%	3.27%	3.27%	1.63%
3	0.27%	4.08%	4.08%	4.08%	0.05%

(b)

Figure 6.5 – a) Diffraction efficiencies of the beam splitter before optimization. b) Diffraction efficiencies of the optimized beam splitter.

6.6 Conclusion

In this chapter, we provide the equations for the adjoint method applied to diffractive optical elements for any merit function based on the efficiency of the diffraction orders. For another merit function, a new set of equations has to be derived and, in order to facilitate the derivation, we emphasize all the steps of the proof.

The adjoint method is adapted to be used with the Fourier modal method, which gives the following advantage. Usually, two simulations of the diffractive optical element are necessary: one with the primary source and the second with the adjoint source. However, at normal incidence, only one simulation is required to obtain the figure of merit and the derivative of the merit function due to the relationship between the Bloch phase of both sources.

In the derivation of the adjoint method, the adjoint source is described in terms of a magnetization and a polarization density in a plane, but, in the Fourier modal method, the weight of the incident plane waves is required. In order to obtain those weights, we provide the relationship between a plane source, described by a magnetization and a polarization density, and the weight of the plane waves generated by this plane source. The notion of plane source developed here seems promising to find the surface susceptibilities for a desired functionality.

As a proof of concept, we optimize a 5x7 beam splitter with a quasi-Newton method in conjunction with the adjoint method. From this example, we show that the gradient obtained by the adjoint method is a smoothed version of the numerical derivative. We demonstrate that, despite this difference, the quasi-Newton method converges quickly to a small value, making the adjoint method suitable for calculating the gradient of the merit function.

6.7 Proofs

6.7.1 Proof of the plane waves generated by a plane source

The plane source described in (6.2) have six variables, M_x, M_y, M_z, P_x, P_y and P_z, and, due to the linearity of the Maxwell equation (6.1), its solutions are the sum of the solutions of the Maxwell equation without source, which are the homogeneous solutions, and a linear combination of six particular solutions, that are not solution of the Maxwell equation without source and whose coefficients depend of the six variable in (6.2). As a reminder, the Maxwell equation (6.1) is

$$\begin{aligned} \nabla \times \vec{E} &= i k_0 (\mu \vec{H} + \vec{M}) \\ \nabla \times \vec{H} &= -i k_0 (\epsilon \vec{E} + \vec{P}), \end{aligned} \tag{6.26}$$

where the source is given by

$$\vec{M} = \begin{pmatrix} M_x \\ M_y \\ M_z \end{pmatrix} e^{i(k_x x + k_y y)} \delta(z) \qquad \vec{P} = \begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} e^{i(k_x x + k_y y)} \delta(z). \tag{6.27}$$

By looking at this equation, the obvious candidates for the particular solutions contain the Dirac function $\delta(z)$ or the unit step function $u(z)$ since $\delta(z)$ appears in the source, $u(z)$ is the derivative of $\delta(z)$ and such candidates cannot be a homogeneous solution of the Maxwell equation (6.26). Moreover, the candidates are assumed to have the same spatial dependency $e^{i(k_x x + k_y y)}$ as the plane source. Hence, the proposed candidates are in the form

$$\vec{E} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} e^{i(k_x x + k_y y)} \delta(z) \qquad \vec{H} = \begin{pmatrix} H_x \\ H_y \\ H_z \end{pmatrix} e^{i(k_x x + k_y y)} \delta(z) \tag{6.28}$$

and

$$\vec{E} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} e^{i(k_x x + k_y y + \gamma z)} u(z) \qquad \vec{H} = \begin{pmatrix} H_x \\ H_y \\ H_z \end{pmatrix} e^{i(k_x x + k_y y + \gamma z)} u(z), \tag{6.29}$$

where E_x, E_y, E_z, H_x, H_y, H_z and γ are unknowns. The particular solution (6.28) cannot be derivated with respect to z since the term $\delta'(z)$ appears. Therefore, E_x, E_y, H_x and H_y are zero. To fit the solution proposed in (6.3), E_z and H_z are replaced by C_1 and C_2 respectively. Inserting the particular solution (6.28) into the Maxwell equation (6.26) gives:

$$\begin{aligned} k_y C_1 &= k_0 M_x & \qquad -k_y C_2 &= k_0 P_x \\ -k_x C_1 &= k_0 M_y & k_x C_2 &= k_0 P_y \\ -\mu k_0 C_2 &= k_0 M_z & -\epsilon k_0 C_1 &= k_0 P_z. \end{aligned} \tag{6.30}$$

Inserting the particular solution (6.29) into the Maxwell equation (6.26) gives

$$\begin{aligned} \begin{pmatrix} k_x \\ k_y \\ \gamma \end{pmatrix} \times \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} u(z) - i \begin{pmatrix} -E_y \\ E_x \\ 0 \end{pmatrix} \delta(z) &= \mu k_0 \begin{pmatrix} H_x \\ H_y \\ H_z \end{pmatrix} u(z) + k_0 \begin{pmatrix} M_x \\ M_y \\ M_z \end{pmatrix} \delta(z) \\ - \begin{pmatrix} k_x \\ k_y \\ \gamma \end{pmatrix} \times \begin{pmatrix} H_x \\ H_y \\ H_z \end{pmatrix} u(z) + i \begin{pmatrix} -H_y \\ H_x \\ 0 \end{pmatrix} \delta(z) &= \epsilon k_0 \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} u(z) + k_0 \begin{pmatrix} P_x \\ P_y \\ P_z \end{pmatrix} \delta(z) \end{aligned} \tag{6.31}$$

By taking only the terms containing the unit step function $u(z)$, the Maxwell equation without source are obtained, meaning that the TM and TE plane waves described in (6.4) and (6.5) are solutions for both propagation directions and γ is equal to $\pm k_z$.

The weight of the four different plane waves is obtained from the equations composed of the terms containing the Dirac function $\delta(z)$. Therefore, the particular solution of the Maxwell equation (6.26) has the form

$$\begin{aligned} \vec{E} &= A_1 u(z) \vec{E}_{TM} - A_2 u(z) \vec{E}_{TM}^{-} + B_1 u(z) \vec{E}_{TE} - B_2 u(z) \vec{E}_{TE}^{-} + C_1 \vec{n}_z e^{i(k_x x + k_y y)} \delta(z) \\ \vec{H} &= A_1 u(z) \vec{H}_{TM} - A_2 u(z) \vec{H}_{TM}^{-} + B_1 u(z) \vec{H}_{TE} - B_2 u(z) \vec{H}_{TE}^{-} + C_2 \vec{n}_z e^{i(k_x x + k_y y)} \delta(z). \end{aligned} \tag{6.32}$$

The solution given in (6.3) is a homogeneous solution of the Maxwell equation (6.26) added to the solution (6.32). The homogeneous solution has been chosen such that the forward propagating waves and the backward propagating waves in (6.26) are only in, respectively, the upper half space (z larger than zero) and the lower half space (z smaller than zero). As it is required, the particular solution (6.32) is a linear combination of six independent functions. The coefficients A_1, A_2, B_1, B_2, C_1 and C_2 are found from the equations (6.30) and (6.31) and are the solution of the system of equations

$$\begin{pmatrix} i k_z s_y & -i k_z s_y & -i\mu k_0 s_x & i\mu k_0 s_x & k_y & 0 \\ -i k_z s_x & i k_z s_x & -i\mu k_0 s_y & i\mu k_0 s_y & -k_x & 0 \\ 0 & 0 & 0 & 0 & 0 & -\mu k_0 \\ -i\epsilon k_0 s_x & -i\epsilon k_0 s_x & -i k_z s_y & -i k_z s_y & 0 & -k_y \\ -i\epsilon k_0 s_y & -i\epsilon k_0 s_y & i k_z s_x & i k_z s_x & 0 & k_x \\ 0 & 0 & 0 & 0 & -\epsilon k_0 & 0 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \\ C_1 \\ C_2 \end{pmatrix} = k_0 \begin{pmatrix} M_x \\ M_y \\ M_z \\ P_x \\ P_y \\ P_z \end{pmatrix}. \tag{6.33}$$

The determinant of the matrix present in (6.33) is $(2\epsilon\mu k_z k_0^2)^2$, meaning that the cases $k_z = 0$, $\epsilon = 0$ and $\mu = 0$ does not have solution for most of the plane sources. After inverting the matrix in equation (6.33), the system of equations becomes

$$k_0 \begin{pmatrix} -\frac{is_y}{2k_z} & \frac{is_x}{2k_z} & 0 & \frac{is_x}{2\epsilon k_0} & \frac{is_y}{2\epsilon k_0} & -\frac{ik_\parallel}{2\epsilon k_z k_0} \\ \frac{is_y}{2k_z} & -\frac{is_x}{2k_z} & 0 & \frac{is_x}{2\epsilon k_0} & \frac{is_y}{2\epsilon k_0} & \frac{ik_\parallel}{2\epsilon k_z k_0} \\ \frac{is_x}{2\mu k_0} & \frac{is_y}{2\mu k_0} & -\frac{ik_\parallel}{2\mu k_z k_0} & \frac{is_y}{2k_z} & -\frac{is_x}{2k_z} & 0 \\ -\frac{is_x}{2\mu k_0} & -\frac{is_y}{2\mu k_0} & -\frac{ik_\parallel}{2\mu k_z k_0} & \frac{is_y}{2k_z} & -\frac{is_x}{2k_z} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{\epsilon k_0} \\ 0 & 0 & -\frac{1}{\mu k_0} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} M_x \\ M_y \\ M_z \\ P_x \\ P_y \\ P_z \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \\ C_1 \\ C_2 \end{pmatrix}, \tag{6.34}$$

which is equivalent to the solution (6.9).

