Advanced modeling of CMOS imagers

Jérémy Grebot

Modélisation avancée des imageurs nanostructurés

Jérémy Grebot

Centre Inria d'Université Côte d'Azur, Équipe ATLANTIS,

Présentée en vue de l'obtention du grade de docteur en Mathématiques d'Université Côte d'Azur
Dirigée par : Stéphane Lanteri
Co-encadrée par : *Claire Scheid, Denis Rideau*
Soutenue le : 23 mai 2024

Devant le jury, composé de :

Henri Benisty,	Président, Professeur des Universités, Université Paris XI
Christophe Sauvan,	Rapporteur, Chargé de recherche, Université Paris-Saclay
Raphaël Clerc,	Rapporteur, Professeur des Universités, Université Jean-Monnet
Denis Rideau,	Co-encadrant, Ingénieur, STMicroelectronics
Claire Scheid,	Co-encadrante, Maître de Conférence, Université Côte d'Azur
Stéphane Lanteri,	Directeur, Directeur de Recherche, Université Côte d'Azur

Résumé: Cette thèse se concentre sur les méthodes numériques utilisées dans l'amélioration de l'absorption lumineuse des capteurs d'images SPAD. En particulier, nous étudions trois aspects méthodologiques, à propos des matériaux utilisés, à propos des solveurs optiques et à propos d'une méthode de design inverse en vue d'obtenir le capteur d'image optimal.

Le premier chapitre étudie l'usage de matériaux innovants, avec une focalisation sur le Silicium, Germanium et leurs alliages (SiGe). Les données expérimentales étant lacunaires, nous proposons un modèle semi-empirique pour la permittivité du SiGe, dépendant de la température et de la concentration de Ge. Une attention spéciale a été apportée à la comparaison avec la méthodologie usuelle, l'interpolation linéaire. Le modèle proposé repose sur des modèles disponibles dans la littérature, pour lesquels certains paramètres spécifiques, les bandes interdites, directes et indirectes, sont extraits par la méthode de Tight-Binding, d'où le nom de semi-empirique. Cette méthodologie a été publiée dans la $51^{ème}$ European Solid-State Device Research Conference (ESSDERC 2021).

Dans le second chapitre, nous comparons la méthode numérique de référence pour résoudre les équations de Maxwell, la méthode FDTD, et deux méthodes alternatives, DGTD et RCWA, pour simuler la réponse optique dans les SPADs fabriqués à STMicroelectronics. Les performances de chacunes de ces méthodes ont été comparées sur des structures d'une complexité croissante. Les profils d'absorption résultant sont ainsi comparés, ainsi que les temps d'exécutions. La contribution principale de ce chapitre est l'écriture, pour les ingénieurs de STMicroelectronics, d'un solveur RCWA entièrement fonctionnel et versatil. Les performances de ce solveur sont comparées au solveur commercial de référence, Lumerical, et le solveur DGTD écrit par l'équipe Atlantis de l'Inria Sophia-Antipolis. Cette comparaison confirme le status de référence de Lumerical. Toutefois cela n'implique pas nécessairement une supériorité intrinsèque à la méthode numérique sous-jacente, la FDTD, comme indiqué en conclusion de ce chapitre.

Dans le troisième, nous proposons une méthodologie de design inverse qui combine un solveur optique avec une méthode statistique d'optimisation pour la découverte efficace des paramètres optimaux maximisant l'absorption lumineuse. En particulier, la structuration est étudiée, vu que celle-ci a apporté les gains d'absorption les plus important récemment. Une structure optimale atteignant une absorption de 83% a été trouvée, dépassant ainsi toutes valeurs analogues présentes dans la littérature. Cette méthodologie a été présentée à la conférence SISPAD 2023.

Mots-clefs Méthodes numériques, RCWA, INRIA, STMicroelectronics, Imageurs, Modélisation.

Summary: This thesis focuses on the numerical methodology for the optical improvement of SPADs devices through the increase of the light absorption. In particular, we study three methodological aspects: about materials, about solvers, and about the process to obtain the best performing device.

The first chapter investigates the usage of innovative material, with a focus on Silicion, Germanium and their alloys (SiGe). In the absence of permittivity data from the literature, we provide a semi-empirical model for the permittivity of SiGe as a function of both temperature and Ge content. A specific attention has been paid on the comparison with the usual methodology applied when facing a lack of measured data, the linear interpolation method. The permittivity model provided relies on usual permittivity models found in litterature. However, specific parameters of these models, both direct and indirect bandgaps, are extracted from band structure computation by Tight-Binding, hence our model is said semi-empirical. This methodology was published in the 51^{st} European Solid-State Device Research Conference (ESSDERC 2021) [1].

In the second chapter, we compare the reference numerical method for solving the Maxwell's equations, the FDTD, and two alternatives, the DGTD and the RCWA, to simulate the optical response of SPADs device fabricated by STMicroelectronics. A benchmark on the structure of increasing complexity is performed. The resulting absorption spectra of the three numerical methods are compared, as well as their time execution. The main contribution of this chapter is the delivery to STMicroelectronics engineers of a fully functionnal and versatile 2D and 3D RCWA solver. The performance of this solver were then compared to the reference FDTD solver, Lumerical, and the Inria DGTD solver, Diogenes. The conclusion of this benchmark confirms the leading position of the Lumerical solver. However, this does not necessary imply a superiority of the underlying numerical method, the FDTD, as discussed in the conclusion of this chapter.

In the third chapter, we propose a inverse-design methodology that combines optical solvers with a statistical learning-based optimization for goal-oriented discovery of the optimal parameters for maximizing light absorption. In particular, the diffractive gratings are studied, since they recently provided an important performance increase. An optimal structure reaching 83% absorption has been found, outclassing all previous absorption level found in litterature. This methodology was presented in the 2023 International Conference on Simulation of Semicondutor Processes and Devices (SISPAD 2023).

Keywords Numerical method, RCWA, INRIA, STMicroelectronics, Imageurs, Modelling.

Contents

Acknowledgement

Je tiens à remercier, pour l'ensemble de son suivi sans failles ni imperfections, mon directeur de thèse Stéphane Lanteri. Ses remarques pertinentes et sa vive acuité ont été des soutiens inébranlables dans l'élaboration et l'écriture de ce manuscrit. Je souhaite également remercier mon encadrant industriel, Denis Rideau, toujours à même de pousser les sujets à leur paroxysme. Notre collaboration fut une force constante et un devenir commun sans égal. Je te remercie pour ces années et ces joies. Enfin, Claire, je te remercie pour ton implication dans cette thèse, envers et contre toutes circonstances. La maturité de ce manuscrit doit beaucoup à tes relectures exigentes et minitieuses.

Je veux également remercier tous celles et ceux qui, trois années durant, m'ont encouragés et accompagnés dans ce long et éreintant travail d'écriture. Marion, Célia, Valentin et Azad pour avoir sonoriser mon immeuble. Constantin, Benjamin W. et Alexandre pour vos multiples messages. Clémence et Héléna évidemment. Robin, Rémi, Pierre-Louis, Gabriel et Yassine pour votre partage constant. Loïc, Antonin, Bruno, Grégoire, Maria, Nour et Thomas D. pour ces belles soirées. Jules et Elvire pour votre force sans égale et ces périples à vélos. Thomas C., Sébastien, Pascal, Guillaume M., Edgar, Olivier, Benjamin V., Gaëlle, Isobel, Andres, Dylan, Pierre, Hélène, Marios et Valérie pour ces moments de partages sur Crolles. Et enfin, Guillaume L., Alexis et Théophile pour m'avoir accompagné dans mon périple sudiste.

Je tiens également à remercier, évidemment, mes parents sans qui rien de tout ceci n'eût été possible, et mon frère dont le soutien, tel une force tranquille et un cap ferme, m'aura beaucoup apporté.

Introduction

In 1865, James Maxwell laid the foundations of modern electromagnetism. Its main contribution are the well-known Maxwell's equations, that formally link electric and magnetic fields. This theoretical breakthrough is nowaday more than 158 years old and numerous studies of these equations have been performed since. Electromagnetism applications are found everywhere and everyday: from wireless communication, including radio transmitter and smartphone, to optical fiber and medical imaging, or even magnetic lift train.

Among the most astonishing applications are the metasurfaces, which are thin etched surfaces that allow modulating in phase the reflected or transmitted wave [2]. With such property, a spherical lense, that concentrate light on a focal point, can be replaced by a single patterned layer, a metasurface, enhancing the miniaturization of CMOS imagers. Also, thanks to optothermal heating [3], nanoparticles are used in medicine to target and kill specifically cancer cells. Optical lidar, used in autonomous cars [4], allows a real time space mapping around the car.

This thesis is concerned with Nanophotonics, the physical science that studies the interactions of light with matter at the nanoscale, and which can be modelled by Maxwell's equations. When the studied structure size is approaching the wavelength of light, specific phenomena occur: geometrical optics, such as raytracing methodology, is no longer relevant and solving Maxwell's equations is mandatory. The well-known diffraction of light passing through a hole, illustrated with Fig. 1, is one example of physical phenomena that is modeled by Nanophotonics.

Motivations of an industrial thesis are economic, and our work does not deviate from the rule: it takes place within the context of continuous improvement of devices fabricated by STMicroelectronics, and in particular of Complementary Metal Oxide Semiconductors (CMOS) imagers. In 2020, the total market-cap is estimated at 20 billions, and STMicroelectronics, which funded this thesis [1], is positioned in fourth place, after Sony, Samsung and Omnivision (see Fig. 2). This market-cap indicates the various applications of CMOS imagers: smartphone, informatic, security, automobile and self-driving cars.

Among the CMOS imagers, our work focuses on the Single Photon Avalanche Diode (SPAD). The main applications are in small range lidar, used in the autofocus for smartphone camera. The latest Iphone, commercialized by Apple, include the SPADs fabricated at STMicroelectronics Crolles. This diode is used as a photodetector for the Near Infrared (NIR) light, in time-of-light device: from the time interval between the emission of an

[1]This thesis was also funded with a subsidy from ANRT (French National Association of Research and Technology) with the CIFRE (Industrial Convention of Training by Research) convention n°2019/1772.

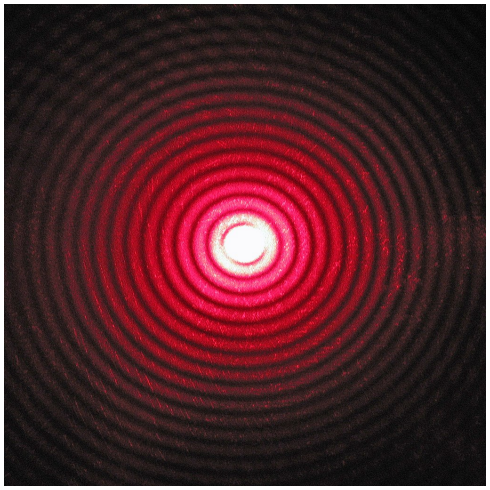


Figure 1: A diffractive pattern of a red laser beam passing through a small circular aperture. Figure taken from [5].

infrared signal, its reflection on the target, and its reception by the SPADs, the distance between the target and the device can be known. Thus, this proximity sensor provides, in real time, the distance between the nearest object to the camera of the smartphone, adjusts automatically the lens of the camera, ensuring that selfies are vivid and clear. Infrared light is mainly used because it is not visible to human eyes. And such device typically uses a signal of wavelength equals to 940 nm.

SPADs, according to their name, are able to detect a single photon, namely to convert a single photon into a current pulse. More precisely, photogenerated electrons are collected and drift toward the avalanche region, where the exponentially increasing impact ionizations generate more carriers, leading to an avalanche of carriers, and then to a current pulse (see Fig 3). For more details on the electrical modeling of SPAD, we advise the recent paper [6], which modeled both the carrier trajectory and the multiple impact ionizations on SPADs device. Our work focuses on the optical modeling of SPADs, and it can be seen as a preliminary step of such electrical modeling. From the optical point of view, the SPADs performances are judged not on the carrier's transport and generation, but only on the percentage of the incident light that is absorbed.

In order to reduce both the development time and cost, numerical methods are mandatory. They make it possible to quickly validate a corrective action, solve complex problems by splitting them into subproblems, optimize the sizing of installations, access non-measurable quantities, better understand the physical phenomena involved, but above all, reduce costs by solving problems before they occur.

In the early 2000s, SPADs were 10 μm to 40 μm wide [8]. The next decades had improved their performance by shrinking their size, up to 4 μm wide [9]. This increase of performance by downscaling semiconductors is famously known as the Moore Law, which predicts that the number of transistors in an integrated circuit doubles every two years.

For predicting SPADs optical performance, this downscaling has a major consequence: the size of the device is now approaching five times the wavelength of interest (940 nm), and some of its parts are even smaller, to the scale of half the wavelength of interest.

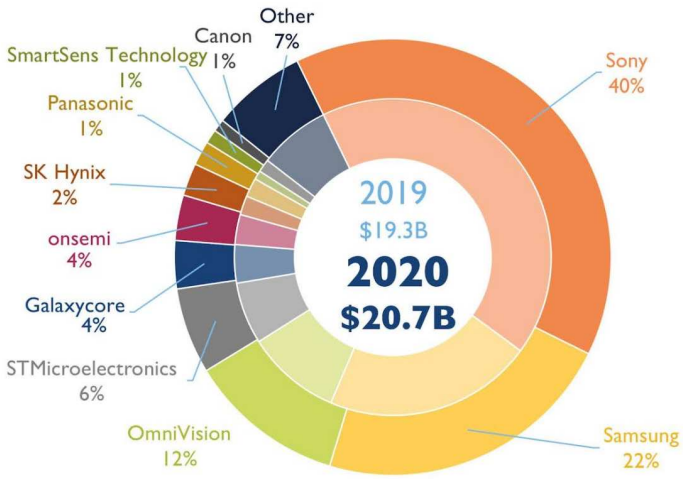


Figure 2: CMOS imagers companies, classified according to their market shares, in 2020. Figure taken from [7].

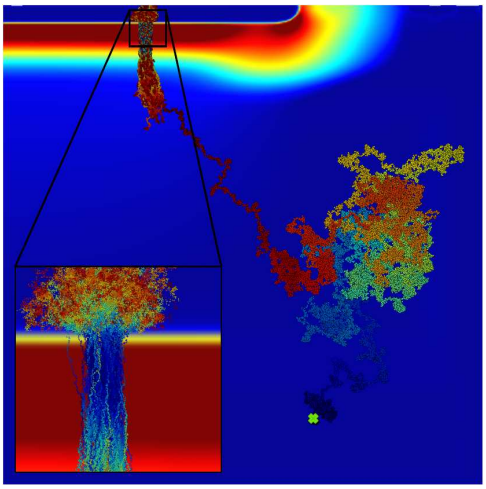


Figure 3: 2D Illustration of a multi-particle simulation of a SPAD. The background represents the strength of the electric field, the red region corresponds to the SPAD main junction. An electron is injected in the volume of the SPAD (yellow cross), it drifts and diffuses towards the junction and triggers several impact ionization events, leading to a self-sustained avalanche. The color of the trajectories' dot represents the intrinsic time of the particles. A zoomed view of the avalanche is shown in the inset. Figure taken from [6].

This means that the numerical methods used to predict the optical response cannot rely on geometrical optics and must solve exactly the Maxwell equations.

When the SPADs were 10 to 40 μm wide, the usual numerical method used was the Transfer-Matrix Method (TMM), which predicts the reflection, transmission and absorption of light within a stack of layers, see [10] for details. Such a numerical method relies on two assumptions: the layers interfaces are smooth, and only the thickness of every layer (only the z dimension) is of interest. The TMM is based on the fact that, according to Maxwell's equations, there are simple continuity conditions for the electric field across boundaries from one medium to the next. If the field is known at the beginning of a layer, the field at the end of the layer can be derived from a simple matrix operation. A stack of layers can then be represented as a system matrix, which is the product of the individual layer matrices. The final step of the method involves converting the system matrix back into reflection and transmission coefficients.

When the SPADs size has approached the wavelength of interest (940 nm), predicting optical performance required numerical methods solving Maxwell's equations. The reference method for many years has been the Finite Difference Time-Domain (FDTD) method or Yee's method [11]. The main reasons for its popularity are its ease of implementation and its computational efficiency. It relies on a cartesian structured discretization of the simulated structure, enhancing the well-known staircasing effect, that can become a significant limitation when the local rounded details must be taken into account. To overcome such limitations, various methods have been proposed, among them the Discontinuous Galerkin Time-Domain (DGTD) method [12]. It is a discontinuous finite element type method that relies on a high-order interpolation of the electromagnetic fields within each cell of an unstructured mesh. Time integration can be achieved using an explicit scheme and no global mass matrix inversion is required to advance the solution at each time-step. This means that this method is well-suited to massively parallel computing. Apart from the time-domain numerical method, we ought to mention the mainly used in frequency-domain, the fourier method named the Right Coupling Wave Analysis (RCWA). The fields are represented as a sum of spatial harmonics [13], and the simulated structure is discretized into uniform layers along the light propagation axis.

All those numerical methods rely on the optical properties of the matter used: the permittivity. From the point of view of an optical engineer, the permittivity is seen as an input for the optical simulations. The precision and prediction of such simulations are intrinsically dependent on the accuracy of such quantities. The vast majority of SPADs are made of Silicon (Si), and thus the permittivity of Silicon must be measured with high precision in order to provide reliable optical simulation. Alternative materials, such as Germanium (Ge), are promising [14], but the permittivity data available in literature are often lacking, especially when measured at various temperatures. In case of lacking data, optical engineers tend to linearly interpolate the permittivity to perform optical simulation.

This thesis focuses on the numerical methodology for the optical improvement of SPADs devices through the increase of the light absorption. In particular, we study three methodological aspects: about materials, about solvers, and about the process to obtain the best performing device.

The first chapter investigates the usage of innovative material, with a focus on Silicion,

Germanium and their alloys (SiGe). In the absence of permittivity data from the literature, we provide a semi-empirical model for the permittivity of SiGe as a function of both temperature and Ge content. A specific attention has been paid on the comparison with the usual methodology applied when facing a lack of measured data, the linear interpolation method. The permittivity model provided relies on usual permittivity models found in litterature. However, specific parameters of these models, both direct and indirect bandgaps, are extracted from band structure computation by Tight-Binding, hence our model is said semi-empirical. This methodology was published in the 51st European Solid-State Device Research Conference (ESSDERC 2021) [1].

In the second chapter, we compare the reference numerical method for solving the Maxwell's equations, the FDTD, and two alternatives, the DGTD and the RCWA, to simulate the optical response of SPADs device fabricated by STMicroelectronics. A benchmark on the structure of increasing complexity is performed. The resulting absorption spectra of the three numerical methods are compared, as well as their time execution. The main contribution of this chapter is the delivery to STMicroelectronics engineers of a fully functionnal and versatile 2D and 3D RCWA solver. The performance of this solver were then compared to the reference FDTD solver, Lumerical, and the Inria DGTD solver, Diogenes. The conclusion of this benchmark confirms the leading position of the Lumerical solver. However, this does not necessary imply a superiority of the underlying numerical method, the FDTD, as discussed in the conclusion of this chapter.

In the third chapter, we propose a inverse-design methodology that combines optical solvers with a statistical learning-based optimization for goal-oriented discovery of the optimal parameters for maximizing light absorption. In particular, the diffractive gratings are studied, since they recently provided an important performance increase. An optimal structure reaching 83% absorption has been found, outclassing all previous absorption level found in litterature. This methodology was presented in the 2023 International Conference on Simulation of Semicondutor Processes and Devices (SISPAD 2023).