6.7.2 Derivation of the adjoint method for periodic diffractive optical elements

In this section, the derivation of equations (6.17c), (6.19), and (6.20) given in section 6.4, which are required in the adjoint method for a periodic diffractive optical element, are provided. The adjoint method can be divided into five steps:

- Description of the merit function in terms of the fields $\vec{E}(\vec{x}')$ and $\vec{H}(\vec{x}')$ at the output plane.
- Computation of the variation of the merit function δF in terms of a variation of the fields $\delta\vec{E}(\vec{x}')$ and $\delta\vec{H}(\vec{x}')$.
- Expression of $\delta\vec{E}(\vec{x}')$ and $\delta\vec{H}(\vec{x}')$ using the Green tensor and a virtual source, which depends of the variation of the material parameters $\delta\epsilon(\vec{x})$ and $\delta\mu(\vec{x})$.
- Application of the properties of the Green tensor due to the reciprocity of the system such that the expression of the adjoint source appears.
- Transformation of the adjoint source into a set of plane waves.

The first step is to describe the transmission efficiency η_m of the order m in terms of the fields $\vec{E}(\vec{x}')$ and $\vec{H}(\vec{x}')$. In general, this field is composed of forward and backward-propagating plane waves with respectively weights t_m and b_m, meaning that

$$\psi = \sum_m t_m \psi_m + b_m \psi_m^-, \tag{6.35}$$

where ψ is a mode described by the fields $\vec{E}(\vec{x}')$ and $\vec{H}(\vec{x}')$, and ψ_m is the plane wave related to the order m.

The weights t_m can be found from the system of equations proposed in section 2.5. In order to have as many equations as unknowns, the system of equations is obtained by setting S_{Rm} and T_{Rm} expressed in (2.28) to zero, giving

$$\begin{aligned} [\psi_m|\psi+\psi^-] \quad &-\sum_v (t_v+b_v)[\psi_m|\psi_v+\psi_v^-] &&= 0\\ [\psi|\psi_m+\psi_m^-]^* &-\sum_v (t_v-b_v)[\psi_v|\psi_m+\psi_m^-]^* &&= 0, \end{aligned} \tag{6.36}$$

where the Poynting operation is defined as

$$[\psi_m|\psi_n] := \frac{1}{2|\Lambda|}\iint_\Lambda (\vec{E}_m(\vec{x}')\times\vec{H}_n^*(\vec{x}')+\vec{E}_n(\vec{x}')\times\vec{H}_m^*(\vec{x}'))\cdot\vec{n}\,d\vec{x}'. \tag{6.37}$$

Since the plane waves are either TE or TM-polarized, they are orthogonal with each other, meaning that $[\psi_v|\psi_m+\psi_m^-]$ is zero if m different than v. Therefore, the system of equations (6.36) reduces to

$$\begin{aligned} [\psi_m|\psi+\psi^-] \quad &-(t_m+b_m)[\psi_m|\psi_m] &&= 0\\ [\psi|\psi_m+\psi_m^-]^* &-(t_m-b_m)[\psi_m|\psi_m]^* &&= 0, \end{aligned} \tag{6.38}$$

and the weights t_m are given by

$$t_m = \frac{[\psi_m|\psi+\psi^-]+[\psi|\psi_m+\psi_m^-]^*}{2[\psi_m|\psi_m]} \tag{6.39}$$

if $[\psi_m|\psi_m]$ is purely real, meaning that the medium is lossless and ψ_m is propagating. Since η_m is given by

$$\eta_m = \frac{\Phi_m}{\Phi_{in}}|t_m|^2 \tag{6.40}$$

with

$$\Phi_m = [\psi_m|\psi_m] = p_m k_0 k_{z,m}, \tag{6.41}$$

η_m becomes

$$\begin{aligned}\eta_m &= \frac{\Phi_m}{\Phi_{in}} \left| \frac{1}{2p_m k_0 k_{z,m}|\Lambda|} \iint_\Lambda (\vec{E}(\vec{x}') \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \vec{H}(\vec{x}')) \cdot \vec{n}_z d\vec{x}' \right|^2 \\ &= \frac{1}{4p_m k_0 k_{z,m}|\Lambda|^2 \Phi_{in}} \left| \iint_\Lambda (\vec{E}(\vec{x}') \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \vec{H}(\vec{x}')) \cdot \vec{n}_z d\vec{x}' \right|^2,\end{aligned} \tag{6.42}$$

which is the same as equation (6.17c).

The second step is the computation of the variation of the merit function δF in terms of a variation of the fields $\delta\vec{E}$ and $\delta\vec{H}$. Since F is a function of the efficiencies η_m, δF is

$$\delta F = \sum_m \frac{\delta F}{\delta\eta_m} \delta\eta_m \tag{6.43}$$

and $\delta\eta_m$ is given by

$$\begin{aligned}\delta\eta_m =& \frac{1}{|\Lambda|\Phi_{in}} \mathrm{Re}\left\{ \iint_\Lambda (\delta\vec{E}(\vec{x}') \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \delta\vec{H}(\vec{x}')) \cdot \vec{n}_z d\vec{x}' \right. \\ & \left. \left(\frac{1}{2p_m k_0 k_{z,m}|\Lambda|} \iint_\Lambda (\vec{E}(\vec{x}') \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \vec{H}(\vec{x}')) \cdot \vec{n}_z d\vec{x}' \right)^* \right\}.\end{aligned} \tag{6.44}$$

The term in the parenthesis is the weight t_m given in (6.39), meaning that

$$\delta\eta_m = \frac{1}{|\Lambda|\Phi_{in}} \mathrm{Re}\left\{ \iint_\Lambda (\delta\vec{E}(\vec{x}') \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \delta\vec{H}(\vec{x}')) \cdot \vec{n}_z d\vec{x}' \; t_m^* \right\}. \tag{6.45}$$

The third step is the expression of $\delta\vec{E}(\vec{x}')$ and $\delta\vec{H}(\vec{x}')$ using the Green tensor. As shown in section 6.3, an infinitesimal variation of the material parameters $\delta\epsilon(\vec{x})$ and $\delta\mu(\vec{x})$ is equivalent to a source, and the effect of this source on the fields at the output plane is given by equation (6.15), where the source is described by $(\vec{E}(\vec{x})\delta\epsilon(\vec{x}), \vec{H}(\vec{x})\delta\mu(\vec{x}))^T$. After replacing $\delta\vec{E}(\vec{x}')$ and $\delta\vec{H}(\vec{x}')$ by the right-hand side of equation (6.15), and multiplying both sides of equation (6.45) by the derivative of the merit function F with respect to η_m, which is a real number, equation (6.45) becomes

$$\begin{aligned}\frac{\delta F}{\delta\eta_m}\delta\eta_m = \frac{1}{|\Lambda|\Phi_{in}} \mathrm{Re}\Bigg\{ \iint_\Lambda \iiint & \left(\hat{G}_{EP}(\vec{x}',\vec{x})\vec{E}(\vec{x}) \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \hat{G}_{HP}(\vec{x},\vec{x})\vec{E}(\vec{x}) \right) \cdot \vec{n}_z \delta\epsilon(\vec{x}) \\ & + \left(\hat{G}_{EM}(\vec{x}',\vec{x})\vec{H}(\vec{x}) \times \vec{H}_m^*(\vec{x}') + \vec{E}_m^*(\vec{x}') \times \hat{G}_{HM}(\vec{x}',\vec{x})\vec{H}(\vec{x}) \right) \cdot \vec{n}_z \delta\mu(\vec{x}) d\vec{x} d\vec{x}' \, \frac{\delta F}{\delta\eta_m} t_m^* \Bigg\}.\end{aligned} \tag{6.46}$$

The fourth step is the application of the properties of the Green tensor given in (6.16). In order to do that, the following properties of the triple product and scalar product are needed:

$$\begin{aligned}(\hat{A}\vec{a}\times\vec{b})\cdot\vec{c} &= (\vec{b}\times\vec{c})\cdot\hat{A}\vec{a} &&= \hat{A}^T(\vec{b}\times\vec{c})\cdot\vec{a}\\ (\vec{a}\times\hat{A}\vec{b})\cdot\vec{c} &= -(\vec{a}\times\vec{c})\cdot\hat{A}\vec{b} &&= -\hat{A}^T(\vec{a}\times\vec{c})\cdot\vec{b}.\end{aligned} \tag{6.47}$$

Applying the properties (6.47) to equation (6.46) gives

$$\begin{aligned}\frac{\delta F}{\delta\eta_m}\delta\eta_m = \frac{1}{|\Lambda|\Phi_{in}}\mathrm{Re}\Bigg\{&\iiint\iint_\Lambda\begin{pmatrix}\hat{G}^T_{EP}(\vec{x}',\vec{x})\\ -\hat{G}^T_{HP}(\vec{x}',\vec{x})\end{pmatrix}^T\frac{\delta F}{\delta\eta_m}t^*_m\begin{pmatrix}\vec{H}^*_m(\vec{x}')\times\vec{n}_z\\ \vec{E}^*_m(\vec{x}')\times\vec{n}_z\end{pmatrix}d\vec{x}'\cdot\vec{E}(\vec{x})\delta\epsilon(\vec{x})\\ -&\iint_\Lambda\begin{pmatrix}-\hat{G}^T_{EM}(\vec{x}',\vec{x})\\ \hat{G}^T_{HM}(\vec{x}',\vec{x})\end{pmatrix}^T\frac{\delta F}{\delta\eta_m}t^*_m\begin{pmatrix}\vec{H}^*_m(\vec{x}')\times\vec{n}_z\\ \vec{E}^*_m(\vec{x}')\times\vec{n}_z\end{pmatrix}d\vec{x}'\cdot\vec{H}(\vec{x})\delta\mu(\vec{x})d\vec{x}\Bigg\}.\end{aligned} \tag{6.48}$$

Then, the properties of the Green tensor (6.16) for a reciprocal system are applied to equation (6.48), leading to

$$\frac{\delta F}{\delta\eta_m}\delta\eta_m = -\frac{k_0}{|\Lambda|\Phi_{in}}\mathrm{Im}\left\{\iiint\vec{E}_{adj,m}(\vec{x})\cdot\vec{E}(\vec{x})\delta\epsilon(\vec{x}) - \vec{H}_{adj,m}(\vec{x})\cdot\vec{H}(\vec{x})\delta\mu(\vec{x})d\vec{x}\right\}, \tag{6.49}$$

where the adjoint fields $\vec{E}_{adj}(\vec{x})$ and $\vec{H}_{adj}(\vec{x})$ are given by

$$\begin{pmatrix}\vec{E}_{adj,m}(\vec{x})\\ \vec{H}_{adj,m}(\vec{x})\end{pmatrix} = \iint_\Lambda\begin{pmatrix}\hat{G}_{EP}(\vec{x},\vec{x}') & \hat{G}_{EM}(\vec{x},\vec{x}')\\ \hat{G}_{HP}(\vec{x},\vec{x}') & \hat{G}_{HM}(\vec{x},\vec{x}')\end{pmatrix}\frac{\delta F}{\delta\eta_m}\frac{t^*_m}{ik_0}\begin{pmatrix}\vec{H}^*_m(\vec{x}')\times\vec{n}_z\\ \vec{E}^*_m(\vec{x}')\times\vec{n}_z\end{pmatrix}d\vec{x}'. \tag{6.50}$$

.

Therefore, the adjoint source, which is a plane source, is

$$\vec{P}_m(\vec{x}') = \frac{\delta F}{\delta\eta_m}\frac{t^*_m}{ik_0}\vec{H}^*_m(\vec{x}')\times\vec{n}_z\delta(z') \qquad \vec{M}_m(\vec{x}') = \frac{\delta F}{\delta\eta_m}\frac{t^*_m}{ik_0}\vec{E}^*_m(\vec{x}')\times\vec{n}_z\delta(z'), \tag{6.51}$$

where $\delta(z)$ is the delta function and, for simplification, the coordinate system in which the vector $\vec{x}'$ is expressed, is chosen such that the output plane is at $z' = 0$. After combining equations (6.43) and (6.49), recognizing the functional derivative and assuming that the variation of the material parameters $\delta\epsilon(\vec{x})$ and $\delta\mu(\vec{x})$ is purely real, the functional derivative of the merit function F with respect to $\delta\epsilon(\vec{x})$ and $\delta\mu(\vec{x})$ is

$$\begin{aligned}\frac{\delta F}{\delta\epsilon(\vec{x})} &= -\frac{k_0}{|\Lambda|\Phi_{in}}\operatorname{Im}\left\{\vec{E}_{adj}(\vec{x})\cdot\vec{E}(\vec{x})\right\}\\ \frac{\delta F}{\delta\mu(\vec{x})} &= \frac{k_0}{|\Lambda|\Phi_{in}}\operatorname{Im}\left\{\vec{H}_{adj}(\vec{x})\cdot\vec{H}(\vec{x})\right\}\end{aligned} \tag{6.52}$$

where the adjoint fields $\vec{E}_{adj}(\vec{x})$ and $\vec{H}_{adj}(\vec{x})$ are generated by the adjoint source

$$\begin{aligned}\vec{P}(\vec{x}) &= \sum_m \vec{P}_m(\vec{x}) = \frac{1}{ik_0}\sum_m \frac{\delta F}{\delta\eta_m} t_m^* \vec{H}_m^*(\vec{x}) \times \vec{n}_z\delta(z)\\ \vec{M}(\vec{x}) &= \sum_m \vec{M}_m(\vec{x}) = \frac{1}{ik_0}\sum_m \frac{F}{\delta\eta_m} t_m^* \vec{E}_m^*(\vec{x}) \times \vec{n}_z\delta(z).\end{aligned} \tag{6.53}$$

Equation (6.52) is the same as equation (6.20).

The last step is the transformation of the adjoint source into a set of plane waves. With methods such as the Finite Difference Time Domain method (FDTD) and Finite Element Method (FEM), the polarization and magnetization densities of the adjoint source can be given as input. However, in the Fourier modal method, the only possible inputs are the weight of the incident plane waves. In order to find those weights, the results obtained in section 6.2 are applied. Since the plane source is assumed to be in a lossless medium defined by ϵ and μ, the plane sources $\vec{P}_m(\vec{x})$ and $\vec{M}_m(\vec{x})$ are given by

$$\begin{aligned}\vec{P}_{m,TM}(\vec{x}') &= \frac{\delta F}{\delta\eta_m}\frac{t_m^*}{ik_0}\begin{pmatrix}\epsilon k_0 s_{x,m}\\ \epsilon k_0 s_{y,m}\\ 0\end{pmatrix} e^{i(k_{x,p}x'+k_{y,p}y')}\delta(z)\\ \vec{M}_{m,TE}(\vec{x}') &= \frac{\delta F}{\delta\eta_m}\frac{t_m^*}{ik_0}\begin{pmatrix}k_z s_{y,m}\\ -k_z s_{x,m}\\ 0\end{pmatrix} e^{i(k_{x,p}x'+k_{y,p}y')}\delta(z)\end{aligned} \tag{6.54}$$

for TM-polarization and

$$\begin{aligned}\vec{P}_{m,TE}(\vec{x}') &= \frac{\delta F}{\delta\eta_m}\frac{t_m^*}{ik_0}\begin{pmatrix}k_z s_{y,m}\\ -k_z s_{x,m}\\ 0\end{pmatrix} e^{i(k_{x,p}x'+k_{y,p}y')}\delta(z)\\ \vec{M}_{m,TE}(\vec{x}') &= \frac{\delta F}{\delta\eta_m}\frac{t_m^*}{ik_0}\begin{pmatrix}-\mu k_0 s_{x,m}\\ -\mu k_0 s_{y,m}\\ 0\end{pmatrix} e^{i(k_{x,p}x'+k_{y,p}y')}\delta(z)\end{aligned} \tag{6.55}$$

for TE-polarization. Because the plane source given in (6.53) is defined from the complex conjugate of $\vec{E}_m$ and $\vec{H}_m$, the tangential components of the k-vector, $k_{x,p}$ and $k_{y,p}$, are given by $(k_{x,p}, k_{y,p}) = (-k_{x,m}, -k_{y,m})$.

From (6.9), the weights of the plane waves generated by the plane source are

$$\begin{aligned}
A_1 &= C_1 = C_2 = 0 \\
A_2 &= \frac{\delta F}{\delta \eta_m} t_m^* (s_{x,p} s_{x,m} + s_{y,p} s_{y,m}) \\
B_1 &= \frac{i}{2\mu k_z} \frac{\delta F}{\delta \eta_m} t_m^* (k_z^2 - \epsilon\mu k_0^2)(s_{x,p} s_{y,m} - s_{y,p} s_{x,m}) \\
B_2 &= -\frac{1}{2\mu k_z} \frac{\delta F}{\delta \eta_m} \frac{t_m^*}{k_0} (k_z^2 + \epsilon\mu k_0^2)(s_{x,p} s_{y,m} - s_{y,p} s_{x,m})
\end{aligned} \tag{6.56}$$

for TM-polarization and

$$\begin{aligned}
A_1 &= \frac{1}{2\epsilon k_z} \frac{\delta F}{\delta \eta_m} \frac{t_m^*}{k_0} (k_z^2 - \epsilon\mu k_0^2)(s_{x,p} s_{y,m} - s_{y,p} s_{x,m}) \\
A_2 &= \frac{1}{2\epsilon k_z} \frac{\delta F}{\delta \eta_m} \frac{t_m^*}{k_0} (k_z^2 + \epsilon\mu k_0^2)(s_{x,p} s_{y,m} - s_{y,p} s_{x,m}) \\
B_1 &= C_1 = C_2 = 0 \\
B_2 &= \frac{\delta F}{\delta \eta_m} t_m^* (s_{x,p} s_{x,m} + s_{y,p} s_{y,m})
\end{aligned} \tag{6.57}$$

for TE-polarization. Except for normal incidence ($k_{/\!/} = 0$), $s_{x,p}$ and $s_{y,p}$ are given by $(s_{x,p}, s_{y,p}) = -(s_{x,m}, s_{y,m})$. For normal incidence, $s_{x,p}$ and $s_{x,m}$ are equal to one, and $s_{y,p}$ and $s_{y,m}$ are equal to zero by convention. Hence, the term $s_{x,p} s_{y,m} - s_{y,p} s_{x,m}$ in (6.56) and (6.57) is zero and the adjoint source emits only in the direction of the diffractive optical element. The weight q_p of the emitted plane waves is given by

$$q_p = c_m \frac{\delta F}{\delta \eta_m} t_m^* \qquad c_m = \begin{cases} 1, \text{ for } k_{/\!/} = 0 \\ -1, \text{ otherwise} \end{cases} , \tag{6.58}$$

which is equation (6.19).