Chapter I

Permittivity of SiGe accounting for strain and temperature

Contents

I.1 Introduction

SiGe alloys are becoming widely used in optoelectronics, for applications such as imaging devices, photodetectors and Single-Photon Avalanche Diodes (SPAD). Their small and tunable bandgap make them candidates for application in Short-Wavelength InfraRed (SWIR) or Near-InfraRed (NIR) range. For instance in [14], one can find electrical simulations and measurements on a Ge-on-Si planar SPAD demonstrating single-photon detection efficiency (PDE) of 38% at 125 K at a wavelength of 1310 nm. The PDE is the overall conversion factor from photons to the number of detectable photoelectrons. In order to deepen the modelling process of these devices, simulating the optical propagation and investigating the influence of temperature or strain on optical absorption are essential.

One example of such study, focusing only on light propagation, can be found in [15] where the authors improved SPAD sensitivity by designing diffractive microlens with FDTD simulations. These optical simulations intrinsically require optical properties as input, such as the complex refractive index $N = n + ik$ or the complex permittivity $\varepsilon = \varepsilon_r + i\varepsilon_i$. However the lack of parametric optical constants data for Si, Ge and their alloys is the main obstacle for performing optical simulations of SiGe SPADs and for improving their design on a realistic set of temperature or strain values. Indeed, usually only electrical simulations are performed, see for instance [14].

On the one hand, measurement data obtained with ellipsometry are usually only available for a discrete set of conditions (temperature, Ge concentration and strain). They are usually reachable with handbooks of optical constants such as [16], or online databases such as [17], where the most precise and up-to-date measurements are compiled with discussion on experimental conditions.

On the other hand, quantum mechanical-based physical models [18, 19, 20, 21] can provide useful information on the optical absorption in semiconductors but they require heavy computations (such as Time-Dependent Density Functional Theory (TDDFT) or Bethe-Salpeter equation) and lack accuracy so far when compared to measurements, in particular when the variation in temperature is considered. This is even more critical for indirect optical transitions where both phonons and bandgap temperature effects must be accurately accounted for in the calculation. Recent works along this line [22] introduce for example extra empirical temperature-dependent bandshift to the rigorous *ab initio* calculation in order to match experimental Silicon optical absorption in the NIR.

A large variety of pragmatic and empirical models exists in the literature and are widely used in domains such as ellipsometry and optical simulations. Such models rely on the oscillators, *i.e.* complex valued functions that respect the Kramers Kronig relations (see Eq. I.6). For instance, various absorption measurement data for Si are compared in [23], leading to an analysis of measurement uncertainties and providing an empirical model of the absorption coefficient of Si for a wavelength range starting from 250 nm up to 1450 nm. It consists of five Tauc-Lorentz oscillators and two gaussian shaped oscillators (see I.2.3). A complete description of these models including Sellmeier, Cauchy, Lorentz, Drude, Gaussian, Tauc-Lorentz or Cody-Lorentz can be found in [24]. Their success in accurately predicting the full spectrum of the dielectric response (being compliant with Kramers-Kronig relations) is well-established. However they require the fitting of several parameters for each condition: material, strain and temperature. To the best of our

knowledge, there is not a general set of parameters and models capable of handling all the above-mentioned conditions in SiGe. Along this line of generalization, we ought to highlight the aforementioned paper of Schinke et al.[23] which models the full spectrum of Si absorption coefficient, and Emminger et al. [25] which provides a temperature-dependent model of the complex permittivity of Ge well-aligned with ellipsometry measurements.

In this chapter, we propose a generalized model that can reproduce (strained) SiGe temperature-dependent optical properties in the NIR and visible spectrum (0-5 eV) and satisfies Kramers-Kronig relations. Extending the empirical approach of Schinke [23] and Emminger [25] with the optical gaps at critical points determined with the Tight-Binding (TB) band structure model [26], the model accurately accounts for SiGe with arbitrary strain.

I.2 Theory

In this section we present the theoretical background of the proposed semi-empirical model for the permittivity of SiGe. In the first subsection, basic definitions are outlined. Then, we introduce the modelling of permittivity from quantum theory, its advantages and its drawbacks. Next, the area of research in physics that focuses on measurements of the optical constants, the ellipsometry, is presented, as well as the concept of critical points of permittivity. Finally we summarize and present all the experimental data available for the permittivity of Si, Ge and their alloys.

I.2.1 Preliminaries

In this subsection are presented the basics definitions of optical constants.

I.2.1.1 Optical constants

Light passing through a material is altered by optical properties of this material. Firstly, each medium is characterized by its capacity to slow down light, which is modelled by its **index of refraction**, denoted as n and defined as $n = \frac{c}{v}$, where c denotes the speed of light in vacuum and v the speed of light in the medium. Secondly, each medium is characterized by its capacity to absorb light, which is modelled by its **extinction coefficient**, denoted as k, which is equal to zero for vacuum.

These two quantities lead to the definition of the **complex valued refractive index** as $N = n + ik$. With the refractive index, one can define, as done in [27], the **complex valued permittivity**, ε and the **absorption coefficient**, denoted as α.

If one considers, in frequency-domain, a linear, isotropic, homogeneous and non-dispersive material, then N, ε and α are real constants. If one considers a dispersive-material then N, ε and α are complex valued frequency-dependent functions.

In the literature, four variables are commonly used as input variables for the permittivity of dispersive media: the wavelength noted λ, the frequency, f, the pulsation ω and

the energy E. All these variables are linked by the following relations in a vacuum:

$$E = \hbar\omega = hf = \frac{hc}{\lambda}, \text{ and } \omega = 2\pi f, \quad \hbar = \frac{h}{2\pi}, \tag{I.1}$$

where h is the plank constant and c is the speed of light in a vacuum. In the relations Eq. I.1, λ is in m, f in s^{-1}, ω in rad·s^{-1} and E in J, since h is in J·s and c is in m·s^{-1}. Another unit for E is commonly used, the electronvolt, noted eV, where one has $1\text{eV} = 1.601e^{-19}\text{J}$. Throughout the rest of this chapter, the energy, in eV, is chosen in order to be consistent with the quantum physics notations.

The permittivity is defined as:

$$\varepsilon(E) = N(E)^2. \tag{I.2}$$

If ε_r, resp. ε_i, designates the real, resp. imaginary, part of ε, then one has:

$$\varepsilon_r(E) = n(E)^2 - k(E)^2 \qquad \text{and} \qquad \varepsilon_i = 2n(E)k(E). \tag{I.3}$$

The absorption coefficient is then given by:

$$\alpha(E) = \frac{4\pi}{hc}Ek(E). \tag{I.4}$$

The relative permittivity is the permittivity of a material expressed as a ratio with the electric permittivity of a vacuum, ε_0:

$$\varepsilon_r = \frac{\varepsilon}{\varepsilon_0}. \tag{I.5}$$

The notation ε_r is now ambiguous. In Eq. I.5 it stands for the relative permittivity while in Eq. I.3, it refers to the real part of the permittivity. In the following, ε_r will be only used to designate the real part of the permittivity. Also, the "permittivity" or the "relative permittivity" are used without distinction to designate the relative permittivity, noted simply ε.

The initial definitions of optical constants being recalled, the next section define the most important property of these optical constants.

I.2.1.2 Kramers-Kronig relations

For a physical system, it is crucial to satisfy the causality principle. Mathematical causality is ensured for permittivity thanks to the Kramers-Kronig (KK) relations. For an extensive presentation of all KK relations we refer the reader to [28]. These relations link ε_r and ε_i, respectively the real part and the imaginary part of permittivity, by the following relations, for ω in $\mathbb{R}^+$:

$$\begin{aligned} \varepsilon_r(\omega) &= \frac{1}{\pi}P\int_0^\infty \frac{\varepsilon_i(\omega')}{\omega' - \omega}\mathrm{d}\omega' \,, \\ \varepsilon_i(\omega) &= \frac{1}{\pi}P\int_0^\infty \frac{\varepsilon_r(\omega')}{\omega' - \omega}\mathrm{d}\omega' \,, \end{aligned} \tag{I.6}$$

where we denoted by P the Cauchy-Principal value, i.e., if $f : \mathbb{R} \to \mathbb{R}$ has only one singularity in $a \in \mathbb{R}$ we define the Cauchy-Principal value as:

$$P \int_{-\infty}^{\infty} f(x)\mathrm{d}x = \lim_{\substack{\eta \to a \\ \eta > 0}} \left(\int_{-\infty}^{a-\eta} f(x)\mathrm{d}x + \int_{a+\eta}^{\infty} f(x)\mathrm{d}x \right) . \tag{I.7}$$

In all generality, one can deduce the real part from the imaginary part and *vice-versa*. But KK relations are based on a non-local operator: in order to reconstruct ε_r from ε_i for instance, one must know ε_i all over $\mathbb{R}^+$. This is a serious limitation of the direct use of the formula since measurements of ε_i, or ε_r are usually provided only on a sub-interval of $\mathbb{R}$. In particular the low convergence of the integral (of order $\frac{1}{\omega}$) increases errors in any extrapolation strategies. See [28] for more details.

I.2.2 Quantum theory

In the following, we briefly present the modelling of permittivity from quantum theory. Firstly, the difference between the classical optical formalism and the quantum one are clarified. The main concept related to quantum-physics, the band structure, is defined in the second section. Then indirect (resp. direct) electronic transitions are explained in the third (resp. fourth) section.

I.2.2.1 From macroscopic to microscopic formalism

The classical and quantum formalism both model the interaction between light and matter. The differences between these two modelling approaches are briefly clarified in the following paragraphs.

The basic definitions of optical constants originate from optics (see [29] for an introduction). In this physical theory, light is a wave and the elementary unit is the "matter" that has properties: the optical constants defined above. Then, with this modelling, it is possible with Maxwell's equations, to compute the light propagation and absorption through any material. A typical experiments is the white light decomposition with a prism or the Young slit experiment, where a laser passing through a double slit exhibits a diffractive pattern. Everyday confirmations of the classic optics theory are also available by looking at a bending stick in water or a rainbow in the sky. Optics theory is well established and no experiment has falsified its predictions. However, what can be said by physicists on the interaction between light and matter is far from being exhausted. With the formalization of quantum physics, a typical theoretical movement in physics epistemological history has occurred: a new modelling that deepens and does not contradict the previous theory. Similarly to the deduction of the laws of thermodynamics by statistical physics, quantum theory is able to deduce the optical properties from a totally different formalism.

The main difference between classical optics theory and the quantum approach is the following: the elementary unit is not "matter" but particles. For instance a bulk of Si is

not a pure and uniform entity but a lattice of atoms, constituted of electrons (that are particles), neutrons and protons (that are made of particles). Similarly, light is not only made of waves but also of photons.

Therefore light absorption on a macroscopic level is now, from a quantum perspective, an interaction between particles, and more precisely an interaction between electrons and photons, since we consider only visible and infrared light in this work (protons and neutrons can interact with ultraviolet light, but this is outside the scope of this work).

Particles are characterized by many properties but only two of them will be the focus of our attention: a particle has a wave vector and an energy. For instance in the case of a plane wave of frequency ν and wavevector $\mathbf{k}$, the corresponding photon has the energy $E = \hbar\nu$, where $\hbar$ is the plank constant.

As aforementioned, light absorption on a microscopic level is the absorption of a photon by an electron. If this interaction involves only these two particles then it results in an increase of energy for the electron. If this interaction also involves a phonon, a particle emitted or absorbed by the vibration of the crystal lattice, then it results in an increase in energy and a modification of the wave vector for the corresponding electron. These two types of interaction are named **direct**, when only a change of energy occurs, and **indirect**, when a change of energy and wave vector occurs. Since this involves a change of state for an electron, we will also refer to these two types of interaction as **direct transition** and **indirect transition** of the state of an electron (see [27]). For further details on this distinction, see Fig. I.3 and section I.2.2.2.

Now that the basic principles are recalled, we will define the levels of energy allowed for electrons, namely the band structure.

I.2.2.2 Band structure

In highs chool, we learn that electrons are small particles moving around an atomic nucleus in circles of different radius. The electrons moving on the furthest radius can escape to become free electrons. Each circle is called a band. From a mathematician's point of view, the band structure is a finite collection of scalar functions on $\mathbb{R}^3$.

Precisely, the band structure describes the possible states of energy for electrons. Each band is a function on the reciprocal space (noted $E(\mathbf{k})$ where $\mathbf{k}$ is the wave vector). For more convenience, bands are defined not on the whole reciprocal space but on its elementary unit: the **Brillouin zone** (BZ). This subset of the reciprocal space is briefly defined in the next paragraph.

A crystal is characterized by a regular array of atoms which repeat periodically in the (real) space. A Bravais lattice is defined as a regular periodic arrangement of points in space, all of them connected by translation vectors:

$$t_n = n_1 t_1 + n_2 t_2 + n_3 t_3. \tag{I.8}$$

The non-coplanar vectors t_1, t_2, t_3 are called primitive translation vectors and n_1, n_2, n_3 are any triplet of integer numbers. The parallelepiped formed by t_1, t_2, t_3 is called the primitive unit cell. In order to fully describe the geometry of a crystal, one needs to

specify how atoms are spread within the unit cell with a set of basis vectors. This basis indicates the coordinate of atoms inside the unit cell.

The reciprocal space is defined as the dual of the real space: given a crystal with primitive translation vectors t_1, t_2, t_3, its reciprocal lattice has three primitive vectors, g_1, g_2, g_3, defined by:

$$t_i \cdot g_j = 2\pi\delta_{ij}. \tag{I.9}$$

The BZ has the property that any point of the cell is closer to the chosen lattice point (say $g \equiv 0$) than to any other. Its shape is directly deduced from the geometry of the direct Bravais lattice.

The BZ of Si and Ge is shown in Fig. I.1. L, U, X, K, W and Γ are points of high symmetry of the BZ, also named **critical points**. For a detailed explanation on how the critical points are chosen, see section I.2.2.5. In Fig. I.1, Γ is in blue, dots and path on the boundary are in red, and paths from the center to the boundary are displayed in green.

In order to visualize a band defined on the BZ, a 1D path through this 3D volume is usually defined: $(L \to \Gamma \to X \to W \to U \to L \to W \to X \to K \to \Gamma)$ (see Fig. I.2 for an example). This 1D path was selected because it fully covers the smallest volume that generates the full BZ. Indeed, the volume defined by the dots Γ, L, U, X, W and K generates the full BZ with symmetry and rotation: first a mirror by the plane Γ, L, K, then three rotations with the axis (Γ, L) generate a 8th of the BZ, finally four rotations with the z axis and a mirror with the x-y plane finish the full BZ generation.

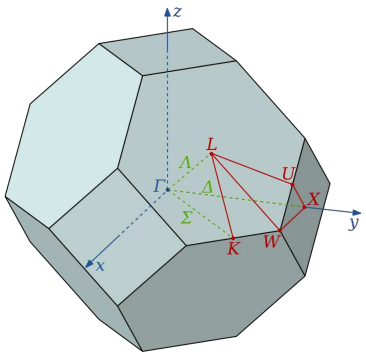


Figure I.1: Brillouin zone of Si and Ge. Figure extrated from `https://fr.wikipedia.org/wiki/Zone_de_Brillouin`.

As aforementioned, bands are scalar functions on the BZ and they describe the electronic state (energy level) that electrons can admit. Without entering into the subtleties of the band structure calculations, we ought to mention the Tight-Binding (TB) method (see [27] for a detailed definition) whose application for determining the band structure of Si, Ge, and their alloys has been done in [26]. The key point, for the present work, is that the band structure can be computed for any Ge content, at any strain, but only at temperature 0 K.

In Fig. I.2, the band structure of bulk Si and Ge, obtained with TB methods, implemented by [26], are shown. The x-axis is a 1D path passing through critical points inside

the BZ (see Fig. I.8). Energy on the y-axis is relative to the maximum of the valence band ($E = E' - \max_{\mathbf{k}\in BZ} E_v(\mathbf{k})$). Each red line is a band. Given an electron of wavevector $\mathbf{k}$ (for instance, $\mathbf{k} = \Gamma$), one can deduce, from the band structure graph, that the admissible relative energy of this electron is in a discrete set depending on which band this electron belongs to. In the following, "energy" will always refer to the "relative energy" defined above.

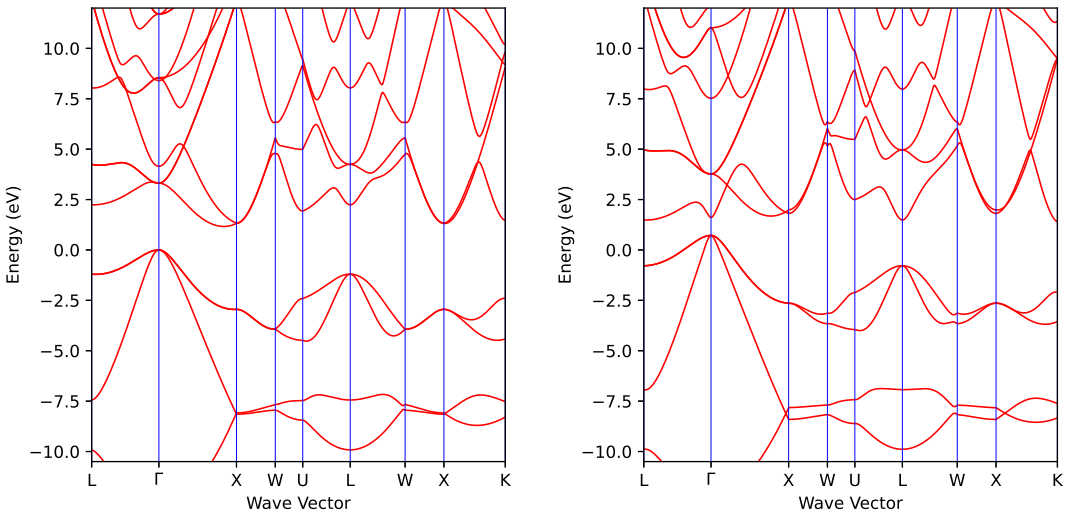


Figure I.2: Band structure of bulk Si (left) and Ge (right) obtained by TB model from [26].

Semiconductors are characterized by the existence of a gap in their band structure. It is clearly visible in Fig. I.2 where there are no bands between 0 eV to ~1 eV for Si and between 0 eV and ~0.9 eV for Ge. Such a forbidden energy interval is called a **bandgap**. The band below this bandgap is called the **valence band** and the one up this bandgap is called the **conduction band**. More generally one can define **direct bandgaps** and **indirect bandgaps** according to the type of interaction considered between two bands. Recalling that direct transitions involve a variation in energy of an electron while keeping its wavevector constant, it can be seen as a vertical arrow on a band structure diagram. Indirect transitions, involving an extra variation on the wave vector, can be seen as a diagonal arrow. The existence of the bandgaps implies that a photon with an energy less than the bandgap cannot achieve a direct transition on an electron situated in the valence band.

The figure Fig. I.3 highlights an indirect transition with a zoom around Γ on the Si band structure, computed with TB of [26]. The figure Fig. I.3 highlights a direct transition with a zoom around Γ of the Ge band structure.

Now that the basic principles have been outlined and the band structure has been defined we can focus, on the aforementioned deduction of the optical constants within the quantum mechanics framework which will be the object of the next two sections. More precisely, we describe interband electronic transitions in materials with a fully empty conduction band.

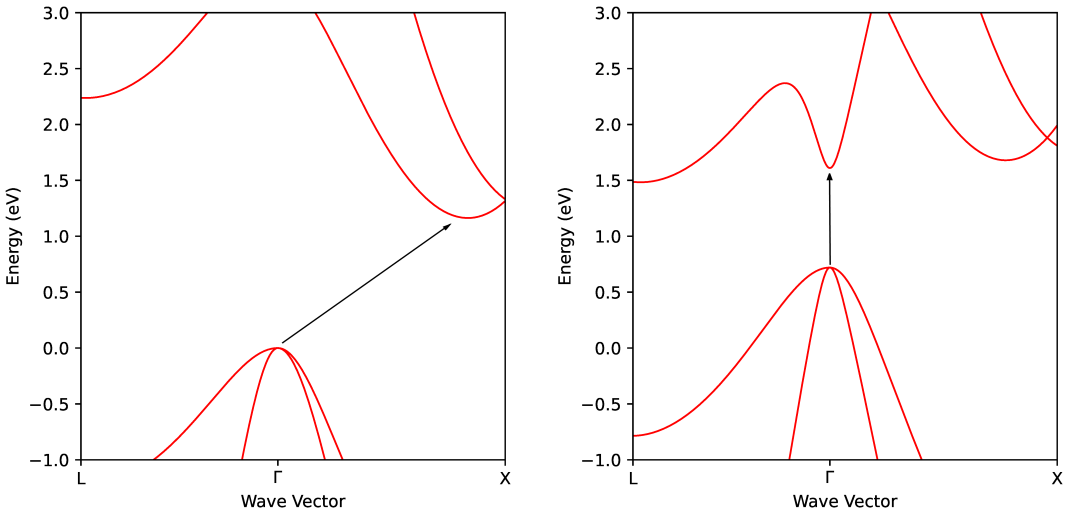


Figure I.3: Distinction between indirect (left) and direct (right) transitions. A bandgap is said direct, resp. indirect, when it corresponds to a direct, resp. indirect transition. For instance, Si, resp. Ge, has an indirect, resp. direct, bandgap in Γ. These figures are zooms of Si (left) and Ge (right) band structure shown in Fig. I.2.

Interband transitions occur when the electron jumps from one band to another. Intraband transitions, where the electron does not change its band, are not relevant for studying optical properties of semiconductors.

Firstly, we present the main formula of the absorption coefficient. Secondly, indirect phonon-assisted transitions are considered. And thirdly, direct transitions are detailed.

I.2.2.3 Absorption coefficient

Following Bassani and Pastori's work in [30], we present in this section the relation between the absorption coefficient and the band structure. We first define quantities associated with radiation of a given frequency ω in a medium of refractive index n.

The **vector potential**, noted $\mathbf{A}$, of the corresponding electromagnetic field, can be written as:

$$\mathbf{A}(\mathbf{r}, t) = A_0 \mathbf{e} \exp(i(\mathbf{k} \cdot \mathbf{r} - \omega t)) + c.c. \tag{I.10}$$

where $\mathbf{e}$ is the polarization vector in the direction of the electric field, $\mathbf{k}$ is the wavevector of the radiation, A_0 is the amplitude and $c.c.$ indicates the complex conjugate of the previous term.

The **average energy density**, u, of a radiation described by the vector potential $\mathbf{A}$ of Eq. I.10, is a real given by:

$$u = \frac{n^2 A_0^2 \omega^2}{2\pi c^2}. \tag{I.11}$$

The **energy flux** is the product of the average energy density and the velocity of the radiation in the medium, $\frac{c}{n}$. So it is a real equal to $u\frac{c}{n}$.

The absorption coefficient (defined in section I.2.1.1) is by definition the energy absorbed per unit time and volume divided by the energy flux:

$$\alpha(\omega) = \frac{\hbar\omega \mathbf{W}(\omega)}{u(c/n)}, \tag{I.12}$$

where $\hbar\omega\mathbf{W}(\omega)$ is the energy absorbed per unit volume and time. The number of transitions is noted $\mathbf{W}(\omega)$. Its computations depend on the type of interaction considered and it is available in [30] for both indirect and direct transitions.

In the two following sections, we present the main results obtained with Eq. I.12 in literature. The notation α_i (resp. α_d) stands for the absorption coefficient accounting for indirect (resp. direct) transitions. The absorption coefficient is the sum of α_i and α_d ($\alpha = \alpha_i + \alpha_d$).

I.2.2.4 Indirect phonon-assisted transitions

Decomposing the absorption coefficient accounting for indirect transitions as a sum on all indirect bandgaps contributions, *i.e.* writing:

$$\alpha_i = \sum_{g \in G} \alpha_{i,g} \tag{I.13}$$

where G is the discrete set of indirect bandgaps, Bassani and Pastori were able in [30] to model $\alpha_{i,g}$, the contribution of the bandgap noted g to the indirect absorption coefficient, thanks to Eq. I.12. Moreover they were able to identify the temperature dependency of α through the Bose-Einstein distribution (see below, Eq. I.17). Thus they retrieved theoretically the empirical model first introduced by Macfarlane in [31].

As already mentioned, indirect transitions involve an electron and a phonon. Their model is the sum of two contributions: according to absorption or emission phonons. Thus we have, given a radiation of frequency ω, a temperature T and an indirect bandgap g:

$$\alpha_{i,g}(\omega, T) = \alpha_{i,g,abs}(\omega, T) + \alpha_{i,g,emi}(\omega, T), \tag{I.14}$$

where for the case of phonon absorption:

$$\alpha_{i,g,abs}(\omega, T) = \begin{cases} 0, & \text{if } \hbar\omega \leq E_G - \hbar\omega_{q_0} \\ A_{abs}\ (\hbar\omega - E_g + \hbar\omega_{q_0})^2\ n_{q_0}(T) & \text{otherwise,} \end{cases} \tag{I.15}$$

and for the case of phonon emission:

$$\alpha_{i,g,em}(\omega, T) = \begin{cases} 0, & \text{if } \hbar\omega \leq E_G + \hbar\omega_{q_0} \\ A_{emi}\ (\hbar\omega - E_g - \hbar\omega_{q_0})^2\ (n_{q_0}(T) + 1), & \text{otherwise,} \end{cases} \tag{I.16}$$

where A_{abs} and A_{emi} are fitting parameters, E_g is the bandgap energy, $\hbar\omega_{q_0}$ is the phonon energy and $n_{q_0}(T)$ is the phonon Bose-Einstein distribution, defined as:

$$n_{q_0}(T) = \frac{1}{e^{\frac{\hbar\omega_{q_0}}{kT}} - 1}, \tag{I.17}$$

where k is the Boltzman constant.

This model leads to a unique formula on $\alpha_{i,g}$:

$$\alpha_{i,g}(\omega,T)=\begin{cases}0, & \text{if} \quad \hbar\omega \leq E_g-\hbar\omega_{q_0}\\ A_{abs}\ n_{q_0}(T)\ (\hbar\omega-E_g+\hbar\omega_{q_0})^2, & \text{if} \quad E_G-\hbar\omega_{q_0}<\hbar\omega\leq E_g+\hbar\omega_{q_0}\\ A_{abs}\ n_{q_0}(T)\ (\hbar\omega-E_g+\hbar\omega_{q_0})^2 \quad + & \\ A_{emi}\ (n_{q_0}(T)+1)\ (\hbar\omega-E_g-\hbar\omega_{q_0})^2, & \text{if} \quad \hbar\omega>E_g+\hbar\omega_{q_0}.\end{cases} \tag{I.18}$$

To summarize, a clear empirical model for the indirect transitions is available on the absorption coefficient.

I.2.2.5 Direct transitions

From Eq. I.12, Bassani and Pastori were able to provide the following equation for the absorption coefficient accounting for direct transitions (see [30]):

$$\alpha_d(\omega)=\frac{e^2}{nc\pi m^2\omega}\sum_{v,c}\int_{BZ}C(\mathbf{k})\delta(E_c(\mathbf{k})-E_v(\mathbf{k})-\hbar\omega)\mathrm{d}\mathbf{k} \tag{I.19}$$

where m is the electron effective mass, E_c (resp. E_v) denotes the conduction (resp. valence) band energy, C is a scalar function that can be supposed constant and δ is the dirac distribution.

In Eq. I.19 only interband extremum contribute significantly to the integral over the BZ. Indeed, suppose given a pair of conduction and valence band (c,v), then the contribution of this pair on the absorption coefficient is proportional to $\frac{1}{\omega}$ and to the joint density of states, noted $J_{cv}(\hbar\omega)$ and defined as,

$$J_{cv}(\hbar\omega)=\int_{BZ}\delta(E_c(\mathbf{k})-E_v(\mathbf{k})-\hbar\omega)\mathrm{d}\mathbf{k}. \tag{I.20}$$

Using in Eq. I.20 the following δ distribution property,

$$\int_a^b g(x)\delta\left[f(x)\right]\mathrm{d}x=\sum_{x_0}g(x_0)\left|\frac{\mathrm{d}f}{\mathrm{d}x}\right|^{-1}_{x=x_0}, \tag{I.21}$$

in which x_0 is a zero of the function $f(x)$ contained in the interval (a,b), one gets,

$$J_{cv}(E)=\int_{E_c(\mathbf{k})-E_v(\mathbf{k})=E}\frac{\mathrm{d}S}{|\nabla_k[E_c(\mathbf{k})-E_v(\mathbf{k})]|} \tag{I.22}$$

where the integral $\mathrm{d}S$ is taken on the surface of the BZ defined by the equation

$$E_c(\mathbf{k})-E_v(\mathbf{k})=E. \tag{I.23}$$

The interband is defined as a function of $\mathbf{k}$, as the difference between the conduction and the valence band:

$$E_{cv}(\mathbf{k})=E_c(\mathbf{k})-E_v(\mathbf{k}). \tag{I.24}$$

Around the extrema of E_{cv}, the joint density of state shows strong variation since $\nabla_k[E_c(\mathbf{k}) - E_v(\mathbf{k})]$ is close to 0.

The previous reasoning, initially formulated in [30], allows to deduce the location of critical points inside the band structure. These critical points are typically located on high symmetry point of the BZ, for instance in Γ. On those critical points, where an interband extrema can be found, are situated the main direct contributions to the absorption coefficient. A direct bandgap is associated to each critical point.

To summarize, the band structure is a collection of band, namely scalar function on the reciprocal space, $\mathbb{R}^3$, reduced to the BZ. Two of these bands are of particular importance: the valence band, E_v, and the conduction band E_c. From the conduction band and valence band, one define another scalar function over the BZ: the interband (see Eq I.24). This scalar function has extrema located over the BZ. These extrema are called the critical points of the band structure. So from a band structure, one can compute the critical points of this band structure.

Then, from the definition of the joint density of states Eq. I.20, the extrema of the interband are the main contribution on the absorption coefficient, thanks to Eq. I.19. A direct bandgap is located on each of these critical points, and it can be computed as the value of the interband at this critical point.

However, no empirical or fundamental models are currently available for modelling the absorption coefficient for direct transitions. Such models aim to directly evaluate the integral over the BZ of Eq.I.19. For instance, recent calculations of Eq. I.19 by *ab-initio* methods have clearly emphasized the drastic impact of excitonic effects [32]. The figure Fig. I.4 illustrates the state of the art in ab-initio computation of the dielectric function and shows the semi-qualitative results obtained so far (see for instance [33, 21, 34]). Models and computations details are out of the scope of the present work.

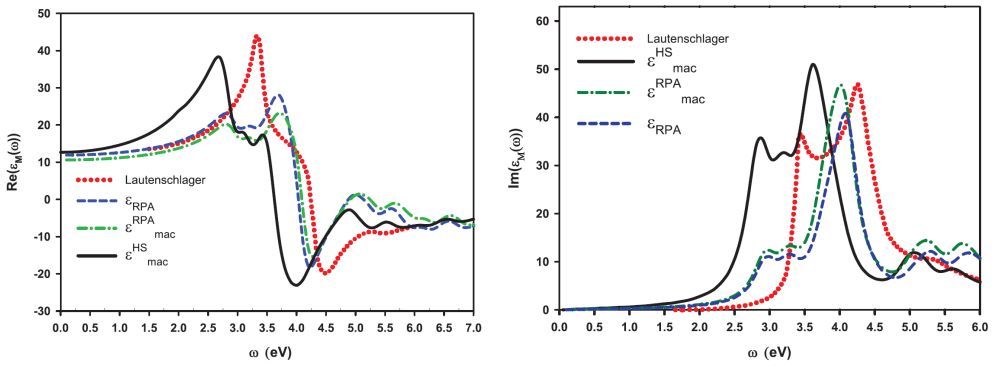


Figure I.4: Comparison between experimental data (dotted red) and various ab-initio models (black, green and blue line) whose definitions are available in [21]. Best results so far are only semi-qualitative. Lautenschlager data are from [35]. Figure extracted from [21].

To conclude, in contrast with the previously mentioned approach for indirect transitions, a clear empirical model for direct transitions is not available and only direct

bandgaps can accurately be computed from the band structure.

I.2.3 Ellipsometry

Even though direct theoretical computation of optical constants is currently not reachable, measurements methods dedicated to optical constants are available and known in literature as Ellipsometry. In this section, we firstly introduce Ellipsometry in general. Secondly, we focus on the dispersion models used in this domain. And finally we present how optical constants measured on a discrete set of conditions (for instance temperature) can usually be extrapolated to a continuous set of conditions.

The permittivity models used in ellipsometry are often called oscillators. An oscillator is a complex-valued function of the energy, wavelength or angular frequency respecting the Kramers-Kronig relations (see Eq. I.6). Also named pole, it refers to, for instance, the Drude model (see Eq. I.26), Tauc-Lorentz model (see Eq. I.30) or the DFPM (see Eq. I.33). In ellipsometry (see section I.2.3), a permittivity model is formulated as the sum of oscillators of different types. We refer to such a model as an oscillator scheme.

I.2.3.1 Introduction

According to [24], Ellipsometry is an optical measurement technique that involves generating a light beam in a known polarization state and reflecting it from a sample having a planar surface. By measuring the polarization state of the specularly reflected beam, the ellipsometry angles (ψ, Δ) can be determined. These angles are specific to the wavelength λ_0 of the light beam and the angle of incidence θ_i of the beam at the sample surface. Upon detailed analysis, the angles (ψ, Δ), along with the associated known values of λ_0 and θ_i, yield information on the sample. Such information for a bulk sample includes the optical properties, i.e. the index of refraction n and the extinction coefficient k, which depend on the wavelength λ_0. Information deduced for samples consisting of one or more thin films having plane-parallel surface/interfaces includes the layer thicknesses d and (n, k) of the components. Considering samples that are isotropic, which describe most structures of interest in photonics applications, (ψ, Δ) are defined by:

$$\tan(\psi)\exp(i\Delta) = \frac{r_p}{r_s}, \tag{I.25}$$

where r_p and r_s are the complex amplitude reflection coefficients for linear p and s-polarization states. For these states, the electric field vibrates parallel (p) and perpendicular (s) to the plane of incidence, defined by the incident and reflected beam propagation directions. Several variations of the ellipsometry experiment have been developed with the goals to obtain a large set of (ψ, Δ) pairs that facilitates data interpretation and to extract as much information as possible on the sample. In spectroscopic ellipsometry, (ψ, Δ) are measured continuously versus the wavelength of the light beam. In real time ellipsometry, (ψ, Δ) are measured versus time at fixed λ_0. In expanded beam imaging spectroscopic ellipsometry, (ψ, Δ) are measured along a line on the surface of the sample using an instrument with a two-dimensional detector array. In general, the most widely used ellipsometers for photonics applications are spectroscopic and span the range from the ultraviolet to the near-infrared (200–2000 nm). Spectroscopic ellipsometry is of great interest in photonics research and development due to its ability to extract $\{d, (n, k)\}$ information for the multiple layers materials and (n, k) for the bulk materials, e.g. wafers or substrates.

I.2.3.2 Dispersion models

In Spectroscopic Ellipsometry, dispersive models are extensively used in order to fit the measured (ψ, Δ) on the wavelength range of interest. In the following is presented a non-exhaustive list of them.

Drude

Drude and Sommerfeld in the late 1800's proposed a model that describes the interaction of time-varying electric fields with free carriers which move freely in conductive materials. The Drude model is given by:

$$\varepsilon_{Drude}(E) \;=\; \varepsilon_r(\infty) \;-\; \frac{A}{E^2 - i\Gamma E} \tag{I.26}$$

where $\varepsilon_r(\infty)$ is the high-frequency dielectric constant, A is the amplitude and Γ is the broadening.

Tauc-Lorentz

Jellison and Modine developed the Tauc-Lorentz (TL) model using the Tauc formula and a Lorentz oscillator [36, 37]. The complex dielectric function is:

$$\begin{aligned} \varepsilon(E) \;&=\; \varepsilon_{r,TL}(E) \;+\; i\,\varepsilon_{i,TL}(E), \\ &=\; \varepsilon_{r,TL}(E) \;+\; i\,(\varepsilon_{i,T}(E)\,\varepsilon_{i,L}(E))\,. \end{aligned} \tag{I.27}$$

Here the imaginary part of the TL, $\varepsilon_{i,TL}$, is given by the product of imaginary part of Tauc's dielectric, $\varepsilon_{i,T}$, defined as, for E in R^+:

$$\varepsilon_{i,T}(E) \;=\; \mathbb{1}_{\{E_g<E\}} A_T \left(\frac{E - E_g}{E}\right)^2, \tag{I.28}$$

where A_T is the amplitude, E_g the bandgap energy, and with Lorentz one, $\varepsilon_{i,L}$, defined as:

$$\varepsilon_{i,L} \;=\; \frac{A_L\; E_0\; C\; E}{\left(E^2 - E_0^2\right)^2 \;+\; C^2E^2}, \tag{I.29}$$

where A_L is the amplitude, C the broadening term and E_0 the central peak energy.

By multiplying Eq. I.28 and Eq. I.29, one gets the imaginary part of the TL model:

$$\varepsilon_{i,TL}(E) \;=\; \begin{cases} \dfrac{1}{E}\dfrac{A\; E_0\; C\,(E - E_g)^2}{\left(E^2 - E_0^2\right)^2 \;+\; C^2E^2} & \text{for} \quad E > E_g \\[2ex] 0 & \text{for} \quad E \leq E_g \end{cases} \tag{I.30}$$

The real part of the TL model can be obtained by applying the KK relations (I.6) to Eq. I.30. The exact analytical formula is available in [37].

Fits of Si and Ge permittivity using only TL oscillators can be found in literature. For instance, in [38], 11, resp. 16, TL oscillators have been used in order to fit Si, resp. Ge, permittivity on the energy range [0.5, 6] eV at room temperature.

Sellmeier

The Sellmeier dispersion function [39] models the real part of the permittivity. It is given as:

$$\varepsilon_{r,Sell}(\lambda) = 1 + \frac{A\lambda^2}{\lambda^2 - \lambda_0^2}, \tag{I.31}$$

or equivalently:

$$\varepsilon_{r,Sell}(E) = 1 + \frac{AE^2}{E^2 - E_0^2}, \tag{I.32}$$

where there are two free parameters for each Sellmeier oscillator: A the amplitude and a resonant wavelength λ_0 (or a resonant energy E_0).

DFPM

The Dielectric Function Parametric Model (DFPM) is also known as PSEMI (Parametric Semiconductors Model). This model was first introduced in [40] and it aims to provide a flexible, generic and parametric model for semiconductor materials. We find in literature examples of it being used: Emminger et. al. [25] fitted ellipsometry data of Ge permittivity at various temperatures with the DFPM. Johs and Herzinger [41] modelled Cadmium-Mercury $Hg_{1-x}Cd_x$ optical constants at various concentrations. Ihn and Kim [42] also modelled the dielectric function of $Cd_{1-x}Mg_xTe$ alloys. We focus on this model since these studies are also modelling optical constant according to condition parameters, such as temperature or material content.

In all previously mentioned studies, the DFPM is always simultaneously used with a Sellmeier model. Since this model is already presented (see Eq. I.32), it is not added in the main equation (Eq. I.33).

In this model the dielectric function is written as the summation of m energy-bounded, Gaussian-broadened polynomials W_j. The starting expressions is given by:

$$D_{FPM}(E) = 1 + i\sum_{j=1}^{m}\int_{E_{min,j}}^{E_{max,j}} W_j(E')\phi_j(E, E', \sigma_j)\mathrm{d}E' \tag{I.33}$$

The next paragraphs aim to describe the different features of Eq. I.33. For a quick presentation only go directly to Eq. I.46.