7 Conclusion

The principal contribution of this thesis is to provide a set of design techniques for binary dielectric metasurfaces that take advantage of internal parameters related to the eigen-modes propagating within the metasurface. Those design techniques are based on concepts, such as ideal metasurface and self-coupling modes, which also allow to get insight on the optical phenomena leading to the metasurface response.

We use the Fourier modal method for the simulation of binary dielectric metasurfaces because this method gives access to the internal parameters related to the eigen-modes, namely their propagation constant, their field profile and the coupling coefficients at the interfaces. In chapter 3, the Fourier modal method has been modified in order to facilitate the access to those internal parameters. In addition, by filtering the eigen-modes present in the metasurface, the response of metasurfaces with the same cross-section and materials, but different thicknesses can be computed in a few milliseconds from the simulation of a single metasurface. This feature greatly facilitates the design of anti-reflective metasurface (section 4.4), the design of metasurface-based half-wave plates (section 4.5), the search of resonances (section 5.3.4) and the exploration of the responses that can be obtained using metasurfaces.

Because of the difference between single-mode metasurfaces and multi-mode metasurfaces in terms of their responses, the design techniques proposed in this work are divided into two different chapters. In chapter 4, we propose two different approaches for the design of single-mode metasurface. The first one is related to the concept of ideal metasurface and it is used to get the main parameters of all possible single-mode metasurfaces from a given functionality. The same can be done based on trajectories on the Poincaré sphere as shown in section 8.11 of [65], but the equations that are given in this work can be applied in a more direct way. Equations are also given for two ideal metasurfaces, or two ideal waveplates, in series.

The second design techniques is based on the concept of Fabry-Pérot cavities and lead a design process for anti-reflective metasurface and metasurface-based half-wave plates. This design process is described step by step in order to greatly facilitate the design of metasurface with such functionality, independently of the materials and geometries involved.

The design techniques given in chapter 5, are based on the concept of self-coupling mode and greatly facilitates the design of resonant metasurfaces. Self-coupling modes are easily computed from the Fourier modal method, allow a systematic characterization of resonances, even if multiple resonances are overlapping, and reduce the number of simulations of resonant metasurfaces required in order to accurately interpolate the response in function of the wavelength, the angle of incidence or any other parameter. The concept of self-coupling mode is applied to four different metasurfaces, highlighting different aspects of its use. The same can be done with the concept of quasi-normal mode [147, 162], which is very similar to the self-coupling modes since a self-coupling mode with a s-value of one is a quasi-normal mode. However, when the Fourier modal method is used for the simulation of a metasurface, the self-coupling modes are easier to compute than the quasi-normal modes. Another way of dealing with resonances in metasurfaces composed of cylinders is to start with the resonances of a cylinder surrounded by a homogeneous medium. This approach is different than the approach based on the self-coupling mode and it would require further work in order to compare those two approaches in the design of resonant metasurfaces.

In most cases, the design techniques in chapters 4 and 5 give a solution only close to the fully optimized metasurface, mainly because the number of parameters that describe the geometry of the metasurface is limited to a small number. In order to optimize the metasurface further, some of the most efficient methods are gradient-based optimizations, which have the drawback that the gradient of the merit function needs to be computed. The two main methods to get the gradient are the algorithmic differentiation and the adjoint method. However, if the Fourier modal method is used for the simulation of the metasurface, the complexity of the Fourier modal method makes the adjoint method more suitable for the computation of the gradient. In chapter 6, we provide the expression of the functional derivative for any merit function based on the efficiency of the transmitted orders when the Fourier modal method is used. Moreover, we clearly write all the steps that need to be done in order to get the functional derivative, and we show that, despite that the adjoint method gives an estimation of the functional derivative, the gradient-based optimization converges rapidly to a solution. The advantage of using the Fourier modal method is that the response of the metasurface and the functional derivative of the merit function are obtained with a single simulation of the metasurface. In addition, the relationship between a plane source and the emitted plane waves are given. Such relationship is needed to get the expression of the functional derivative, but it can also be used to find the surface susceptibilities for a given functionality.

The Poynting operation, introduced in chapter 2, is an operation defined by its properties and it has a strong relationship with the power flow. It can be used for the reformulation of the boundary condition and for modes orthonormalization. One of its advantage are that the reflection and transmission coefficients can be expressed without computing the fields related to the modes, leading to equations that can be applied for a wide range of systems. In this work, the Poynting operation is used to express the efficiency of the transmitted orders based on the fields, to analyze the contribution of each eigen-mode to the power flow and to orthonormalize the eigen-modes.

This work not only provides design techniques, but also a better idea on the functionalities single-mode and multi-mode metasurfaces can have along with trade-offs. Due the diversity in their response, multi-mode metasurfaces are particularly interesting and, with the help of the Fourier modal method implemented in this work and the concept of self-coupling mode, it is now easier to design performant multi-mode metasurfaces for a wide range of functionalities.

Bibliography

[1] Max Herzberger. Optics from euclid to huygens. *Applied Optics*, 5(9):1383, sep 1966.

[2] Abdelghani Tbakhi and Samir S. Amr. Ibn al-haytham: Father of modern optics. *Annals of Saudi Medicine*, 27(6):464–467, nov 2007.

[3] Tomoyuki Matsuyama, Toshiro Ishiyama, and Yasuhiro Omura. Nikon projection lens update. may 2004.

[4] Nan Fu, , Yanxiang Liu, Xiaolong Ma, and Zhanfeng Chen. EUV lithography: State-of-the-art review. *Journal of Microelectronic Manufacturing*, 2(2):1–6, 2019.

[5] J. C. Maxwell. VIII. a dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London*, 155:459–512, dec 1865.

[6] F. Hopkinson and David Rittenhouse. An optical problem, proposed by mr. hopkinson, and solved by mr. rittenhouse. *Transactions of the American Philosophical Society*, 2:201, 1786.

[7] Nicolas Bonod and Jérôme Neauport. Diffraction gratings: from principles to applications in high-intensity lasers. *Advances in Optics and Photonics*, 8(1):156, mar 2016.

[8] D. Gabor. Microscopy by reconstructed wave-fronts. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 197(1051):454–487, jul 1949.

[9] B. R. Brown and A. W. Lohmann. Complex spatial filtering with binary masks. *Applied Optics*, 5(6):967, jun 1966.

[10] Sam Van der Jeught and Joris J.J. Dirckx. Real-time structured light profilometry: a review. *Optics and Lasers in Engineering*, 87:18–31, dec 2016.

[11] Florian Rößler, Tim Kunze, and Andrés Fabián Lasagni. Fabrication of diffraction based security elements using direct laser interference patterning. *Optics Express*, 25(19):22959, sep 2017.

[12] Daniel Werdehausen, Isabelle Staude, Sven Burger, Jörg Petschulat, Toralf Scharf, Thomas Pertsch, and Manuel Decker. Design rules for customizable optical materials based on nanocomposites. *Optical Materials Express*, 8(11):3456, oct 2018.

[13] Winston E. Kock. Metallic delay lenses. *Bell System Technical Journal*, 27(1):58–82, jan 1948.

[14] J. B. Pendry, A. J. Holden, W. J. Stewart, and I. Youngs. Extremely low frequency plasmons in metallic mesostructures. *Physical Review Letters*, 76(25):4773–4776, jun 1996.

[15] D. R. Smith, Willie J. Padilla, D. C. Vier, S. C. Nemat-Nasser, and S. Schultz. Composite medium with simultaneously negative permeability and permittivity. *Physical Review Letters*, 84(18):4184–4187, may 2000.

[16] J. B. Pendry. Negative refraction makes a perfect lens. *Physical Review Letters*, 85(18):3966–3969, oct 2000.

[17] F. Falcone, T. Lopetegi, M. A. G. Laso, J. D. Baena, J. Bonache, M. Beruete, R. Marqués, F. Martín, and M. Sorolla. Babinet principle applied to the design of metasurfaces and metamaterials. *Physical Review Letters*, 93(19), nov 2004.

[18] Shuang Zhang, Wenjun Fan, N. C. Panoiu, K. J. Malloy, R. M. Osgood, and S. R. J. Brueck. Experimental demonstration of near-infrared negative-index metamaterials. *Physical Review Letters*, 95(13), sep 2005.

[19] Fu Min Huang, Tsung Sheng Kao, Vassili A. Fedotov, Yifang Chen, and Nikolay I. Zheludev. Nanohole array as a lens. *Nano Letters*, 8(8):2469–2472, aug 2008.

[20] Xingjie Ni, Satoshi Ishii, Alexander V Kildishev, and Vladimir M Shalaev. Ultra-thin, planar, babinet-inverted plasmonic metalenses. *Light: Science & Applications*, 2(4):e72–e72, apr 2013.

[21] N. Yu, P. Genevet, M. A. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, and Z. Gaburro. Light propagation with phase discontinuities: Generalized laws of reflection and refraction. *Science*, 334(6054):333–337, sep 2011.