First, the function $\phi_j(E, E', \sigma_j)$ under the integral describes the broadening of the j^{th} polynomial W_j. This function is given by:

$$\phi_j(E, E', \sigma_j) = \int_0^{+\infty} \exp\{(is\,[E - E' + i\gamma_j(s)])\}\mathrm{d}s - \int_0^{+\infty} \exp\{[is\,(E + E' + i\gamma_j(s)])\}\mathrm{d}s \tag{I.34}$$

and the definition of γ_j is given by the broadening considered: $\gamma_j = 2\sigma_j^2 s$ for Gaussian broadening ($\sigma_j \in \mathbb{R}$), or $\gamma_j = \Gamma_j$ for Lorentzian broadening ($\Gamma_j \in \mathbb{R}$). In case of Gaussian broadening, one can transform to the variables:

$$\xi_{1,j} = \frac{E - E'}{2\sqrt{2}\sigma_j} \quad \text{and} \quad \xi_{2,j} = \frac{E + E'}{2\sqrt{2}\sigma_j}. \tag{I.35}$$

Transforming Eq. I.34 with Laplace transforms for the case of Gaussian broadening one arrives at the following expressions:

$$\begin{aligned}\phi_j(E, E', \sigma_j) &= \sqrt{\frac{\pi}{8\sigma_j^2}}\left[\exp\left(-\xi_{1,j}^2\right)\operatorname{erfc}\left(-i\xi_{1,j}\right) - \exp\left(-\xi_{2,j}^2\right)\operatorname{erfc}\left(-i\xi_{2,j}\right)\right], \\ &= \sqrt{\frac{\pi}{8\sigma_j^2}}\left[\Psi(\xi_{1,j}) - \Psi(\xi_{2,j})\right], \end{aligned} \tag{I.36}$$

where erfc designates the complementary error function given by $\operatorname{erfc}(z) = 1 - \operatorname{erf}(z)$, for all $z \in \mathbb{C}$, where

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp\left(-t^2\right)\mathrm{d}t \quad \text{and} \quad \Psi(z) = \exp\left(-z^2\right)\operatorname{erfc}\left(-iz\right). \tag{I.37}$$

In order to compute the erfc function, one can use a package freely available[1].

Returning to Eq. I.33, we now focus on the m energy-bounded polynomials W_j. In theory, the choice of m and $\{W_j \mid j \in [[1,\ m]]\}$ relies on the user. In practice, m is chosen by the user and for W_j, for instance in [40], we have the following definitions. Suppose that j is a fixed integer.

Given,

$$\begin{aligned} &A^j,\ E_C^j,\ E_L^j,\ E_U^j \quad \in \quad \mathbb{R}^{+*} \quad \text{such as} \quad E_L^j < E_C^j < E_U^j, \\ &F_U^j,\ F_L^j,\ A_{sym}^j,\ A_{UM}^j,\ A_{LM}^j,\ L_{2d}^j,\ U_{2d}^j,\ \sigma^j \quad \in \quad [0, 1], \end{aligned} \tag{I.38}$$

the interval $[E_L^j,\ E_U^j]$ is the support of the energy-bounded polynomial W_j, *ie* we have:

$$W_j(E) \ = \ 0 \quad \text{if} \quad E \notin [E_L^j,\ E_U^j]. \tag{I.39}$$

and by definition $E_{min,j} \ := E_L^j$ and $E_{max,j} \ := E_U^j$. This support is divided in four parts, *ie* introducing,

$$E_{UM}^j \ = \ E_U^j \ + \ (E_C^j - E_U^j)F_U^j, \tag{I.40}$$

[1]See `http://ab-initio.mit.edu/wiki/index.php/Faddeeva_Package`. This package, written in cpp, also provides wrappers for C, Octave, Matlab, Python, R, Scilab and Julia.

$$E^j_{LM} \;=\; E^j_L \;+\; (E^j_C - E^j_L)F^j_L,$$

we have the following partition:

$$[E^j_L,\; E^j_U] \;=\; [E^j_L,\; E^j_{LM}] \;\cup\; [E^j_{LM},\; E^j_C] \;\cup\; [E^j_C,\; E^j_{UM}] \;\cup\; [E^j_{UM},\; E^j_U]. \tag{I.41}$$

In order to respect notations of [40], we introduce the following variable changes: given E in $\mathbb{R}^{+*}$:

$$\begin{aligned}
y_{j,1} &:= \frac{1}{E^j_{LM} - E^j_L} E \;-\; \frac{E^j_L}{E^j_{LM} - E^j_L} \\
y_{j,2} &:= \frac{1}{E^j_C - E^j_{LM}} E \;-\; \frac{E^j_{LM}}{E^j_C - E^j_{LM}} \\
y_{j,3} &:= \frac{1}{E^j_{UM} - E^j_C} E \;-\; \frac{E^j_{UM}}{E^j_{UM} - E^j_C} \\
y_{j,4} &:= \frac{1}{E^j_U - E^j_{UM}} E \;-\; \frac{E^j_U}{E^j_U - E^j_{UM}}
\end{aligned} \tag{I.42}$$

and we introduce six intermediate variables:

$$\begin{aligned}
c^j_L &:= L^j_{2d} \frac{A^j_{LM}}{1 - A^j_{LM}} \left(\frac{E^j_C - E^j_{LM}}{E^j_{LM} - E^j_L} \right)^2, \\
d^j_L &:= \frac{1}{1 - A^j_{LM}} \left(\frac{E^j_C - E^j_{LM}}{E^j_{LM} - E^j_L} A^j_{LM} \left(E^j_C - E^j_L \right) \frac{L^j_{2d}}{E^j_{LM} - E^j_L} + \frac{1}{E^j_C - E^j_{LM}} \right), \\
c^j_U &:= U^j_{2d} \frac{A^j_{UM}}{1 - A^j_{UM}} \left(\frac{E^j_C - E^j_{UM}}{E^j_{UM} - E^j_U} \right)^2, \\
d^j_U &:= \frac{1}{1 - A^j_{UM}} \left(\frac{E^j_C - E^j_{UM}}{E^j_{UM} - E^j_U} A^j_{UM} \left(E^j_C - E^j_U \right) \frac{U^j_{2d}}{E^j_{UM} - E^j_U} + \frac{1}{E^j_C - E^j_{UM}} \right), \\
A^j_L &:= A^j(1 - A^j_{sym}), \\
A^j_U &:= A^j(1 + A^j_{sym}).
\end{aligned} \tag{I.43}$$

Finally, for E in $\mathbb{R}^{+*}$, W_j is given by the following equation:

$$\begin{aligned}
W_j(E) \;=\; & \mathbb{1}_{[E^j_L,\, E^j_{LM}]} P^j_1(y_{j,1}(E)) \;+\; \mathbb{1}_{[E^j_{LM},\, E^j_C]} P^j_2(y_{j,2}(E)) \;+ \\
& \mathbb{1}_{[E^j_C,\, E^j_{UM}]} P^j_3(y_{j,3}(E)) \;+\; \mathbb{1}_{[E^j_{UM},\, E^j_U]} P^j_4(y_{j,4}(E))
\end{aligned} \tag{I.44}$$

where P^j_1, P^j_2, P^j_3 and P^j_4 respectively are defined as function of $y_{j,1}$, $y_{j,2}$, $y_{j,3}$ and $y_{j,4}$ respectively, as:

$$\begin{aligned}
P^j_1(y_{j,1}) \;&= A^j_L \left(A^j_{LM}(1 - L^j_{2d}) y_{j,1} \;+\; A^j_{LM} L^j_{2d} y^2_{j,1} \right), \\
P^j_2(y_{j,2}) \;&= A^j_L \left(A^j_{LM} \;+\; \frac{1 - A^j_{LM}}{1 - c^j_L - d^j_L} y_{j,2} \;+ c^j_L {y_{j,2}}^2 \;+ d^j_L {y_{j,2}}^4 \right),
\end{aligned} \tag{I.45}$$

$$P_3^j(y_{j,3}) = A_U^j \left(A_{UM}^j + \frac{1 - A_{UM}^j}{1 - c_U^j - d_U^j} y_{j,2} + c_U^j {y_{j,2}}^2 + d_U^j {y_{j,2}}^4 \right),$$
$$P_4^j(y_{j,4}) = A_U^j \left(A_{UM}^j (1 - U_{2d}^j) y_{j,1} + A_{UM}^j U_{2d}^j y_{j,1}^2 \right).$$

To summarize, given six control points of $\mathbb{R}^2$ defined as:

$$\begin{aligned} C_1 &= (E_L^j,\ 0) \\ C_2 &= (E_{LM}^j,\ \widehat{A_L^j}) \\ C_3 &= (E_C^j,\ A_L^j) \\ C_4 &= (E_C^j,\ A_U^j) \\ C_5 &= (E_{UM}^j,\ \widehat{A_U^j}) \\ C_6 &= (E_U^j,\ 0) \end{aligned} \tag{I.46}$$

where $\widehat{A_U^j} = A_U^j A_{UM}^j$ and $\widehat{A_L^j} = A_L^j A_{LM}^j$, W_j is a real function that passes through all these six control points, is a polynomial of order 1, 2 or 4 on each subinterval of $[E_L^j, E_U^j]$ (defined by Eq. I.41), is continuous and differentiable in E_{UM}^j and E_{LM}^j, and is discontinous in E_C^j if A_{sym}^j is not equal to zero.

An example of one W_j function is available in Fig. I.5.

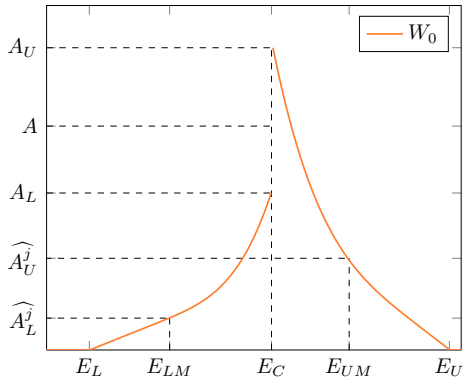


Figure I.5: Schematic construction of W_0 from the six control points coordinates. Parameters are available in table I.1.

In practice, a DFPM model is usually defined so the support of each oscillator overlaps. The final choice of oscillator supports definition *in fine* relies on the user. For instance, assuming that three oscillators are used ($m = 3$), the following relations could be chosen to define oscillator support:

$$E_C^0 < E_C^1 < E_C^2, \tag{I.47}$$
$$E_U^0 = E_C^1, \tag{I.48}$$

$$
\begin{aligned}
E_L^1 &= E_C^0, \\
E_U^1 &= E_C^2, \\
E_L^2 &= E_C^1.
\end{aligned}
$$

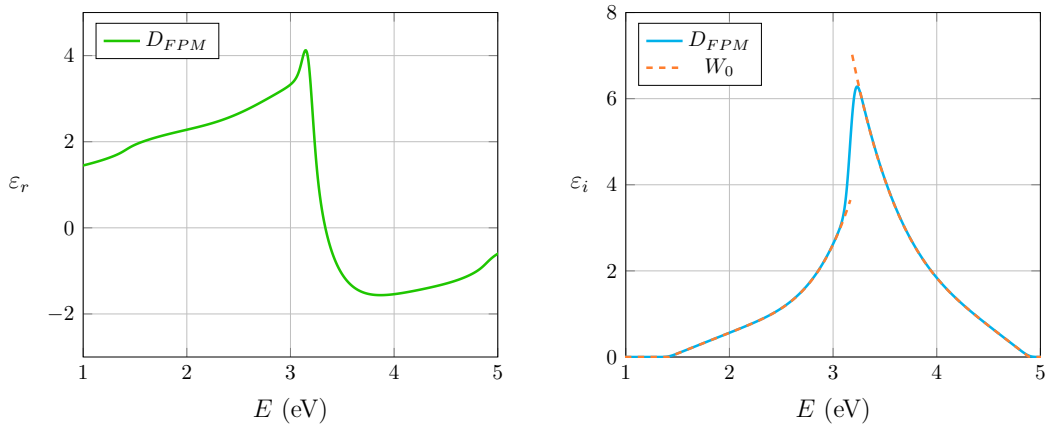
In order to actually compute this model, one remarks that two variable changes are needed. Firstly, W_j is expressed in Eq. I.44 as a variable of $\{y_i \mid i \in [1,4]\}$ and must be expressed as a function of E. Secondly, W_j, as a function of E, need to be expressed in term of $\xi_{1,j}$ and $\xi_{2,j}$ defined in Eq. I.35. The final step consists in evaluating the following integral,

$$
I_n^{(i,j)} = \int_{E_{min,j}}^{E_{max,j}} {\xi_{i,j}}^n \, \Psi(\xi_{i,j}) \tag{I.49}
$$

for i in $[[1,2]]$, n in $[[0,4]]$, j in $[[1,m]]$, where $\Psi(z)$ is defined by Eq. I.37 and where the n exponent on $\xi_{i,j}$ denotes the usual polynomial function. In order to compute $I_n^{(i,j)}$, 1D precomputed table of Ψ are used in literature (see [40, 25, 41]). We do not use such tables in our implementation and rather we divide the integration interval $[E_{min,j},\ E_{max,j}]$ in seven subintervals and apply a twelve dots Gauss-Legendre quadrature formula on each subinterval.

In the following, we illustrate the DFPM with simple examples of first, only one oscillator, and then, a summation of three oscillators.

In Fig. I.6, one can see an example of a single DFPM oscillator ($m = 1$). The input values are given in Table I.1. In Fig. I.6b, the orange dashed line is $W_0(E)$ and the blue line is the actual imaginary part of the oscillator. One remarks the smoothing operated by the convolution of W_0 by ϕ_0 in order to get the imaginary part. The corresponding real part is shown in Fig. I.6a.



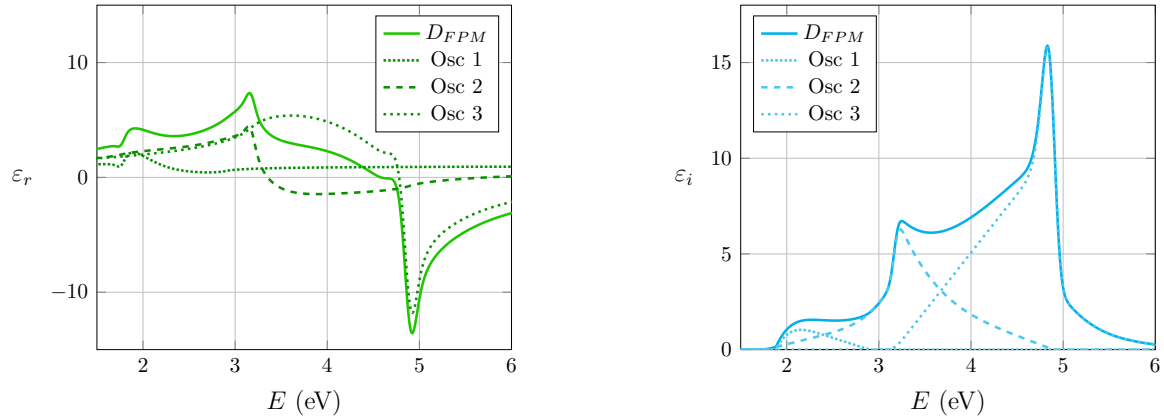
(a) Real part.

(b) Imaginary part.

Figure I.6: A DFPM with one oscillator. Parameters are available in Table I.1.

In Fig. I.7, one can see a DFPM model composed of three oscillators ($m = 3$). The resulting real, respectively imaginary, part of the model is the sum of each oscillator real,

respectively imaginary, part. The parameters are available in Table I.1. The definition of the three oscillators support follow the scheme introduced by Eq. I.47.



(a) Real part.

(b) Imaginary part.

Figure I.7: A DFPM with three oscillators. Parameters are available in table I.1.

Fig.	A	E_C	E_U	E_L	F_U	F_L	A_{sym}	A_{UM}	A_{LM}	L_{2d}	U_{2d}	σ
I.5, I.6	5.19	3.17	4.89	1.76	0.1	0.3	0.35	0.045	0.15	0	0	45.28
I.7, osc 1	0.93	1.76	2.89	1.76	0.4	0.5	0	0.75	0.5	0	0	40
I.7, osc 2	5.19	3.17	4.89	1.76	0.1	0.3	0.35	0.045	0.15	0	0	45.28
I.7, osc 3	12.22	4.89	6.5	3.17	0.1	0.8	-0.72	0.024	0.4	0	0	64.7
I.10a, $x = 0$	19.67	3.34	4.80	1.12	0.61	0.92	0.12	0.21	0.56	0	0	81.4
I.10a, $x = 1$	29.67	5.34	6.30	3.12	0.43	0.83	0.2	0.21	0.05	0	0	61.8
I.10b, $x = 0$	89.06	4.80	6.5	3.26	0.61	0.89	0	0.22	0.27	0	0	124
I.10b, $x = 1$	52.02	6.30	7.5	3.26	0.61	0.75	0	0.07	0.27	0	0	124

Table I.1: Parameters of DFPM models shown in various figures.

I.2.3.3 Critical point model

Both E_g for TL model and E_C from the DFPM represent a bandgap of the considered material and thus accept a physical interpretation. In order to extract these specific parameters from ellipsometry measurements, a model, the critical point parabolic band (CPPB), has been proposed in [43]. It allows, as a preliminary step, to retrieve bandgaps by fitting the second derivative of the measured real and imaginary part of the permittivity.

A example can be found for instance in [25] and is shown in Fig. I.8. The authors first extracted critical points (E_C) with the CPPB model, and then fitted a DFPM with height oscillators on measurements using all previously extracted E_C.

The CPPB function models the second derivative of the dielectric function near critical points using five parameters: amplitude A, phase projection factor θ, threshold energy E_g, broadening parameter Γ and exponent μ:

$$\begin{aligned} \varepsilon(\omega) &= B - \frac{Ae^{i\theta}}{(\hbar\omega - E_g + i\Gamma)^{\mu}} \quad \text{if } \mu \in \{-\frac{1}{2}, \frac{1}{2}\} \\ \varepsilon(\omega) &= B - \frac{Ae^{i\theta}}{(\hbar\omega - E_g + i\Gamma)^{\mu}} \quad \text{if } \mu = 0 \end{aligned} \tag{I.50}$$

where μ has three discrete values: $\mu \in \{-\frac{1}{2}, 0, \frac{1}{2}\}$.

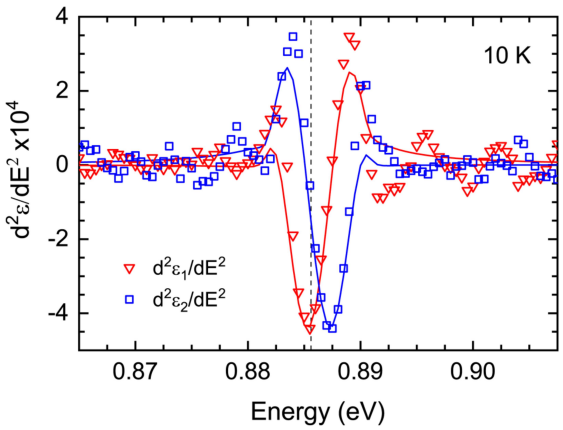


Figure I.8: Example of critical point identification from the numerically calculated second derivatives of the real part (triangles) and the imaginary part (squares) of the measured permittivity of Ge at 10 K. The solid lines represent the best fit of the second derivative of Eq. I.50 to the data, and the vertical black dashed line indicates the extracted E_0 threshold energy. Figure extracted from [25].

I.2.3.4 Parametric ellipsometry

Parametric ellipsometry aims to provide models of optical constants varying on one or multiple extra parameters, usually related to material concentration or temperature. In our problem, we wish to provide a unified model for the permittivity of Si, Ge and their alloys at different temperatures. So two extra parameters are considered: the Ge content (x in $[0, 1]$), and the temperature, noted T.

Measurements are done on particular conditions and can be repeated on a wide range in order to cover the set of interest. However this set of measured conditions is always *discrete* and an interpolation methodology is needed in order to extend available data on a *continuous* set of conditions.

The first methodology is simply to perform a linear interpolation of the permittivities at fixed energy. For instance, suppose given, on energy range $I = [1, 5]$ eV, both Si permittivity (noted $\varepsilon_{Si}(E)$) and Ge permittivity (noted $\varepsilon_{Ge}(E)$), then the dielectric function of a $SiGe$ alloy of concentration $x_0 \in]0, 1[$ is given by, for all E in I:

$$\varepsilon_{Si_{1-x_0}Ge_{x_0}}(E) \;=\; (1 - x_0)\varepsilon_{Si}(E) \;+\; x_0\varepsilon_{Ge}(E). \tag{I.51}$$

This method is extensively used in application when no data are available, for instance in [44, 45, 46]. However, it produces results not even close to actual measurements, as shown in Fig. I.9.

The second methodology aims at both more flexibility and more precision and consists in linearly interpolating oscillator parameters.

For instance, suppose given one TL oscillator for Si, noted $O_{1,Si}$, depending on parameters:

$$A_{Si},\; C_{Si},\; E_{(g,\,Si)},\; E_{(0,\,Si)}, \tag{I.52}$$

and one TL for Ge, noted $O_{1,Ge}$, depending on parameters:

$$A_{Ge},\; C_{Ge},\; E_{(g,\,Ge)},\; E_{(0,\,Ge)}. \tag{I.53}$$

The corresponding oscillator $O_{1,Si_{1-x}Ge_x}$ is then depending on parameters:

$$A_{Si_{1-x}Ge_x},\; C_{Si_{1-x}Ge_x},\; E_{(g,\,Si_{1-x}Ge_x)},\; E_{(0,\,Si_{1-x}Ge_x)}, \tag{I.54}$$

which are obtained by linear interpolation, *i.e* :

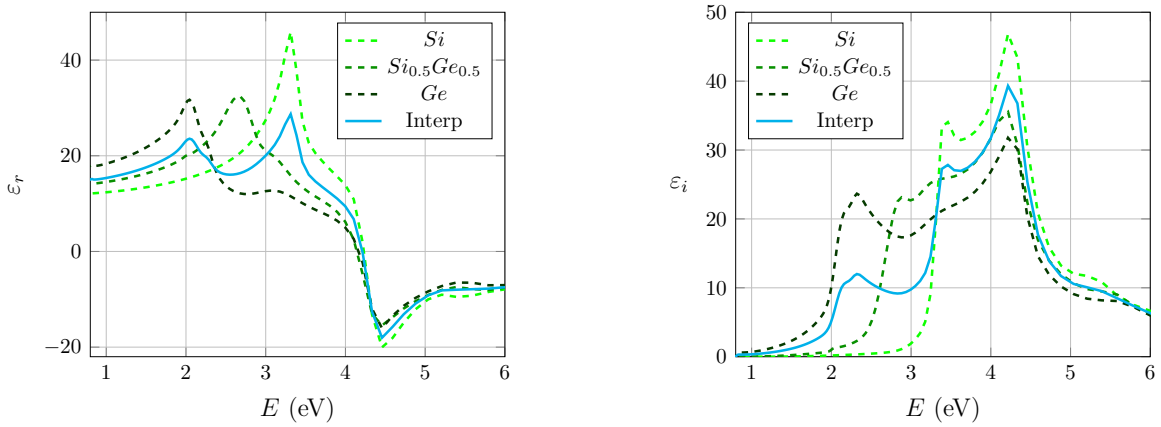
$$\begin{aligned}
A_{Si_{1-x}Ge_x} &= (1 - x)A_{Si} \;+\; xA_{Ge} \\
C_{Si_{1-x}Ge_x} &= (1 - x)C_{Si} \;+\; xC_{Ge} \\
E_{(0,Si_{1-x}Ge_x)} &= (1 - x)E_{(0,Si)} \;+\; xE_{(0,Ge)} \\
E_{(g,Si_{1-x}Ge_x)} &= (1 - x)E_{(g,Si)} \;+\; xE_{(g,Ge)}.
\end{aligned} \tag{I.55}$$

This methodology assumes that ε_{Si} and ε_{Ge} are modelled with the same number of oscillators, identified by pairs. Such oscillator identification is called an **oscillator scheme**.

Linear parameters interpolation is then performed on each oscillator defined by the scheme.

This methodology makes sense only when the oscillator parameters are in direct correspondence with geometrical characteristics of the oscillator curve. For instance, the DFPM model has been created in order to apply this methodology because all its parameters define six geometrical control points, as illustrated in Fig. I.5. Actual interpolation of DFPM oscillators is shown in Fig. I.10.

However, the TL oscillator parameters are more ambiguous: only E_g is a geometrical parameter for determining the support of the oscillator. The geometrical amplitude is not only function of A when $C \ll 1$, leading to erratic interpolation as shown in Fig. I.11. On the contrary, when $C \gg 1$, the interpolation of parameters leads to realistic interpolation, as shown in Fig. I.12.



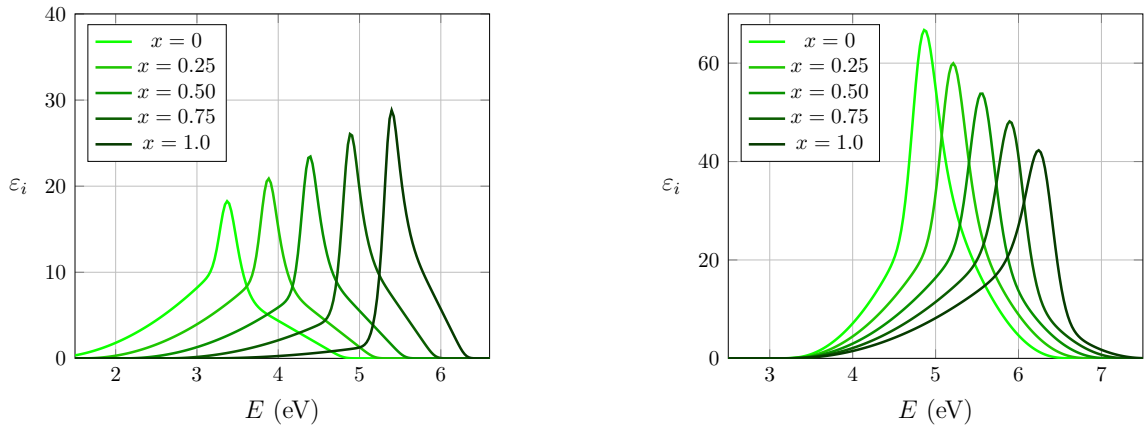
(a) Real part.

(b) Imaginary part.

Figure I.9: Performing a linear interpolation on the permittivity of Si and Ge leads to drastic errors. The differences between the prediction by linear interpolation (blue curve) and the actual measurement (medium green dashed curve) is clear. Dashed lines are measures from [47]. The blue line is obtained by linearly interpolating the Si and Ge permittivity (Eq. I.51 with $x_0 = 0.5$).

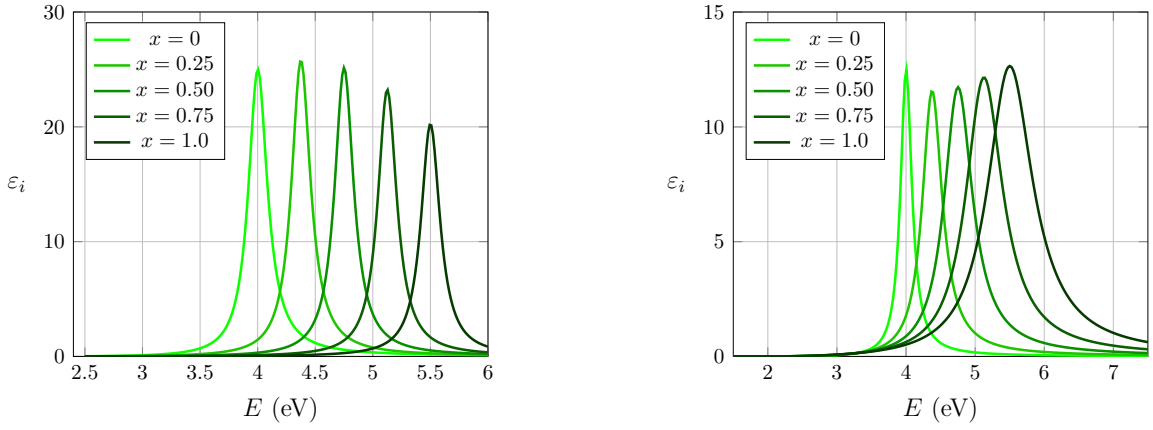
Fig.	x	A	C	E_i	E_g
I.12a	$x = 0$	30	10	5.5	1.5
I.12a	$x = 1$	30	10	5	2
I.12b	$x = 0$	30	10	5.5	1.5
I.12b	$x = 1$	30	10	5	2
I.11a	$x = 0$	20	0.2	4.5	2
I.11a	$x = 1$	10	0.2	5.5	2
I.11b	$x = 0$	10	0.2	4.5	2
I.11b	$x = 1$	25	0.8	5	2

Table I.2: TL parameters of various figures.



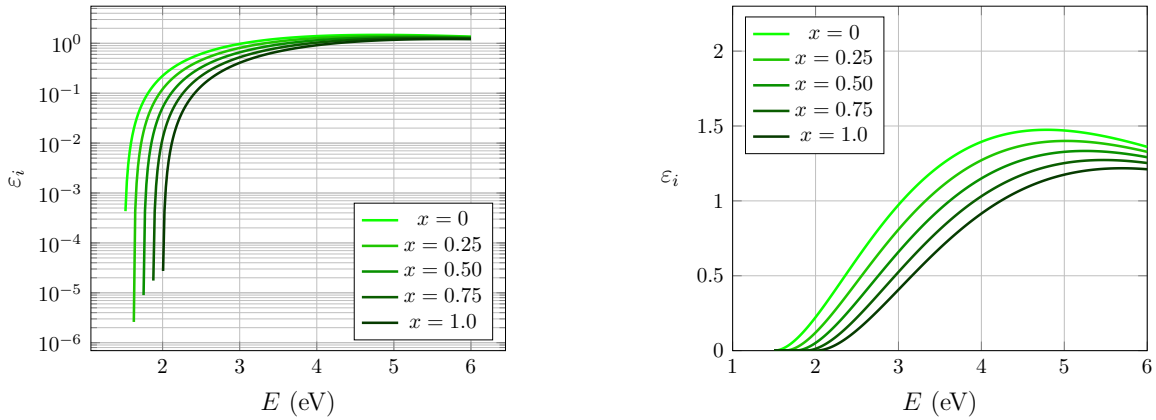
(a) (b)

Figure I.10: Linear parameters interpolations between two DFPM oscillators with parameters given by table I.1. The intermediate oscillators ($x = 0.25,\ 0.50,\ 0.75$) are exactly smooth translations between the curve $x = 0$ and $x = 1$. DFPM oscillator parameters of curves $x = 0.25,\ 0.50,\ 0.75$ are obtained similarly to Eq. I.55.



(a) (b)

Figure I.11: Linear parameters interpolations between two TL oscillators with parameters given by table I.2. Since the two initial TL oscillators ($x = 0$ and $x = 1$) have both $C \ll 1$, the interpolated oscillators do not behave coherently. For instance, the peak of curves $x = 0.25,\ 0.50,\ 0.75$ is lower or upper than the peak of curve $x = 0$ and $x = 1$. TL parameters of curves $x = 0.25,\ 0.50,\ 0.75$ are obtained with Eq. I.55.



(a) Log scale.

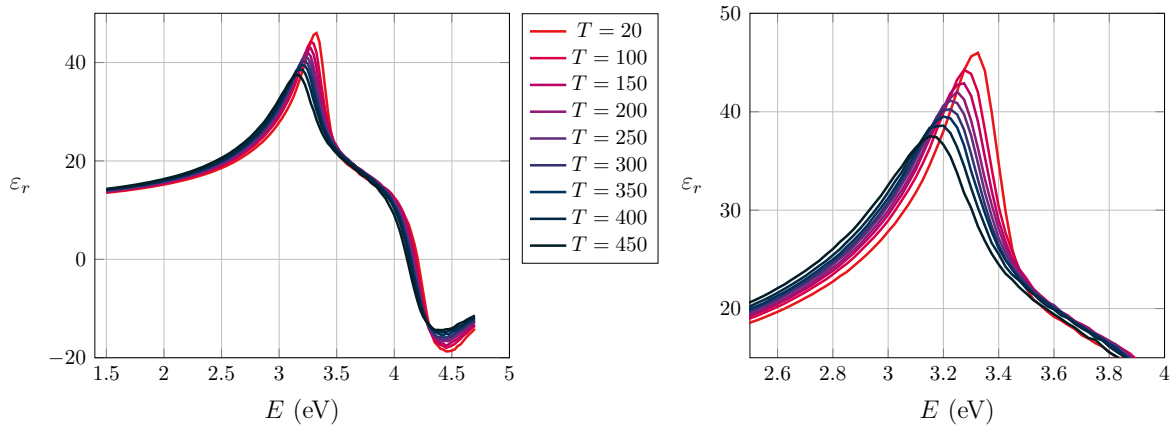
(b) Linear scale.

Figure I.12: Linear parameters interpolations between two TL oscillators with parameters given by table I.2. Since the two initial TL oscillators ($x = 0$ and $x = 1$) have both $C \gg 1$, the interpolated oscillators behave coherently, *i.e.* the curves $x = 0.25,\ 0.50,\ 0.75$ are smooth transition between the curve $x = 0$ and $x = 1$. TL parameters of curves $x = 0.25,\ 0.50,\ 0.75$ are obtained with Eq. I.55.

I.2.4 Available data

Here are presented data that will be used to fit our model. This list is not exhaustive and this work does not include an experimental data analysis. Three sets of data are chosen. The permittivity of pure Si according to various temperature is taken from [48] and shown in Fig. I.13 and I.14. The permittivity of Si, Ge and their alloys at 300 K are taken from [47], and shown in Fig. I.15 and I.16. The absorption coefficient of Si, Ge, and their alloys, at various temperatures, is taken from [31] and shown in Fig. I.17.

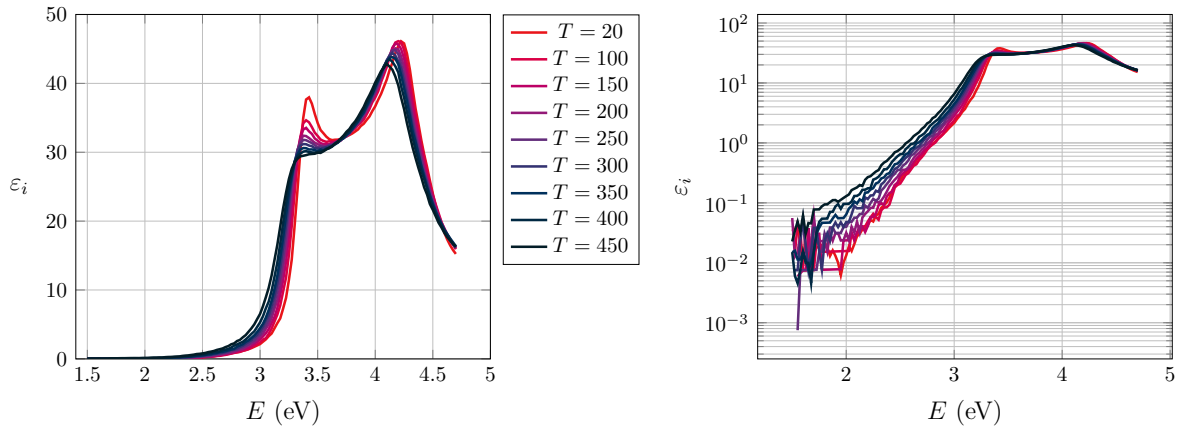
To our knowledge, data for the pure Ge permittivity according to temperature are not found in the literature. Thus, the main goal of our model is to fit data of Fig. I.15, I.16 and I.17.



(a) Linear scale.

(b) Linear scale, zoom.

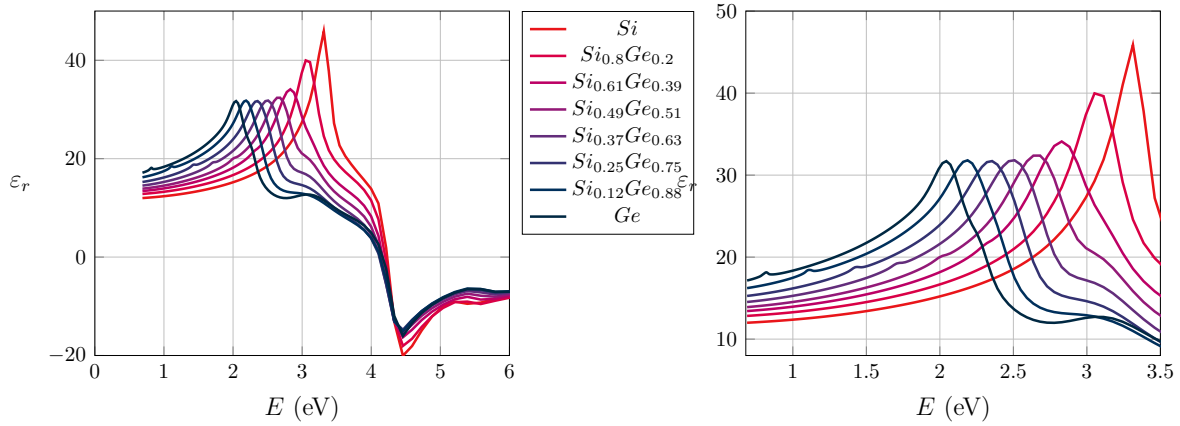
Figure I.13: Experimental measurements according to temperature of the real part of pure Si permittivity, from [48].



(a) Linear scale.

(b) Log scale.

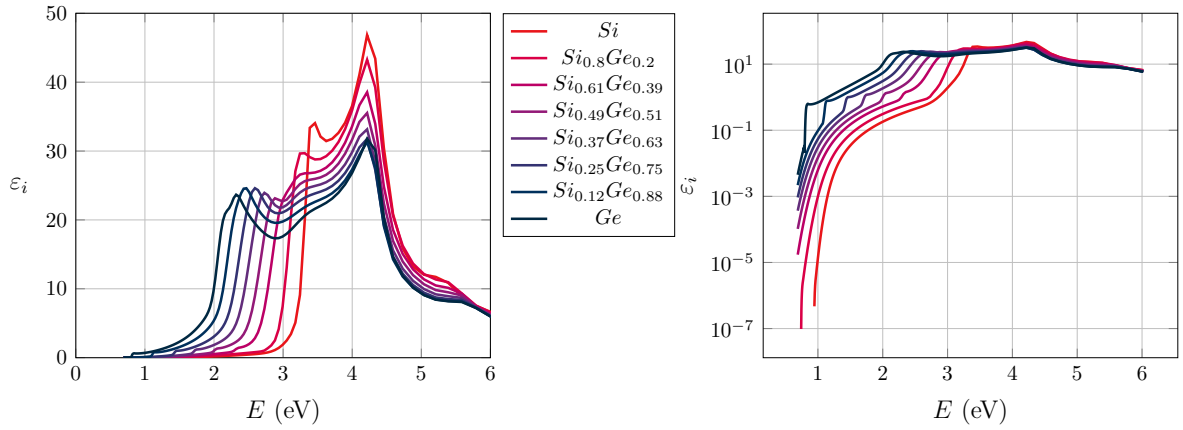
Figure I.14: Experimental measurements according to temperature of the imaginary part of pure Si permittivity, from [48]. In I.14b, the uncertainties in low energy are clearly visible. These data cannot be used for the NIR range (1-2.5 eV).



(a) Linear scale.

(b) Linear scale (zoom).

Figure I.15: Experimental real part of SiGe alloys permittivity at room temperature from [47].