[22] Xiaobin Hu and Xin Wei. Metallic metasurface for high efficiency optical phase control in transmission mode. *Optics Express*, 25(13):15208, jun 2017.

[23] W. Stork, N. Streibl, H. Haidner, and P. Kipfer. Artificial distributed-index media fabricated by zero-order gratings. *Optics Letters*, 16(24):1921, dec 1991.

[24] Philippe Lalanne, Simion Astilean, Pierre Chavel, Edmond Cambril, and Huguette Launois. Design and fabrication of blazed binary diffractive elements with sampling periods smaller than the structural cutoff. *Journal of the Optical Society of America A*, 16(5):1143, may 1999.

[25] Ze'ev Bomzon, Gabriel Biener, Vladimir Kleiner, and Erez Hasman. Space-variant pancharatnam–berry phase optical elements with computer-generated subwavelength gratings. *Optics Letters*, 27(13):1141, jul 2002.

[26] Seyedeh Mahsa Kamali, Amir Arbabi, Ehsan Arbabi, Yu Horie, and Andrei Faraon. Decoupling optical function and geometrical form using conformal flexible dielectric metasurfaces. *Nature Communications*, 7(1), may 2016.

[27] M. Khorasaninejad, A. Y. Zhu, C. Roques-Carmes, W. T. Chen, J. Oh, I. Mishra, R. C. Devlin, and F. Capasso. Polarization-insensitive metalenses at visible wavelengths. *Nano Letters*, 16(11):7229–7234, oct 2016.

[28] Zhenpeng Zhou, Juntao Li, Rongbin Su, Beimeng Yao, Hanlin Fang, Kezheng Li, Lidan Zhou, Jin Liu, Daan Stellinga, Christopher P. Reardon, Thomas F. Krauss, and Xuehua Wang. Efficient silicon metasurfaces for visible light. *ACS Photonics*, 4(3):544–551, feb 2017.

[29] You Zhou, Ivan I. Kravchenko, Hao Wang, J. Ryan Nolen, Gong Gu, and Jason Valentine. Multilayer noninteracting dielectric metasurfaces for multiwavelength metaoptics. *Nano Letters*, 18(12):7529–7537, nov 2018.

[30] Nasir Mahmood, Inki Kim, Muhammad Qasim Mehmood, Heonyeong Jeong, Ali Akbar, Dasol Lee, Murtaza Saleem, Muhammad Zubair, Muhammad Sabieh Anwar, Farooq Ahmad Tahir, and Junsuk Rho. Polarisation insensitive multifunctional metasurfaces based on all-dielectric nanowaveguides. *Nanoscale*, 10(38):18323–18330, 2018.

[31] D. Lin, P. Fan, E. Hasman, and M. L. Brongersma. Dielectric gradient metasurface optical elements. *Science*, 345(6194):298–302, jul 2014.

[32] Kun Huang, Zhaogang Dong, Shengtao Mei, Lei Zhang, Yanjun Liu, Hong Liu, Haibin Zhu, Jinghua Teng, Boris Luk'yanchuk, Joel K.W. Yang, and Cheng-Wei Qiu. Silicon multi-meta-holograms for the broadband visible light. *Laser & Photonics Reviews*, 10(3):500–509, apr 2016.

[33] Mohammadreza Khorasaninejad, Wei Ting Chen, Robert C. Devlin, Jaewon Oh, Alexander Y. Zhu, and Federico Capasso. Metalenses at visible wavelengths: Diffraction-limited focusing and subwavelength resolution imaging. *Science*, 352(6290):1190–1194, jun 2016.

[34] M. Khorasaninejad, W. T. Chen, A. Y. Zhu, J. Oh, R. C. Devlin, D. Rousso, and F. Capasso. Multispectral chiral imaging with a metalens. *Nano Letters*, 16(7):4595–4600, jun 2016.

[35] Yuanyuan Liu, Huiying Zhou, and Jin Zhang. Modulation of spin-dependent diffraction based on dielectric metasurfaces. *Scientific Reports*, 10(1), may 2020.

[36] J. P. Balthasar Mueller, Noah A. Rubin, Robert C. Devlin, Benedikt Groever, and Federico Capasso. Metasurface polarization optics: Independent phase control of arbitrary orthogonal states of polarization. *Physical Review Letters*, 118(11), mar 2017.

[37] Katie E. Chong, Lei Wang, Isabelle Staude, Anthony R. James, Jason Dominguez, Sheng Liu, Ganapathi S. Subramania, Manuel Decker, Dragomir N. Neshev, Igal Brener, and

Yuri S. Kivshar. Efficient polarization-insensitive complex wavefront control using huygens' metasurfaces based on dielectric resonant meta-atoms. *ACS Photonics*, 3(4):514–519, mar 2016.

[38] Yanliang He, Ying Li, Junmin Liu, Xiaoke Zhang, Yao Cai, Yu Chen, Shuqing Chen, and Dianyuan Fan. Switchable phase and polarization singular beams generation using dielectric metasurfaces. *Scientific Reports*, 7(1), jul 2017.

[39] Hongqiang Zhou, Basudeb Sain, Yongtian Wang, Christian Schlickriede, Ruizhe Zhao, Xue Zhang, Qunshuo Wei, Xiaowei Li, Lingling Huang, and Thomas Zentgraf. Polarization-encrypted orbital angular momentum multiplexed metasurface holography. *ACS Nano*, 14(5):5553–5559, apr 2020.

[40] Amir Arbabi, Yu Horie, Mahmood Bagheri, and Andrei Faraon. Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission. *Nature Nanotechnology*, 10(11):937–943, aug 2015.

[41] Zhongyi Guo, Lie Zhu, Kai Guo, Fei Shen, and Zhiping Yin. High-order dielectric metasurfaces for high-efficiency polarization beam splitters and optical vortex generators. *Nanoscale Research Letters*, 12(1), aug 2017.

[42] Zhongyi Guo, Haisheng Xu, Kai Guo, Fei Shen, Hongping Zhou, Qingfeng Zhou, Jun Gao, and Zhiping Yin. High-efficiency visible transmitting polarizations devices based on the GaN metasurface. *Nanomaterials*, 8(5):333, may 2018.

[43] Manuel Decker, Isabelle Staude, Matthias Falkner, Jason Dominguez, Dragomir N. Neshev, Igal Brener, Thomas Pertsch, and Yuri S. Kivshar. High-efficiency dielectric huygens' surfaces. *Advanced Optical Materials*, 3(6):813–820, feb 2015.

[44] Katie E. Chong, Isabelle Staude, Anthony James, Jason Dominguez, Sheng Liu, Salvatore Campione, Ganapathi S. Subramania, Ting S. Luk, Manuel Decker, Dragomir N. Neshev, Igal Brener, and Yuri S. Kivshar. Polarization-independent silicon metadevices for efficient optical wavefront control. *Nano Letters*, 15(8):5369–5374, jul 2015.

[45] Wenyu Zhao, Huan Jiang, Bingyi Liu, Jie Song, Yongyuan Jiang, Chengchun Tang, and Junjie Li. Dielectric huygens' metasurface for high-efficiency hologram operating in transmission mode. *Scientific Reports*, 6(1), jul 2016.

[46] Bo Wang, Fengliang Dong, Qi-Tong Li, Dong Yang, Chengwei Sun, Jianjun Chen, Zhiwei Song, Lihua Xu, Weiguo Chu, Yun-Feng Xiao, Qihuang Gong, and Yan Li. Visible-frequency dielectric metasurfaces for multiwavelength achromatic and highly dispersive holograms. *Nano Letters*, 16(8):5235–5240, jul 2016.

[47] Chul-Soon Park, Vivek Raj Shrestha, Wenjing Yue, Song Gao, Sang-Shin Lee, Eun-Soo Kim, and Duk-Yong Choi. Structural color filters enabled by a dielectric metasurface incorporating hydrogenated amorphous silicon nanodisks. *Scientific Reports*, 7(1), may 2017.

[48] Ishwor Koirala, Sang-Shin Lee, and Duk-Yong Choi. Highly transmissive subtractive color filters based on an all-dielectric metasurface incorporating TiO2 nanopillars. *Optics Express*, 26(14):18320, jul 2018.

[49] Xiaofei Zang, Fengliang Dong, Fuyong Yue, Chunmei Zhang, Lihua Xu, Zhiwei Song, Ming Chen, Pai-Yen Chen, Gerald S. Buller, Yiming Zhu, Songlin Zhuang, Weiguo Chu, Shuang Zhang, and Xianzhong Chen. Polarization encoded color image embedded in a dielectric metasurface. *Advanced Materials*, 30(21):1707499, mar 2018.

[50] Jonas Berzinš, Stefan Fasold, Thomas Pertsch, Stefan M. B. Bäumer, and Frank Setzpfandt. Submicrometer nanostructure-based RGB filters for CMOS image sensors. *ACS Photonics*, 6(4):1018–1025, mar 2019.

[51] Zhijie Ma, Yi Li, Yang Li, Yandong Gong, Stefan A. Maier, and Minghui Hong. All-dielectric planar chiral metasurface with gradient geometric phase. *Optics Express*, 26(5):6067, feb 2018.

[52] Zhenyu Yang, Zhaokun Wang, Yuxi Wang, Xing Feng, Ming Zhao, Zhujun Wan, Liangqiu Zhu, Jun Liu, Yi Huang, Jinsong Xia, and Martin Wegener. Generalized hartmann-shack array of dielectric metalens sub-arrays for polarimetric beam profiling. *Nature Communications*, 9(1), nov 2018.