(a) linear scale.

(b) Log scale.

Figure I.16: Experimental imaginary part of SiGe alloys permittivity at room temperature from [47].

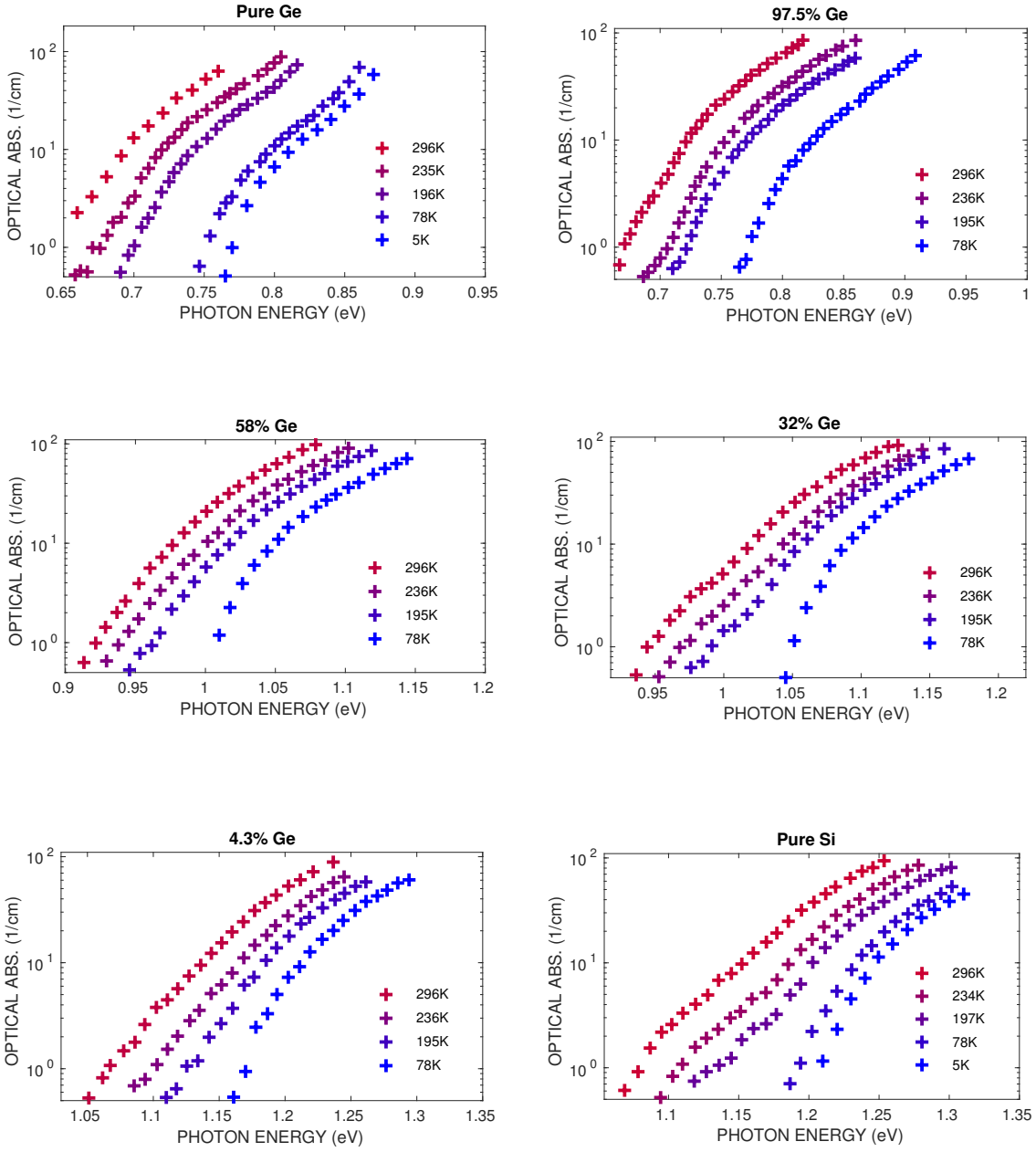


Figure I.17: Braunstein *et al.* [31] experimental data on the absorption coefficient, at various temperatures and Ge contents.

I.3 Model

The proposed semi-empirical model for the permittivity of SiGe accounting for Ge content and temperature is defined in this section. First, we describe the basic principles. Secondly, the modelling of both the indirect and direct transitions is clarified, followed by a detailed explanation of conditions variation.

I.3.1 Principle: parametric ellipsometry

We aim at providing an unified semi-empirical model for the permittivity of Silicon, Germanium and their alloys at various temperatures. The model must accept three input variables, the Ge content, noted x, the temperature, noted T, and the Energy, noted E. For the following, we choose energy as the variable instead of the angular frequency, noted ω, and the wavelength, noted λ. Corresponding variable changes are available in Eq. I.1. The model, providing permittivity value, is complex-valued.

Recalling that this is equivalent to a double parametric ellipsometry problem, both in x and T, we choose the approach described in the section I.2.3.4. The following points summarize the corresponding framework:

1. Choose a number of oscillators *i.e.* define an oscillator scheme.

2. Fit both pure Si and pure Ge at minimum and maximum temperature with this scheme.

3. Perform a double parameter interpolation (see Eq. I.55) in order to retrieve the $Si_{1-x}Ge_x$ permittivity at a given temperature, T, and a given Ge content, x.

As a preliminary step, we will perform a simpler task by fixing the temperature at $T_0 = 20$ °C. The framework becomes:

1. Choose a number of oscillators *i.e.* define an oscillator scheme.

2. Fit both pure Si and pure Ge, at fixed temperature, T_0, with this scheme.

3. Perform a single parameter interpolation (see Eq. I.55) in order to retrieve the $Si_{1-x}Ge_x$ permittivity at temperature T_0, for a given Ge content, x.

The remaining question lies in the oscillator scheme definition, which is explained in the following section.

I.3.2 Principle: oscillators scheme

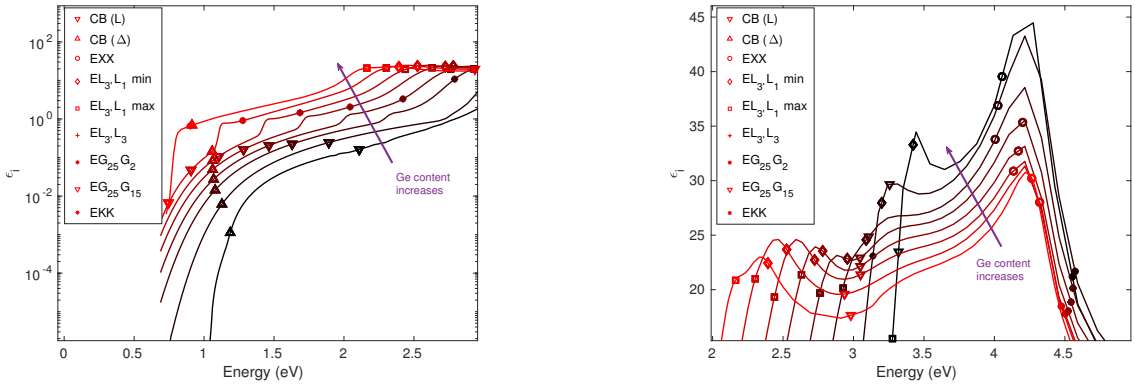
In this section is explained how the oscillator scheme is chosen and formulated.

As a reminder, an oscillator is a complex-valued function of the energy, wavelength or angular frequency respecting the Kramers-Kronig relations (see Eq. I.6). Also named pole, it refers to, for instance, the Drude model (see Eq. I.26), Tauc-Lorentz model (see Eq. I.30) or the DFPM (see Eq. I.33). In ellipsometry (see section I.2.3), a permittivity model is formulated as the sum of oscillators of different types. We refer to such a model as an oscillator scheme.

In order to define an oscillator scheme, we **decide** to identify each oscillator to a specific bandgap of the band structure. Each oscillator will thus represent transitions associated to a specific bandgap. Such a hypothesis is justified by Fig. I.18 where the bandgap variation on Ge content is displayed on the imaginary part of the permittivity of SiGe alloys. The bandgaps clearly follow the trend, peaks and gaps, of the permittivity. The bandgaps variation according to Ge content are computed with the TB model from [26] and their values are shown in Fig. I.19. From the band structure computation of each Ge content, the bandgaps are then extracted at critical points, as shown, for instance for pure Ge, in Fig. I.20.

This approach is coherent with the contribution of Bassani and Pastori that already claimed that only interband extrema contribute significantly to the optical constants for direct transitions (see section I.2.2.5).

Such methodology, *i.e.* to deduce, by TB, bandgaps at critical points and then use these computed bandgaps as physical parameters for oscillators, is strictly in competition with the induction of bandgaps done on ellipsometry measurements (see section I.2.3.4). However, computing the band structure to deduce bandgaps is ultimately more precise: In the 3-3.2 eV energy range, the CPPB (see Eq. I.50), could only identify one bandgap when the TB solver allows the computation of four bandgaps, as shown in Fig. I.21. This implies that optical transitions are more precisely described by the band structure computation from TB.



(a) Low Energy, log scale

(b) High Energy, linear scale

Figure I.18: Bandgaps at critical points on experimental permittivity imaginary part data from [47]. Pure Si permittivity and bandgaps are in black, pure Ge permittivity and bandgaps in red. Bandgaps variation on Ge content follow the trend, peaks and gaps, of the permittivity. For instance, the EXX bandgap clearly marks the 4.2 peak of the permittivity and the EKK bangap emphasizes the varying low energy peak. Three bandgaps, EL_3L_1 min/max and $EG_{25}G_2$, are contributing to the various peaks on the energy range 2-3.5 eV. All markers abscissa are determined by the bandgap value. Then, all markers are directly placed on the corresponding permittivity curve. The bandgaps are computed with the TB model from [26] and their values are shown in Fig. I.19.

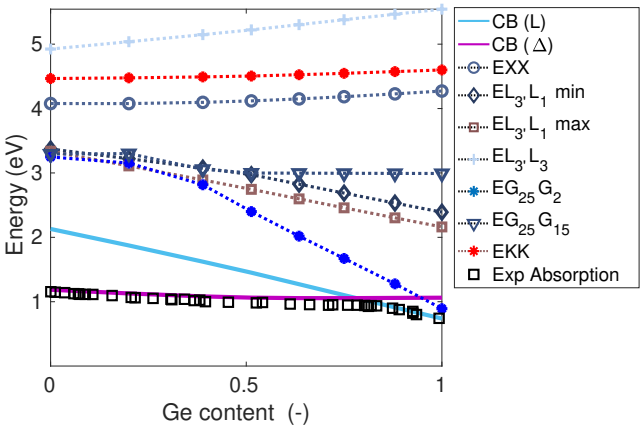


Figure I.19: TB predictions of the main direct (dashed lines) and indirect (solid lines) bandgaps at critical points of Si, Ge, and their alloy. Also shown (black square) are the experimental data at 4.3K from [49]. Except for the EL_3L_1min and EL_3L_1max bandgap, the spin-orbits splitting are averaged.

I.3.3 Indirect transitions modelling

Indirect transitions refer to indirect bandgaps, namely the CB(L) and CB(Δ) bandgaps (see Fig. I.19). From literature (see section I.2.2.4), an empirical model for the absorption coefficient, see Eq. I.18, is available. This model exhibits a $(\hbar\omega - E_g \pm \hbar\omega_q)^2$ dependency that is identical to the TL model (Eq. I.30) for the imaginary part of the permittivity

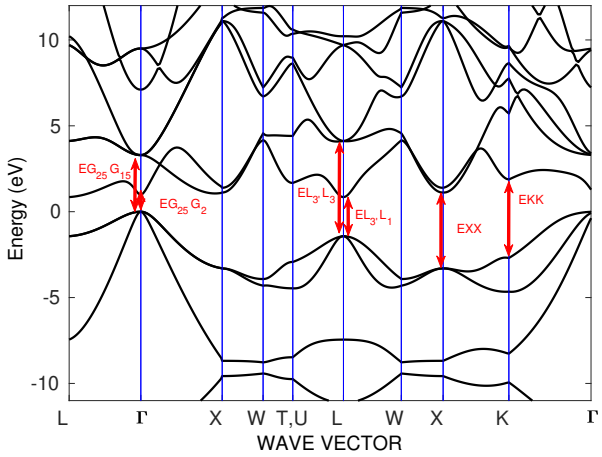


Figure I.20: Band structure of Ge computed with TB model. Main optical transitions (bandgaps) are shown by vertical arrows. For explanation about how to read this figure, please refer to section I.2.2.2.

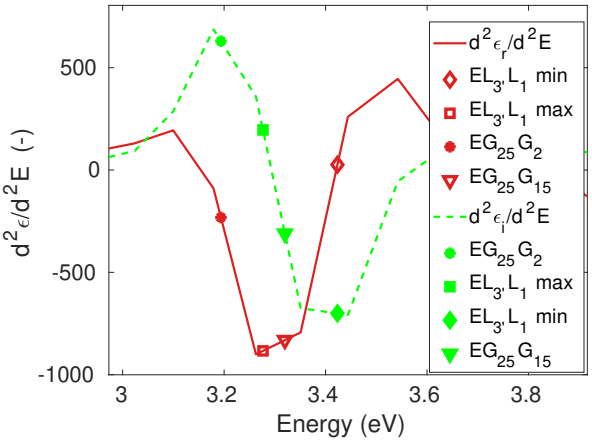


Figure I.21: Second order derivative of the real (red curve) and imaginary part (green curve) of the permittivity measured in Si from [47] and position of gaps at critical points of Si in the 3-4 eV range calculated with TB.

(see optical constant relations in section I.2.1.1).

A TL oscillator is identified to a bandgap through its E_g parameter.

Thus, we select the TL oscillators to model the indirect transitions. Recalling that every indirect bandgap is decomposed through emitting and absorption phonons (see Eq. I.14), we shall fit four TL oscillators for indirect transition modelling.

I.3.4 Direct transitions modelling

Since no empirical model can be found in the existing literature for direct transitions (see section I.2.2.5), we choose a more pragmatic approach combining the flexible DFPM model (Eq. I.33) with the gaps values at critical points computed by TB. The E_C parameters of all DFPM are set to the bandgaps values computed by TB, thus identifying

a DFPM oscillator to a bandgap.

Seven direct bandgaps are identified (see Fig. I.19 and Fig. I.20), leading to seven DFPM oscillators.

To summarize, in order to model the permittivity, TL and DFPM oscillators are used. TL (resp. DFPM) represents the indirect (resp. direct) transitions. Each oscillator is associated (through E_g for TL oscillators, and E_C for DFPM) to a bandgap, computed thanks to TB prediction. The oscillators decomposition of our model, fitted on the pure Ge permittivity, is visible in Fig. I.22 and Fig. I.23. The fitted model, shown as a red line, is the sum of all the oscillators, shown as dashed colored line. The fit at low energy, corresponding to the indirect transitions, is shown in Fig. I.22. The fit at high energy, corresponding to the direct transitions, is shown in Fig. I.23.

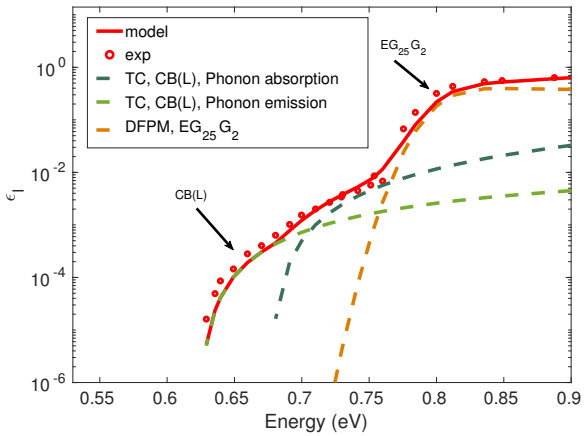


Figure I.22: Low energy oscillators decomposition for the imaginary part of permittivity in our model for pure Ge. Each dashed line corresponds to an oscillator: the pale and dark green ones are TL oscillators associated with the indirect CB(L) bandgap, the orange one is a DFPM oscillator associated with the $EG_{25}G_2$ direct bandgap. The solid curve (our model) is obtained by summing up each oscillator.

I.3.5 Real part modeling

As advised in the presentation of the DFPM (see section I.2.3.2), one Sellmeier oscillator (see Eq. I.32) is added to the sum of the real part of all previously mentioned TL and DFPM oscillators.

I.3.6 Temperature variations

As explained in section I.3.1, we choose the parametric ellipsometry methodology. So all parameters, excepted the bandgaps (E_g for TL, E_C for DFPM), of all TL and DFPM oscillators are linearly interpolated.

The bandgaps variation on Ge content are described by Fig. I.19. However TB models cannot include temperature and provide bandgaps at 0K only (see section I.2.2.2). The remaining temperature variations of bandgaps are described in the following paragraphs.

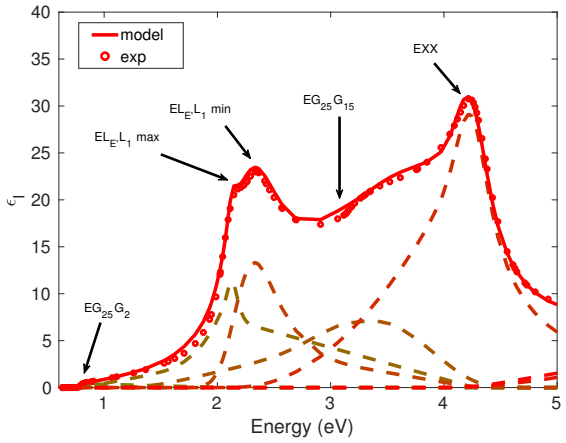


Figure I.23: High energy oscillators decomposition for the imaginary part of permittivity in our model for pure Ge. Each dashed line corresponds to a DFPM oscillator defined in Eq. I.33. The peak point of each DFPM oscillator curve is positioned at the bandgap of the corresponding critical point. The solid curve (our model) is obtained by summing up each oscillator plus the four TL (not visible on this scale).

The temperature dependency is included in this model through the band gap $E_g(T)$ temperature dependency, which is well described by the following formula:

$$E_g(T) = E_g(0) - \alpha \frac{T^2}{\beta + T} \tag{I.56}$$

where α and β are fitting parameters and are given in Table I.3. This formula is widely used in literature and was firstly proposed by [50].

	$E_g(0)$	α	β
Si	1.147	5.8e-4	636
Ge	0.71	3.4e-4	235

Table I.3: Fitting parameter for bandgap temperature dependence of Eq. I.56

We used Macfarlane [31] and Braunstein [51] experimental data to fit α and β for pure Si and pure Ge.

I.3.7 Parameters

In this section are available all parameters of the model at temperature of 300K. For DFPM oscillator, the "Left" and "Right" parameters determine the support of the oscillator, *i.e.* E_{min} and E_{max}. "Left" and "Right" are integer. "Left" determines E_{min}: if "Left" is equal to 1, then E_{min} is equal to E_C of the oscillator number 1. Similarly, "Right" designates E_{max}: if "Right" is equal to 7, then E_{max} is equal to E_C of the oscillator number 7. For instance, for pure Si (see table I.5), the support of the oscillator number 3 is the interval $[E_{C,\,osc\,0},\ E_{C,\,osc\,5}] = [1.25,\ 4.21]$. In all DFPM parameters tables, the oscillator number 0 is a fictional one used to define the support of all other oscillators. For a full definition of the DFPM model, see section I.2.3.2.

Nb osc	Bandgap	ε_∞	A	C	E_0	E_g
0	$CB(L)$ abs	0.00	3.00	5.00	3.00	2.173
1	$CB(L)$ emit	0.00	8.00	5.00	3.00	2.226
2	$CB(\Delta)$ abs	0.00	0.40	20.00	4.00	1.049
3	$CB(\Delta)$ emit	0.00	6.00	2.99	4.00	1.151

Nb osc	Bandgap	Left	Right	A	σ	E_C
0	-	0.00	0.00	1.00	1.00	1.25
1	$EG_{25}G_2$	1.00	2.00	0.55	50.00	3.26
2	$EL_3L_1\ max$	1.00	5.00	20.67	81.47	3.35
3	$EL_3L_1\ min$	0.00	5.00	12.71	42.25	3.35
4	$EG_{25}G_{15}$	4.00	5.00	11.84	256.58	3.26
5	EXX	4.00	7.00	80.06	124.70	4.21
6	EKK	5.00	7.00	5.03	136.77	5.39
7	EL_3L_3	5.00	7.00	5.68	200.00	6.50

Table I.5: Pure Si fitting DFPM parameters (Part 1) for direct transitions modeling. E_C are computed thanks to TB prediction.

Nb osc	A_{sym}	F_L	A_{LM}	L_{2d}	F_U	A_{UM}	U_{2d}
0	0.29	0.62	0.13	0.00	0.40	0.75	0.00
1	0.00	0.50	0.50	0.00	0.50	0.15	0.00
2	0.12	0.78	0.02	1.00	0.61	0.22	0.00
3	0.14	0.62	0.00	0.00	0.35	0.38	0.00
4	-0.35	0.40	0.45	0.00	0.15	0.28	0.00
5	-0.53	0.90	0.28	1.00	0.62	0.17	1.00
6	0.00	0.05	0.01	0.00	0.95	0.50	0.00
7	0.00	0.50	0.50	0.00	0.50	0.50	0.00

Table I.6: Pure Si fitting DFPM parameters (Part 2) for direct transitions modeling. E_C are computed thanks to TB prediction.

A	E_0
1.03	0.00

Table I.7: Pure Si fitting parameters of a single Sellmeier oscillator (see Eq. I.56).

Nb osc	Bandgap	ε_∞	A	C	E_0	E_g
0	$CB(L)$ abs	-0.12	1.00	46.45	2.00	0.623
1	$CB(L)$ emit	0.00	1.70	3.15	2.00	0.676
2	$CB(\Delta)$ abs	0.00	9.00	20.00	2.00	0.913
3	$CB(\Delta)$ emit	0.00	4.20	3.15	2.00	0.957

Table I.8: Pure Ge fitting TL parameters for indirect transitions modeling. E_g are computed thanks to TB prediction and Eq. I.56 (abs and emit refer to absorbing and emitting phonons).

Nb osc	Bandgap	Left	Right	A	σ	E_C
0	-	0.00	0.00	0.33	20.00	1.40
1	$EG_{25}G_2$	1.00	2.00	0.82	20.00	0.80
2	$EL_3L_1\ max$	1.00	5.00	9.92	30.58	2.11
3	$EL_3L_1\ min$	0.00	5.00	8.20	200.00	3.18
4	$EG_{25}G_{15}$	4.00	5.00	22.35	150.00	2.18
5	EXX	4.00	7.00	34.85	124.70	4.27
6	EKK	5.00	7.00	3.45	136.77	5.67
7	EL_3L_3	5.00	7.00	4.68	200.00	6.50

Table I.9: Pure Ge fitting DFPM parameters (Part 1) for direct transitions modeling. E_C are computed thanks to TB prediction.

Nb osc	A_{sym}	F_L	A_{LM}	L_{2d}	F_U	A_{UM}	U_{2d}
0	0.29	0.62	0.13	0.00	0.40	0.75	0.00
1	-0.50	0.00	0.00	0.00	0.50	0.50	0.00
2	0.00	0.30	0.09	0.00	0.89	0.65	0.00
3	0.00	0.30	0.18	0.00	0.40	0.55	0.00
4	0.00	0.52	0.07	0.00	0.40	0.08	0.00
5	-0.35	0.90	0.45	1.00	0.62	0.21	1.00
6	0.00	0.05	0.03	0.00	0.95	0.50	0.00
7	0.00	0.50	0.50	0.00	0.50	0.50	0.00

Table I.10: Pure Ge fitting DFPM parameters (Part 2) for direct transitions modeling. E_C are computed thanks to TB prediction.

A	E
1.03	0.00

Table I.11: Pure Ge fitting parameters of a single Sellmeier oscillator (see Eq. I.56).

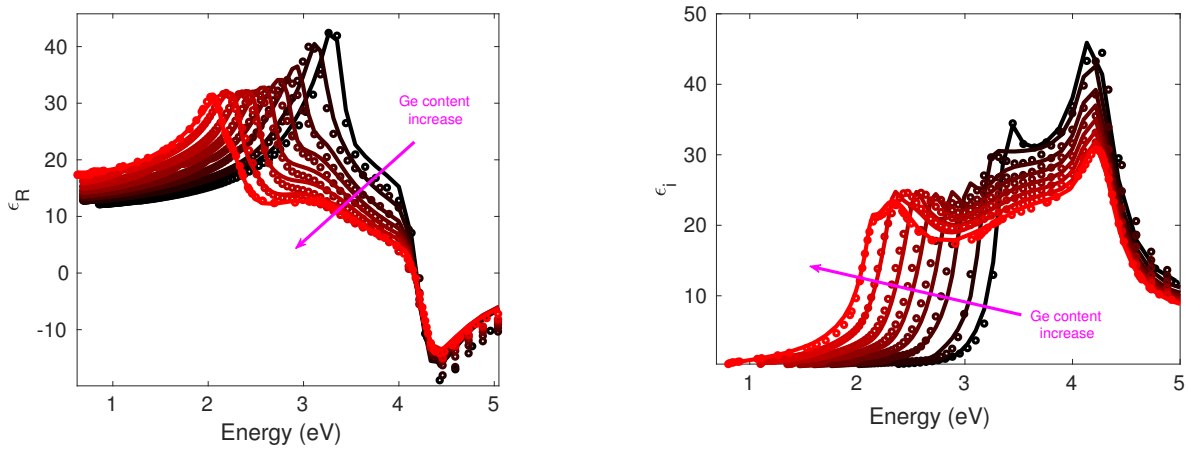
I.4 Results

In this section we present the main result of our study, that consist in the fit of our model on data from Nolot [47] (at 300K) for high energy and from Braunstein [51] (at various temperature) for low energy.

All the parameters are provided in section I.3.7. The fit of pure Si and pure Ge are shown in Fig. I.24(a) and I.24(b). Compared to a usual linear interpolation model (see Fig. I.9), our model demonstrates an overall good agreement to the fitting data. On Fig. I.24(a), the fitting error are concentrated around 3.8 eV, for the lower concentration of Ge, the darker curves.

On Fig. I.25, the prediction of our model are compared to the experimental data of the absorption coefficient from [51]. The parameters of section I.3.7 are used, kept as constant, in conjonction of the energy bandgap model of section I.3.6. Only the E_g parameters of the TL oscillators are varying according to Eq. I.56.

Fitting our model on the high energy, temperature dependent, permittivity of Si, Ge, and their alloys, was not performed due to the lack of available data, as mentionned in section I.2.4.



(a) Low Energy, log scale

(b) High Energy, linear scale

Figure I.24: Real (left) and imaginary (right) part of SiGe at room temperature of various Ge content (0, 0.2 , 0.389 , 0.513 , 0.635 , 0.75 , 0.8 , 1). Dots are experimental data from [47] and the solid line is our model.

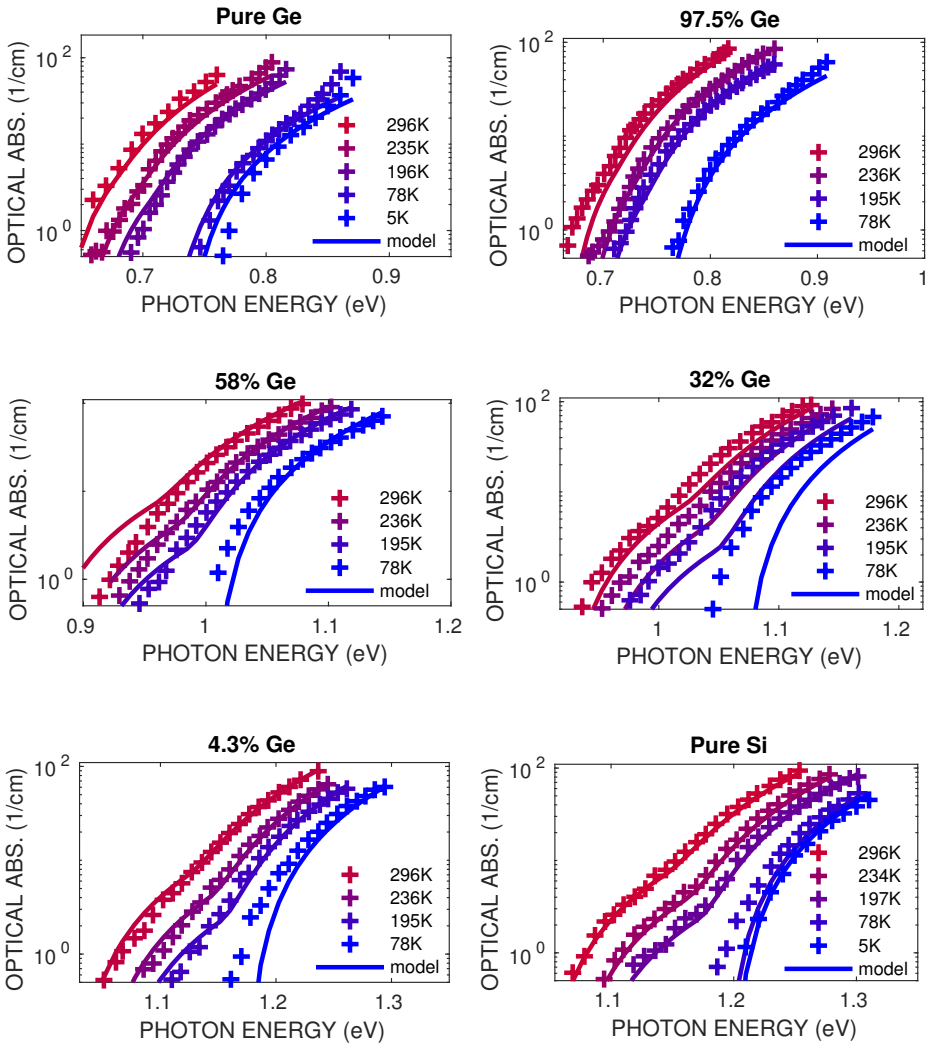


Figure I.25: Braunstein *et al.* [31] experimental data (represented by symbols) at various temperatures and Ge contents compared to the predictions of our model.

I.5 Numerical application

I.5.1 Motivation

In the following, our model, at 300 K, is compared with the usual linear interpolation methodology detailed in section I.2.3.4 (see Eq. I.51). A typical SPAD pixel has been chosen, with a layer of Silicon Germanium alloy, and the absorption in the active region is computed with our 2D in-house RCWA software (see section II.5). All the pixel dimensions are provided on the next section. Then the RCWA convergence study is performed. Finally, comparisons between our model, the usual linear interpolation methodology and the measured permittivity are presented on various Silicon Germanium alloys.

I.5.2 Pixel geometry

In order to test our permittivity model, we choose a 2D nanostructured SPAD pixel surrounded by deep isolation trench (DTI) and covered by an antireflective layer of Silicon oxide (SiO2). The bottom of the pixel is reflexive due to the presence of a metallic layer (Cu). This structure is similar to the one presented in [52]. The actual simulated pixel can be seen in Fig. I.26.

From top to bottom, the dimensions are (in μm):

- 2 thick air layer (in light pink);
- 0.15 thick SiO2 layer (in red);
- 0.4 thick grating layer, with SiO2 DTI. On x axis, the intervals composed of SiO2 are:
  - DTI: $[-2.5,\ -2.25]$, $[2.25,\ 2.5]$,
  - Grating: $[-2.1,\ -1.9]$, $[-1.7,\ -1.5]$, $[-1.3,\ -1.1]$, $[-0.9,\ -0.7]$, $[-0.5,\ -0.3]$, $[-0.1,\ 0.1]$, $[0.3,\ 0.5]$, $[0.7,\ 0.9]$, $[1.1,\ 1.3]$, $[1.5,\ 1.7]$, $[1.9,\ 2.1]$;
- 3.9 thick Si layer (in yellow), with DTI;
- 0.5 thick SiGe layer (in green), with DTI;
- 0.2 thick Si layer (in yellow), with DTI;
- 0.25 thick Cu layer (in blue).

The pixel has a 5 μm period and is centered at 0.

The permittivity data used are Palik [53] for SiO2, [47] or our model for Si, Ge and SiGe, McPeak [54] for Cu.

The figure of merits are:

- The reflection on top of the pixel;
- The absorption in all Si and SiGe layers;
- The euclidean norm of the electric and the magnetic fields.

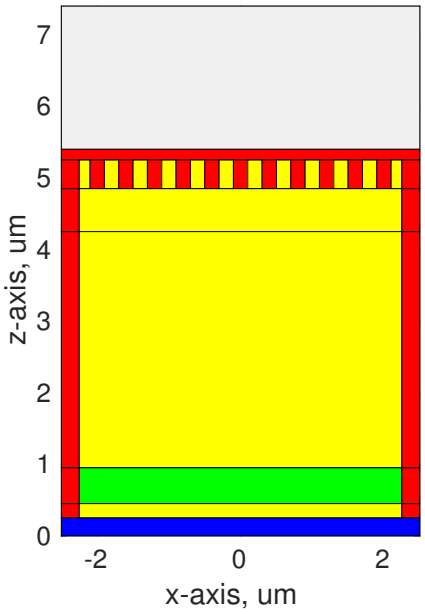


Figure I.26: 2D SPAD pixel simulated. Each color corresponds to a material: light pink for air, red for SiO2, yellow for Si, green for SiGe and blue for Cu.

I.5.3 Convergence study

Since the pixel geometry is matching a cartesian grid, *i.e.* there isn't any curved surface, as a lense for instance, the layer definition of the structure is straightforward and does not contain approximations. Thus, the only source of numerical approximation in RCWA simulations on this structure is the plane wave truncation, also referred as the mode truncation.

In order to define a sufficiently precise mode truncation, we first run, on a Si filled pixel, RCWA solving at 940 nm at various mode truncations. In Fig. I.27 is plotted the computed reflection and absorption at different mode truncation. From this figure of merit, 61 modes seem sufficient. To test this threshold, we also computed, between 900 nm and 1000 nm, on 201 evenly spaced wavelengths, the reflection and absorption with 61 and 101 modes truncations. Fig. I.28 exhibits no differences in the reflection and absorption spectrum, confirming that 61 modes are enough. This conclusion is also visible in Fig. I.29 where the euclidean norm of the electric and magnetic fields at 940 nm exhibits no significant differences between the 61 and the 101 modes truncations.

In total, 455 RCWA simulations (50 for Fig. I.27, 402 for Fig. I.28 and 3 for Fig. I.29) were run for a wall time inferior to 10 minutes.

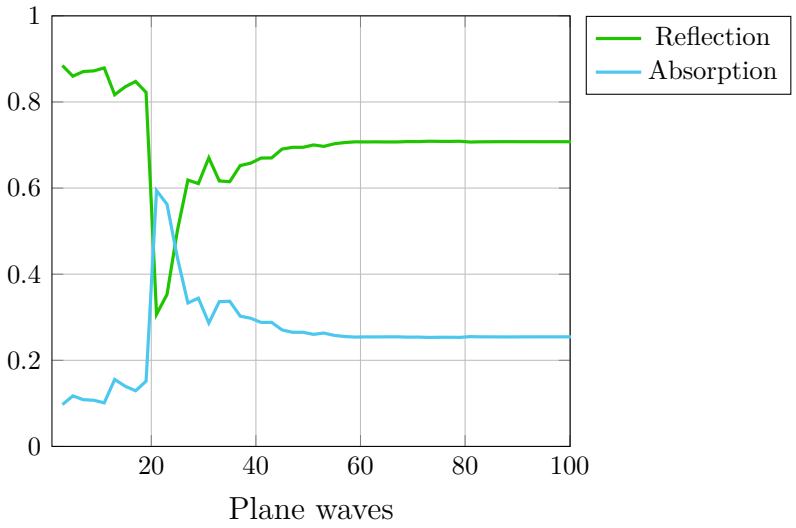


Figure I.27: Absorption and reflection at 940 nm, at various plane waves truncations.

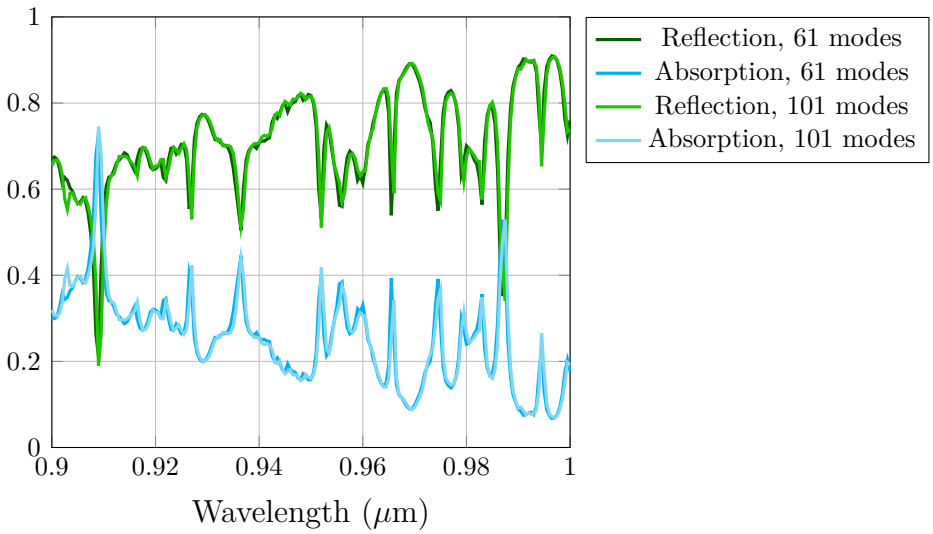
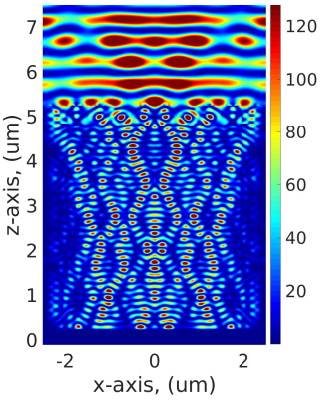
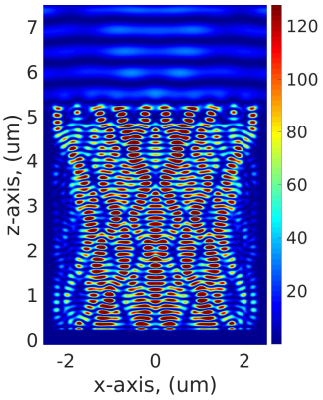


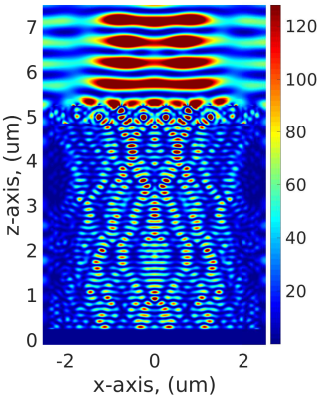
Figure I.28: Absorption (blue) and reflection (green) spectrum on the wavelength range 900-1000 nm, for 61 modes truncations (dark colored lines) and 101 modes truncations (pale colored lines).



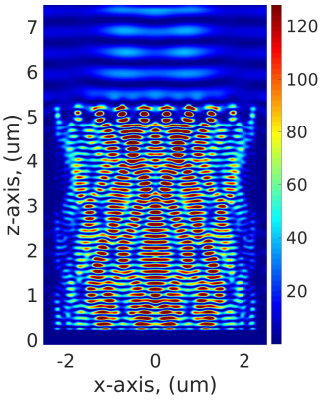
(a) $||E||^2$, 41 modes.