[53] Andreas Tittl, Aleksandrs Leitis, Mingkai Liu, Filiz Yesilkoy, Duk-Yong Choi, Dragomir N. Neshev, Yuri S. Kivshar, and Hatice Altug. Imaging-based molecular barcoding with pixelated dielectric metasurfaces. *Science*, 360(6393):1105–1109, jun 2018.

[54] Sheng Liu, Aleksandr Vaskin, Sadhvikas Addamane, Benjamin Leung, Miao-Chan Tsai, Yuanmu Yang, Polina P. Vabishchevich, Gordon A. Keeler, George Wang, Xiaowei He, Younghee Kim, Nicolai F. Hartmann, Han Htoon, Stephen K. Doorn, Matthias Zilk, Thomas Pertsch, Ganesh Balakrishnan, Michael B. Sinclair, Isabelle Staude, and Igal Brener. Light-emitting metasurfaces: Simultaneous control of spontaneous emission and far-field radiation. *Nano Letters*, 18(11):6906–6914, oct 2018.

[55] Jongwon Lee, Nishant Nookala, J. Sebastian Gomez-Diaz, Mykhailo Tymchenko, Frederic Demmerle, Gerhard Boehm, Markus-Christian Amann, Andrea Alù, and Mikhail A. Belkin. Ultrathin second-harmonic metasurfaces with record-high nonlinear optical response. *Advanced Optical Materials*, 4(5):664–670, feb 2016.

[56] Sheng Liu, Michael B. Sinclair, Sina Saravi, Gordon A. Keeler, Yuanmu Yang, John Reno, Gregory M. Peake, Frank Setzpfandt, Isabelle Staude, Thomas Pertsch, and Igal Brener. Resonantly enhanced second-harmonic generation using III–v semiconductor all-dielectric metasurfaces. *Nano Letters*, 16(9):5426–5432, aug 2016.

[57] Polina P. Vabishchevich, Sheng Liu, Michael B. Sinclair, Gordon A. Keeler, Gregory M. Peake, and Igal Brener. Enhanced second-harmonic generation using broken symmetry III–v semiconductor fano metasurfaces. *ACS Photonics*, 5(5):1685–1690, jan 2018.

[58] Mark Lawrence, David R. Barton, and Jennifer A. Dionne. Nonreciprocal flat optics with silicon metasurfaces. *Nano Letters*, 18(2):1104–1109, jan 2018.

[59] Aristi Christofi, Yuma Kawaguchi, Andrea Alù, and Alexander B. Khanikaev. Giant enhancement of faraday rotation due to electromagnetically induced transparency in all-dielectric magneto-optical metasurfaces. *Optics Letters*, 43(8):1838, apr 2018.

[60] Edgar O. Owiti, Hanning Yang, Peng Liu, Calvine F. Ominde, and Xiudong Sun. Polarization converter with controllable birefringence based on hybrid all-dielectric-graphene metasurface. *Nanoscale Research Letters*, 13(1), feb 2018.

[61] Lifeng Li. New formulation of the fourier modal method for crossed surface-relief gratings. *Journal of the Optical Society of America A*, 14(10):2758–2767, oct 1997.

[62] P. Lalanne, J.P. Hugonin, and P. Chavel. Optical properties of deep lamellar gratings: A coupled bloch-mode insight. *Journal of Lightwave Technology*, 24(6):2442–2449, jun 2006.

[63] O. D. Miller. *Photonics design: from fundamental solar cell physics to computational inverse design*. PhD thesis, University of California, 2012.

[64] Francois Quentel, Jim Fieret, Andrew S. Holmes, and Sylvain Paineau. Multilevel diffractive optical element manufacture by excimer laser ablation and halftone masks. In Malcolm C. Gower, Henry Helvajian, Koji Sugioka, and Jan J. Dubowski, editors, *Laser Applications in Microelectronic and Optoelectronic Manufacturing VI*. SPIE, jun 2001.

[65] Jan Balthasar Mueller. *Polarization in Nanophotonics*. PhD thesis, Harvard University, 2016.

[66] Raphaël Pestourie, Carlos Pérez-Arancibia, Zin Lin, Wonseok Shin, Federico Capasso, and Steven G. Johnson. Inverse design of large-area metasurfaces. *Optics Express*, 26(26):33732, dec 2018.

[67] Edwin K. P. Chong and Stanislaw H. Żak. An introduction to optimization. feb 2008.

[68] Ehsan Arbabi, Amir Arbabi, Seyedeh Mahsa Kamali, Yu Horie, and Andrei Faraon. Controlling the sign of chromatic dispersion in diffractive optics with dielectric metasurfaces. *Optica*, 4(6):625, jun 2017.

[69] M. Khorasaninejad, Z. Shi, A. Y. Zhu, W. T. Chen, V. Sanjeev, A. Zaidi, and F. Capasso. Achromatic metalens over 60 nm bandwidth in the visible and metalens with reverse chromatic dispersion. *Nano Letters*, 17(3):1819–1824, feb 2017.

[70] C.F.R. Mateus, M.C.Y. Huang, L. Chen, C.J. Chang-Hasnain, and Y. Suzuki. Broad-band mirror using a subwavelength grating. *IEEE Photonics Technology Letters*, 16(7):1676–1678, jul 2004.

[71] Uriel Levy, Emanuel Marom, and David Mendlovic. Thin element approximation for the analysis of blazed gratings: simplified model and validity limits. *Optics Communications*, 229(1-6):11–21, jan 2004.

[72] Philippe Lalanne and Pierre Chavel. Metalenses at visible wavelengths: past, present, perspectives. *Laser & Photonics Reviews*, 11(3):1600295, may 2017.

[73] Donald C. O'Shea, Thomas J. Suleski, Alan D. Kathman, and Dennis W. Prather. *Diffractive Optics: Design, Fabrication, and Test.* SPIE, dec 2003.

[74] Joseph W. Goodman. *Introduction to Fourier optics.* Roberts and Company Publishers, 2nd edition, 1996.

[75] A. Taflove. *Computational Electrodynamics: The Finite-Difference Time-Domain Method.* Artech House, 2005.

[76] S. Burger, J. Pomplun, and F. Schmidt. Finite element methods for computational nano-optics. In *Encyclopedia of Nanotechnology*, pages 1191–1197. 2016.

[77] A. C. Polycarpou. *Introduction to the Finite Element Method in Electromagnetics.* Morgan & Claypool, 2006.

[78] J. Jin. *The Finite Element Method in Electromagnetics.* John Wiley & Sons, 2nd edition, 2002.

[79] E. Noponen and J. Turunen. Eigenmode method for electromagnetic synthesis of diffractive elements with three-dimensional profiles. *Journal of the Optical Society of America A*, 11(9):2494–2502, 1994.

[80] W. Iff. *Rigorous Fourier Methods Based on Numerical Integration for the Calculation of Diffractive Optical Systems.* PhD thesis, Friedrich-Alexander-Universität, 2016.

[81] H. Kim, J. Park, and B. Lee. *Fourier Modal Method and Its Applications in Computational Nanophotonics.* CRC Press, 2017.

[82] P. Lalanne. Improved formulation of the coupled-wave method for two-dimensional gratings. *Journal of the Optical Society of America A*, 14(7):1592–1598, 1997.

[83] Thaibao Phan, David Sell, Evan W. Wang, Sage Doshay, Kofi Edee, Jianji Yang, and Jonathan A. Fan. High-efficiency, large-area, topology-optimized metasurfaces. *Light: Science & Applications*, 8(1), may 2019.

[84] Steven J. Byrnes, Alan Lenef, Francesco Aieta, and Federico Capasso. Designing large, high-efficiency, high-numerical-aperture, transmissive meta-lenses for visible light. *Optics Express*, 24(5):5110, mar 2016.

[85] Tianji Liu, Rongyang Xu, Peng Yu, Zhiming Wang, and Junichi Takahara. Multipole and multimode engineering in mie resonance-based metastructures. *Nanophotonics*, 9(5):1115–1137, mar 2020.

[86] Assefaw H. Gebremedhin and Andrea Walther. An introduction to algorithmic differentiation. *WIREs Data Mining and Knowledge Discovery*, 10(1), oct 2019.

[87] Filip Šrajer, Zuzana Kukelova, and Andrew Fitzgibbon. A benchmark of selected algorithmic differentiation tools on some problems in computer vision and machine learning. 2018. ArXiv:1807.10129v1.

[88] Christopher M. Lalau-Keraly, Samarth Bhargava, Owen D. Miller, and Eli Yablonovitch. Adjoint shape optimization applied to electromagnetic design. *Optics Express*, 21(18):21693, sep 2013.

[89] David Sell, Jianji Yang, Sage Doshay, Rui Yang, and Jonathan A. Fan. Large-angle, multifunctional metagratings based on freeform multimode geometries. *Nano Letters*, 17(6):3752–3757, may 2017.

[90] Sean Molesky, Zin Lin, Alexander Y. Piggott, Weiliang Jin, Jelena Vucković, and Alejandro W. Rodriguez. Inverse design in nanophotonics. *Nature Photonics*, 12(11):659–670, oct 2018.

[91] Olivier Ripoll, Ville Kettunen, and Hans Peter Herzig. Review of iterative fourier-transform algorithms for beam shaping applications. *Optical Engineering*, 43(11):2549, nov 2004.

[92] B. Saleh and M. Teich. *Fundamentals of Photonics*. Canada Wiley Interscience, 2nd edition, June 2007.

[93] Elyas Bayati, Raphaël Pestourie, Shane Colburn, Zin Lin, Steven G. Johnson, and Arka Majumdar. Inverse designed metalenses with extended depth of focus. *ACS Photonics*, 7(4):873–878, mar 2020.

[94] L. Li. Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings. *J. Opt. Soc. Am. A.*, 1996.

[95] T. Antonakakis, F. I. Baida, A. Belkhir, K. Cherednichenko, S. Cooper, et al. *Gratings: Theory and Numeric Applications*. AMU,(PUP), CNRS, ECM, 2nd edition, 2014.

[96] F. Zolla, G. Renversez, A. Nicolet, B. Kuhlmey, S. Guenneau, and D. Felbacq. *Foundations of Photonic Crystal Fibres*. Imperial College Press, 2005.

[97] P. Genevet, F. Capasso, F. Aieta, M. Khorasaninejad, and R. Devlin. Recent advances in planar optics: from plasmonic to dielectric metasurfaces. *Optica*, 4(1):139–152, 2017.

[98] G. Sztefka and H. P. Nolting. Bidirectional eigenmode propagation for large refractive index steps. *IEEE Photonics Technology Letters*, 5(5):554–557, 1993.

[99] M. Mrozowski. *Guided electromagnetic waves: properties and analysis*. Research Studies Press, 1997.

[100] A. D. Yaghjian. Bidirectionality of reciprocal, lossy or lossless, uniform or periodic waveguides. *IEEE Microwave and Wireless Components Letters*, 17(7):480–482, 2007.

[101] P. R. McIsaac. Bidirectionality in gyrotropic waveguide. *IEEE Transactions on Microwave Theory and Techniques*, 24(4):223–226, 1976.

[102] D. Tihon, C. Craeye, C. Guclu, F. Capolino, and S. Withington. Orthogonality properties of eigenmodes inside z-invariant periodic structures. *9th International Congress on Advanced Electromagnetic Materials in Microwaves and Optics*, pages 505–507, 2015.

[103] G. Lecamp, J. P. Hugonin, and P. Lalanne. Theoretical and computational concepts for periodic optical waveguides. *Opt. Express*, 15(18):11042–11060, 2007.

[104] A. W. Snyder and J. D. Love. *Optical waveguide theory*. Springer, 1983.

[105] W. Smigaj, P. Lalanne, J. Yang, T. Paul, C. Rockstuhl, and F. Lederer. Closed-form expression for the scattering coefficients at an interface between two periodic media. *Appl. Phys. Lett.*, 98(11), 2011.

[106] T. Paul, C. Menzel, W. Smigaj, C. Rockstuhl, P. Lalanne, and F. Lederer. Reflection and transmission of light at periodic layered metamaterial films. *Phys. Rev. B*, 84(11), 2011.

[107] T. Kaiser, S. Bin Hasan, T. Paul, T. Pertsch, and C. Rockstuhl. Impedance generalization for plasmonic waveguides beyond the lumped circuit model. *Phys. Rev. B*, 88(3), 2013.

[108] C. Wang. Electromagnetic power flow, fermat's principle, and special theory of relativity. *Optik*, 126(20):2703–2705, 2015.

[109] D. A. Jacobs, A. E. Miroshnichenko, Y. S. Kivshar, and A. B. Khanikaev. Photonic topological chern insulators based on tellegen metacrystals. *New J. Phys.*, 17(12), 2015.

[110] K. J. Garcia. Calculating component coupling coefficients. *Laser focus world*, 36(8):51–56, 2000.

[111] C.-L. Chen. *Foundations for guided-wave optics*. Wiley-Interscience, 2007.

[112] K. Okamoto. *Fundamentals of optical waveguides*. Elsevier, 2nd edition, 2006.

[113] R. E. Collin. *Field theory of guided waves*. Wiley-Interscience, 2nd edition, 1991.

[114] J. A. Kong. Theorems of bianisotropic media. *Proceedings of the IEEE*, 60(9):1036–1046, 1972.

[115] J. A. Kong. Image theory for bianisotropic media. *IEEE Transactions on Antennas and Propagation*, 19(3):451–452, 1971.

[116] T. F. Jablonski. Complex modes in open lossless dielectric waveguides. *J. Opt. Soc. Am. A*, 11(4):1272–1282, 1994.

[117] R. E. Wagner and W. J. Tomlinson. Coupling efficiency of optics in single-mode fiber components. *Applied Optics*, 21(15):2671–2688, 1982.

[118] W. Shin. *3D Finite-Difference Frequency-Domain Method for Plasmonics and Nanophotonics.* PhD thesis, Stanford University, 2013.

[119] W. C. Gibson. *The Method of Moments in Electromagnetics.* Chapman & Hall/CRC, 2007.

[120] M. G. Moharam and T. K. Gaylord. Three-dimensional vector coupled-wave analysis of planar-grating diffraction. *J. Opt. Soc. Am.*, 1983.

[121] L. Li. Use of fourier series in the analysis of discontinuous periodic structures. *J. Opt. Soc. Am.*, 13(9):1870–1876, 1996.

[122] Bertil Nistad and Johannes Skaar. Causality and electromagnetic properties of active media. *Physical Review E*, 78(3), sep 2008.

[123] Niels Gregersen, Stephan Reitzenstein, Caroline Kistner, Micha Strauss, Christian Schneider, Sven Hofling, Lukas Worschech, Alfred Forchel, Torben Roland Nielsen, Jesper Mork, and Jean-Michel Gerard. Numerical and experimental study of the q factor of high-q micropillar cavities. *IEEE Journal of Quantum Electronics*, 46(10):1470–1483, oct 2010.

[124] Jakob Rosenkrantz de Lasson, Philip Trøst Kristensen, Jesper Mørk, and Niels Gregersen. Roundtrip matrix method for calculating the leaky resonant modes of open nanophotonic structures. *Journal of the Optical Society of America A*, 31(10):2142, sep 2014.

[125] N. S. Nye, A. Swisher, C. Bungay, S. Tuenge, T. Mayer, D. N. Christodoulides, and C. Rivero-Baleine. Design of broadband anti-reflective metasurfaces based on an effective medium approach. In Jay N. Vizgaitis, Bjørn F. Andresen, Peter L. Marasco, Jasbinder S. Sanghera, and Miguel P. Snyder, editors, *Advanced Optics for Defense Applications: UV through LWIR II.* SPIE, may 2017.

[126] Viktoriia E. Babicheva, Mihail I. Petrov, Kseniia V. Baryshnikova, and Pavel A. Belov. Reflection compensation mediated by electric and magnetic resonances of all-dielectric metasurfaces [invited]. *Journal of the Optical Society of America B*, 34(7):D18, apr 2017.

[127] E. Slivina, A. Abass, D. Bätzner, B. Strahm, C. Rockstuhl, and I. Fernandez-Corbaton. Insights into backscattering suppression in solar cells from the helicity-preservation point of view. *Physical Review Applied*, 12(5), nov 2019.

[128] Yuan Dong, Zhengji Xu, Nanxi Li, Jinchao Tong, Yuan Hsing Fu, Yanyan Zhou, Ting Hu, Qize Zhong, Vladimir Bliznetsov, Shiyang Zhu, Qunying Lin, Dao Hua Zhang, Yuandong Gu, and Navab Singh. Si metasurface half-wave plates demonstrated on a 12-inch CMOS platform. *Nanophotonics*, 9(1):149–157, nov 2019.

[129] Dennis H. Goldstein. *Polarized Light.* CRC Press, dec 2017.

[130] F. Ratajczyk, W.A. Woźniak, and P. Kurzynowski. Transformation of polarization state of the light using wave plates with arbitrary phase difference. *Optics Communications*, 183(1-4):1–5, sep 2000.

[131] Karol Salazar-Ariza and Rafael Torres. Trajectories on the poincaré sphere of polarization states of a beam passing through a rotating linear retarder. *Journal of the Optical Society of America A*, 35(1):65, dec 2017.

[132] Henry Hurwitz and R. Clark Jones. A new calculus for the treatment of optical SystemsII proof of three general equivalence theorems. *Journal of the Optical Society of America*, 31(7):493, jul 1941.

[133] S. N. Savenkov. Jones and mueller matrices: structure, symmetry relations and information content. In A. A. Kokhanovsky, editor, *Light Scattering Reviews 4*. Springer, Berlin, Heidelberg, 2009.

[134] S. Pancharatnam. Generalized theory of interference, and its applications. *Proceedings of the Indian Academy of Sciences - Section A*, 44(5):247–262, nov 1956.

[135] M. V. Berry. Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 392(1802):45–57, mar 1984.

[136] A. Perot and C. Fabry. On the application of interference phenomena to the solution of various problems of spectroscopy and metrology. *The Astrophysical Journal*, 9:87–115, 1899.

[137] S. G. Lipson, H. Lispon, and D. S. Tannhauser. *Optical physics*. Cambridge University Press, 3rd edition, 1995.

[138] Xufeng Jing, Chengfei Chu, Chenxia Li, Haiyong Gan, Yingwei He, Xincui Gui, and Zhi Hong. Enhancement of bandwidth and angle response of metasurface cloaking through adding antireflective moth-eye-like microstructure. *Optics Express*, 27(15):21766, jul 2019.

[139] E.F. Kuester and C.L. Holloway. Comparison of approximations for effective parameters of artificial dielectrics. *IEEE Transactions on Microwave Theory and Techniques*, 38(11):1752–1755, 1990.

[140] M. G. Moharam. Coupled-wave analysis of two-dimensional dielectric gratings. In Ivan Cindrich, editor, *Holographic Optics: Design and Applications*. SPIE, apr 1988.

[141] R. Stoian. *Investigations of the dynamics of material removal in ultrashort pulsed laser ablation of dielectrics*. PhD thesis, Freie Universität Berlin, 2001.

[142] Salvatore Campione, Sheng Liu, Lorena I. Basilio, Larry K. Warne, William L. Langston, Ting S. Luk, Joel R. Wendt, John L. Reno, Gordon A. Keeler, Igal Brener, and Michael B.

Sinclair. Broken symmetry dielectric resonators for high quality factor fano metasurfaces. *ACS Photonics*, 3(12):2362–2367, nov 2016.

[143] Andrey B. Evlyukhin, Carsten Reinhardt, Andreas Seidel, Boris S. Luk'yanchuk, and Boris N. Chichkov. Optical response features of si-nanoparticle arrays. *Physical Review B*, 82(4), jul 2010.

[144] P. T. Leung, S. Y. Liu, and K. Young. Completeness and time-independent perturbation of the quasinormal modes of an absorptive and leaky cavity. *Physical Review A*, 49(5):3982–3989, may 1994.

[145] Q. Bai, M. Perrin, C. Sauvan, J-P Hugonin, and P. Lalanne. Efficient and intuitive method for the analysis of light scattering by a resonant nanostructure. *Optics Express*, 21(22):27371, nov 2013.

[146] Jianji Yang, Harald Giessen, and Philippe Lalanne. Simple analytical expression for the peak-frequency shifts of plasmonic resonances for sensing. *Nano Letters*, 15(5):3439–3444, apr 2015.

[147] P. Lalanne, W. Yan, A. Gras, C. Sauvan, J.-P. Hugonin, M. Besbes, G. Demésy, M. D. Truong, B. Gralak, F. Zolla, A. Nicolet, F. Binkowski, L. Zschiedrich, S. Burger, J. Zimmerling, R. Remis, P. Urbach, H. T. Liu, and T. Weiss. Quasinormal mode solvers for resonators with dispersive materials. *Journal of the Optical Society of America A*, 36(4):686, apr 2019.

[148] Andre-Pierre Blanchard-Dionne and Michel Meunier. Multiperiodic nanohole array for high precision sensing. *Nanophotonics*, 8(2):325–329, oct 2018.

[149] Jeffrey C. Lagarias, James A. Reeds, Margaret H. Wright, and Paul E. Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1):112–147, jan 1998.

[150] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, jan 1965.

[151] Xinjie Yu and Mitsuo Gen. Introduction to evolutionary algorithms. 2010.

[152] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, jan 2016.

[153] X. Garcia-Santiago, P.-I. Schneider, C. Rockstuhl, and S. Burger. Shape design of a reflecting surface using bayesian optimization. *Journal of Physics: Conference Series*, 963:012003, feb 2018.

[154] Christopher L. Holloway, Derik C. Love, Edward F. Kuester, Joshua A. Gordon, and David A. Hill. Use of generalized sheet transition conditions to model guided waves on

metasurfaces/metafilms. *IEEE Transactions on Antennas and Propagation*, 60(11):5173–5186, nov 2012.

[155] Karim Achouri, Mohamed A. Salem, and Christophe Caloz. General metasurface synthesis based on susceptibility tensors. *IEEE Transactions on Antennas and Propagation*, 63(7):2977–2991, jul 2015.

[156] Nima Chamanara, Karim Achouri, and Christophe Caloz. Efficient analysis of metasurfaces in terms of spectral-domain GSTC integral equations. *IEEE Transactions on Antennas and Propagation*, 65(10):5340–5347, oct 2017.

[157] Xuchen Wang. *Surface-impedance engineering for advanced wave transformations.* PhD thesis, Aalto university, 2020.

[158] Mohamed A. Salem and Christophe Caloz. Manipulating light at distance by a metasurface using momentum transformation. *Optics Express*, 22(12):14530, jun 2014.

[159] Dong Cheon Kim, Andreas Hermerschmidt, Pavel Dyachenko, and Toralf Scharf. Adjoint method and inverse design for diffractive beam splitters. 2020. ArXiv:2001.04943v1.

[160] Walter Murray and Kien-Ming Ng. An algorithm for nonlinear optimization problems with binary variables. *Computational Optimization and Applications*, 47(2):257–288, dec 2008.

[161] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, Ltd, 2nd edition, may 2000.

[162] Minh Duy Truong, Guillaume Demesy, Frederic Zolla, and Andre Nicolet. Dispersive quasi-normal mode (DQNM) expansion in open and periodic nanophotonic structures. In *2019 22nd International Conference on the Computation of Electromagnetic Fields (COMPUMAG)*. IEEE, jul 2019.

Publications

Reviewed papers

Kevin Müller, Raoul Kirner, Wilfried Noell, Toralf Scharf, and Reinhard Voelkel. Design principles of scanning multi-aperture imaging systems: phase-space diagram, system efficiency and contrast. *OSA Continuum*. [submitted, under review]

Kevin Müller, Ivan Fernandez-Corbaton, Fernando Negredo, and Carsten Rockstuhl. The Poynting operation: An electro-magnetic power-flow based operation for lossless z-invariant media. *Journal of Physics: Photonics*. [rejected, under review]

Proceedings

Raoul Kirner, Kevin Mueller, Pauline Malaurie, Uwe Vogler, Wilfried Noell, Toralf Scharf, and Reinhard Voelkel. Array imaging system for lithography. In *Optical System Alignment, Tolerancing, and Verification X*. International Society for Optics and Photonics, 2016.

Kevin Müller. Self-Coupling Modes in Periodic Resonant Metasurfaces. In *Integrated Photonics Research, Silicon and Nanophotonics*. Optical Society of America, 2018.

Kevin Müller. Analysis of Resonances in Periodic Metasurfaces through the Concept of Self-Coupling Mode. In *2019 Thirteenth International Congress on Artificial Materials for Novel Wave Phenomena (Metamaterials)*. IEEE, 2019.

Conference presentations

The Poynting operation: a tool to analyze and orthonormalize modes in metasurface. *EOS Topical Meeting on Diffractive Optics*, Joensuu, 2017. (poster presentation)

Binary dielectric metasurface for infra-red applications. *36th EUPROMETA School*, Karlsruhe, 2018. (poster presentation)

Self-Coupling Modes in Periodic Resonant Metasurfaces. *OSA Advanced Photonics Congress*, Zürich, 2018. (poster presentation)

Analysis of resonances in periodic metasurfaces through the concept of self-coupling modes. *12th Annual Meeting Photonics Devices*, Berlin, 2019. (**oral** presentation)

Interpolation of resonant metasurface's spectra using the concept of self-coupling mode. *Swiss NanoConvention*, Lausanne, 2019. (poster presentation)

Analysis of Resonances in Periodic Metasurfaces through the Concept of Self-Coupling Mode. *13th International Congress on Artificial Materials for Novel Wave Phenomena*, Rome, 2019. (**oral** presentation)

Kevin Müller

E-mail: kevin.mueller@alumni.epfl.ch
Language: French (native), English

Nationality: Swiss
Date of Birth: September 16, 1990
Place of Birth: Neuchâtel

Education and employment

until July 2008	Obligatory school and "lycée" in Neuchâtel
Sep 2008 – July 2010	Undergraduate study in EPFL (1st and 2nd year) in the Microengineering faculty
Sep 2010 – Apr 2011	3rd year as an exchange student at Waterloo University (Canada)
July 2011 – Apr 2012	Military service as "Soldat de transmission"
Sep 2012 – July 2013	Graduate study at EPFL in Optics. (1st year)
Aug 2013 – Jan 2014	Internship at SUSS MicroOptics
Principal activity:	Design of diffractive optical elements
Feb 2014 – July 2014	Graduate study at EPFL (Master project)
Oct 2014 – May 2015	Internship at SUSS MicroOptics
Principal activity:	Design of diffractive optical elements and lens system
May 2015 – Sep 2015	Martial art study in the Si Ping city Shaolin martial art academy (China)
Oct 2015 – Jan 2016	Internship at SUSS MicroOptics
Principal activity:	Design of diffractive optical elements and lens system
Jan 2016 – July 2016	Scientist at EPFL in the Optics & Photonics Technology laboratory
Sep 2016 – now	PhD thesis on binary dielectric metasurface
Supervisors:	Dr. Toralf Scharf and Dr. Wilfried Noell
Sep 2016 – Mar 2017	In the group of Prof. Carsten Rockstuhl at KIT
Apr 2017 – Sep 2018	At SUSS MicroOptics
Oct 2018 – now	In the group of Prof. Olivier Martin at EPFL

Semester and master projects

Sep 2012 – Dec 2012
High precision measurement of light characteristics by speckle decorrelation (Dr. Toralf Scharf)

Feb 2013 – July 2013
Engineering of scattering in multiple resonance plasmonic structures (Prof. Olivier Martin)

Feb 2014 – July 2014
Nanoparticle assembly: scalable process development from nanoscale molding to hardware/software implementation (Prof. Jürgen Brugger)

Interests

Shaolin Kung Fu, discofox (dance), running, video game, computer programming, calisthenic , chess