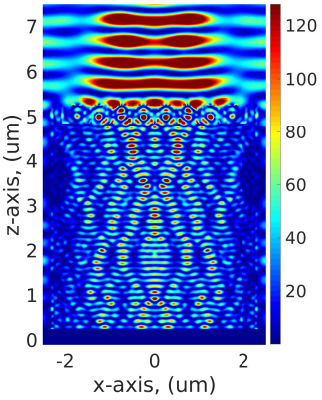
(b) $||H||^2$, 41 modes.



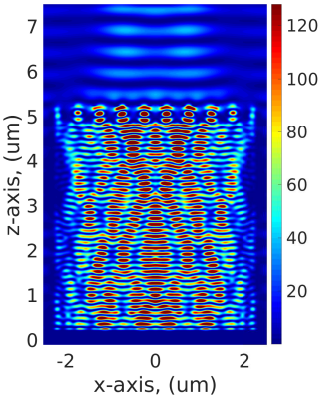
(c) $||E||^2$, 61 modes.



(d) $||H||^2$, 61 modes.



(e) $||E||^2$, 101 modes.



(f) $||H||^2$, 101 modes.

Figure I.29: Squared norm of both the electric field, E, and the magnetic field, H, at three mode truncations (41, 61 and 101). No difference can be seen between the figures of 61 and 101 modes truncations.

I.5.4 Results

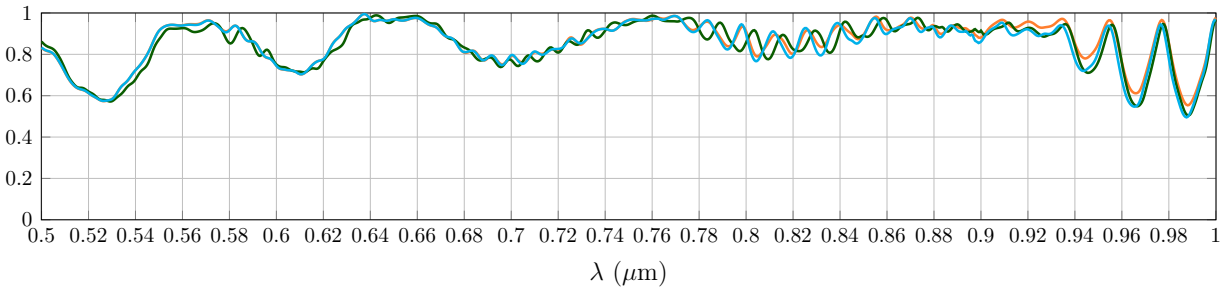
In Fig. I.30 and I.31, are available the absorption spectrum, at room temperature, on the pixel defined in section I.5.2, of three methods:

- blue lines are the simulations using the experimental data from [47];
- orange lines are the simulations using the linear interpolation methodology (see Eq. I.51);
- dark green lines are the simulations using our semi-empirical model.

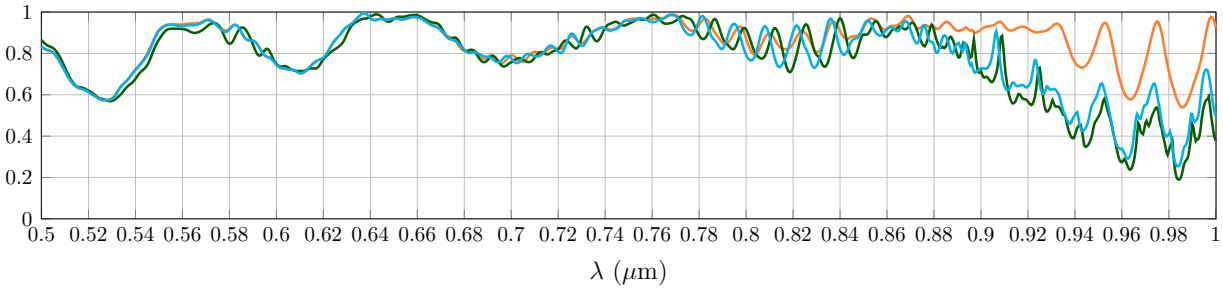
One can observe the drastic divergence of the usual interpolation methodology in the wavelength range from ~850 nm to 1000 nm, compared to the reference blue line, while our model, despite exhibiting a phase shift, follows the reference. The less Ge content in the SiGe layer, the wider is the interval of divergence of usual interpolation methodology. This divergence is visible in the figures of all Ge concentration except for the 88% one. On the visible range, the usual methodology performs better than our model that is exhibiting oscillations. This is mainly due to the absence of absorption in the SiGe layer. Indeed, Silicon is well absorbing light in the visible range and all light is absorbed in the other layers, as it can be seen in Fig. I.32 for 63% of Ge content.

In order to produce one absorption spectrum, 1001 RCWA simulations were run on evenly-spaced wavelengths between 500 nm and 1000 nm. Since each of the 6 Ge contents contains 3 absorption spectrum, 18018 simulations were run, for a wall time of approximately 3 hours.

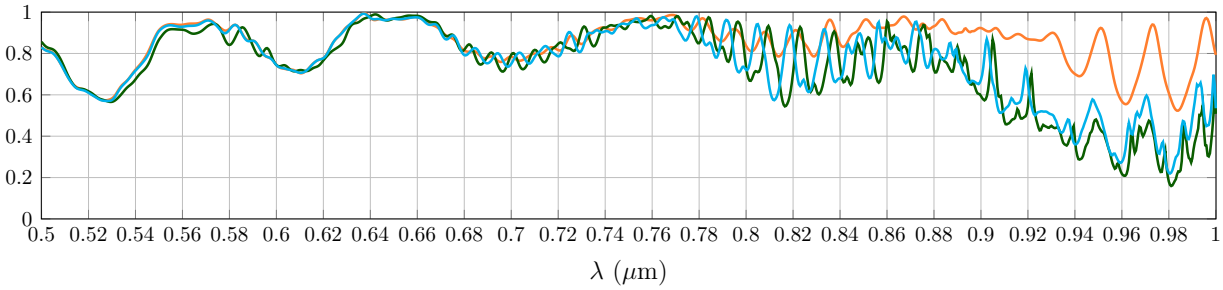
Finally, this study demonstrates the improvement of accuracy that our model of SiGe permittivity is achieving on optical simulations.



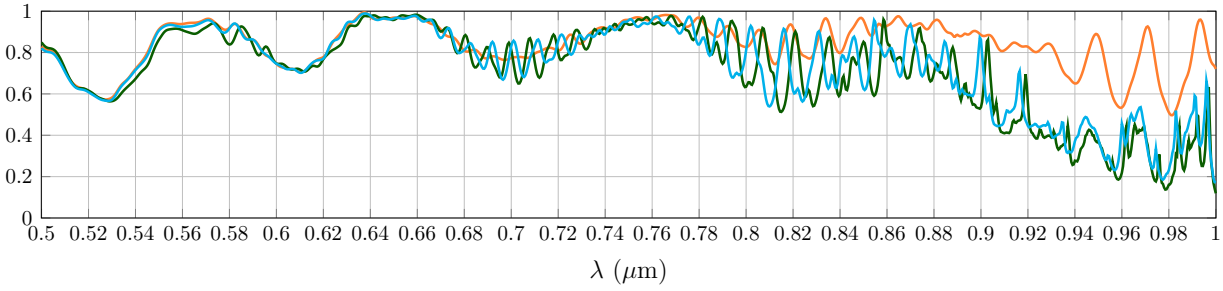
(a) 88% Ge



(b) 75% Ge

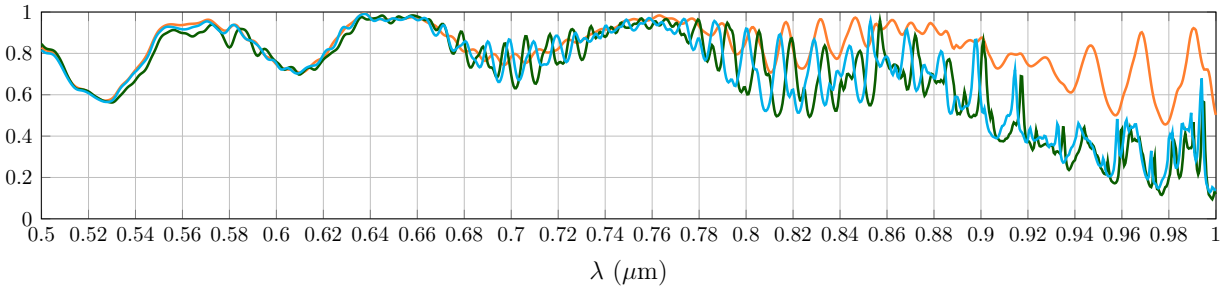


(c) 63% Ge

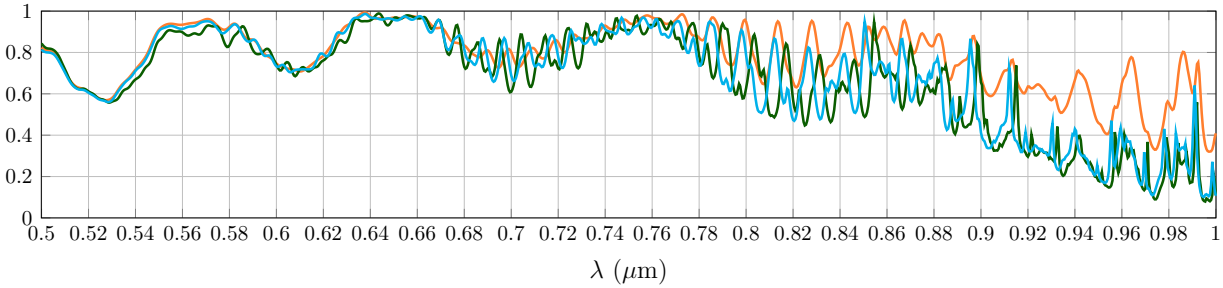


(d) 51% Ge

Figure I.30: Comparisons, at various Ge contents, of the absorption spectrum between the usual linear interpolation methodology (orange line) and our model (dark green line) with the reference of experimental data (blue line) from [47]. Only the permittivity of the SiGe layer is changed according to the model used.



(a) 39% Ge



(b) 20% Ge

Figure I.31: Comparisons, at various Ge contents, of the absorption spectrum between the usual linear interpolation methodology (orange line) and our model (dark green line) with the reference of experimental data (blue line) from [47]. Only the permittivity of the SiGe layer is changed according to the model used.

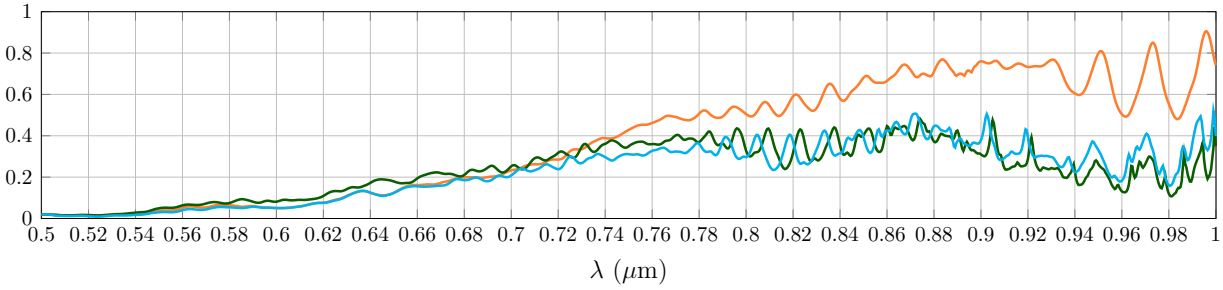


Figure I.32: Absorption spectrum of the SiGe layer only, at 63% Ge, for the usual linear interpolation methodology (orange line), our model (dark green line) and the experimental data (blue line) from [47]. The absorption in the range 500 nm 750 nm is low, confirming that the other layers of Si absorb the majority of the light.

Chapter II

Benchmark of numerical methods

Contents

II.1 Introduction

In general, numerical methods make it possible to solve complex problems, quickly validate a corrective action, optimize the geometrical properties of the device, access non-measurable quantities and better understand the physical phenomena involved. Fabricating CMOS imagers relies on a cost reduction policy, and thus efficient optical solvers are necessary.

In this chapter, three numerical methods for solving Maxwell's equations are compared on a pixel-like device. Benchmarking of the numerical methods can be found in the literature, for instance on nanometric scatterers [55], or on photonics crystal [56], but, to our knowledge, we cannot find such study for the particular application of pixel-like structures. The aim of this chapter is, after presenting the three numerical methods involved, to be able to choose the appropriate one for the study of the optical response of CMOS imagers.

Two stakes justify such study. Firstly, the company STMicroelectronics, funding this thesis, must be able to select the most efficient and appropriate software for its extensive studies on CMOS pixels. Secondly, the INRIA team Atlantis wants to know whether their software suite, DIOGENeS, implementing an innovative numerical method, can be used to improve the numerical treatment of light propagation in complex CMOS imagers. Both entities are thus highly interested in the results of a benchmark between the various numerical methods for solving optics in CMOS imager.

Such a benchmark is empirical and not analytical, in the sense that the comparisons are performed on execution time and simulation outputs, and not on evaluating formal algorithms complexity.

Firstly, the three numerical methods compared, namely the Finite Difference Time Domain (FDTD), the Discontinuous Galerkin Time Domain (DGTD) and the Rigorous Coupling Wave Analysis (RCWA) methods, are presented. Secondly, the comparison of these methods on structure of increasing complexity are performed.

II.2 Preliminaries

II.2.1 Notations for vector fields calculus

In this section, an operator, $\boldsymbol{\nabla}$, and three associated operations are defined. These notations allow compact notations and are constantly used in the rest of this chapter.

We define the vector operator $\boldsymbol{\nabla}$ by

$$\boldsymbol{\nabla} \;:=\; \left(\frac{\partial}{\partial x},\, \frac{\partial}{\partial y},\, \frac{\partial}{\partial z}\right), \tag{II.1}$$

that can be multiplied by a scalar function: given $T : \mathbb{R}^3 \mapsto \mathbb{R}$, we have

$$\boldsymbol{\nabla} T \;:=\; \left(\frac{\partial T}{\partial x},\, \frac{\partial T}{\partial y},\, \frac{\partial T}{\partial z}\right), \tag{II.2}$$

also known as grad(T). The operator $\boldsymbol{\nabla}$ can also be used with the usual scalar product and cross product: given $\mathbf{h} : \mathbb{R}^3 \mapsto \mathbb{R}^3$, we have

$$\boldsymbol{\nabla} \cdot \mathbf{h} \;:=\; \frac{\partial h_x}{\partial x} \;+\; \frac{\partial h_y}{\partial y} \;+\; \frac{\partial h_z}{\partial z}, \tag{II.3}$$

$$\boldsymbol{\nabla} \times \mathbf{h} \;:=\; \left(\frac{\partial h_z}{\partial y} - \frac{\partial h_y}{\partial z},\, \frac{\partial h_x}{\partial z} - \frac{\partial h_z}{\partial x},\, \frac{\partial h_y}{\partial x} - \frac{\partial h_x}{\partial y}\right), \tag{II.4}$$

also known as the divergence and the curl of the function $\mathbf{h}$. These two operators can be combined, we have in particular:

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \mathbf{h}) \;=\; 0, \tag{II.5}$$

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \mathbf{T}) \;=\; 0, \tag{II.6}$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) \;=\; \mathbf{A}\,(\mathbf{B} \cdot \mathbf{C}) \;-\; (\mathbf{A} \cdot \mathbf{B})\,\mathbf{C}, \tag{II.7}$$

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \mathbf{h}) \;=\; \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{h}) - \Delta \mathbf{h}. \tag{II.8}$$

for $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ vectors of $\mathbb{R}^3$, where $\Delta\mathbf{h}$ is the Laplacian of the function $\mathbf{h}$.

II.2.2 Fourier transform

In this section the well known definition of Fourier transform is recalled, as well as some of its properties that will be used later in this chapter.

Given d a positive integer and f an integrable function on $\mathbb{R}^d$ with value in $\mathbb{C}$, *i.e.* if $f \in L^1\left(\mathbb{R}^d\right)$, the Fourier transform of f, denoted $\hat{f}$, is:

$$\hat{f}\,(\boldsymbol{\xi}) \;=\; \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\mathrm{i}\,\langle \boldsymbol{\xi}, \mathbf{x} \rangle} f(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{II.9}$$

where $\boldsymbol{\xi} \in \mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ denotes the usual scalar product of $\mathbb{R}^d$.

The Fourier transform has the following properties:

- It is $\mathbb{C}$-linear: if f, g are integrable and $a \in \mathbb{C}$, then one has:

$$\widehat{af+g} = a\hat{f} + \hat{g}. \tag{II.10}$$

- If $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^d$ and $f \in L^1$, and one denotes:

$$\tau_{\mathbf{y}} f(\mathbf{x}) = f(\mathbf{x}-\mathbf{y}) \quad \text{and} \quad e_{\mathbf{y}} f(\mathbf{x}) = e^{\mathrm{i}\,\langle \mathbf{y},\mathbf{x}\rangle} f(\mathbf{x}), \tag{II.11}$$

  then,

$$\widehat{\tau_{\mathbf{y}} f} = e_{-\mathbf{y}}\hat{f} \quad \text{and} \quad \widehat{e_{\mathbf{y}} f} = \tau_{\mathbf{y}}\hat{f}. \tag{II.12}$$

  In other words, the Fourier transform of a translation, $\tau_{\mathbf{y}}$, of a function, f, is equivalent to a phase shift, $e_{\mathbf{y}}$, applied on $\hat{f}$, and inversely.

- If M is a matrix of $\mathrm{GL}_d(\mathbb{R})$ and if $g(\mathbf{x}) = f(M^{-1}\mathbf{x})$, for $x \in \mathbb{R}^d$, where $f \in L^1\left(\mathbb{R}^d\right)$, then

$$\hat{g}(\boldsymbol{\xi}) = |\mathrm{det}M|\, \hat{f}(M^t\boldsymbol{\xi}), \tag{II.13}$$

  where M^t is the transposed matrix of M. This property will be particularly useful to compute the Fourier transform of the indicator function of an ellipse in $\mathbb{R}^2$ (see appendix C).

- If $f \in L^1$ a function such as $\mathbf{x} \mapsto |\mathbf{x}| f(\mathbf{x})$ is integrable, then $\hat{f}$ is in $\mathcal{C}^1\left(\mathbb{R}^d, \mathbb{C}\right)$ and, for all $j \in \{1, 2,\ ...\ d\}$, one has:

$$\frac{\partial \hat{f}}{\partial \xi_j}(\boldsymbol{\xi}) = -\int_{\mathbb{R}^d} \mathrm{i}\, x_j e^{-\mathrm{i}\,\langle \boldsymbol{\xi},\mathbf{x}\rangle} f(\mathbf{x}) \frac{\mathrm{d}\mathbf{x}}{(2\pi)^{d/2}}, \tag{II.14}$$

  that is the Fourier transform of $\mathbf{x} \mapsto \mathrm{i}\, x_j f(\mathbf{x})$.

- If $f \in L^1 \cap \mathcal{C}^1\left(\mathbb{R}^d, \mathbb{C}\right)$, such as $\partial f/\partial x_j$ is integrable for all $j \in \{1, 2, ..., d\}$, then one has:

$$\widehat{\frac{\partial f}{\partial x_j}}(\boldsymbol{\xi}) = -\mathrm{i}\, \xi_j \hat{f}\,(\boldsymbol{\xi})\,. \tag{II.15}$$

II.3 Maxwell equations

In this section, Maxwell's equations, both in time-domain and in frequency-domain, are presented. We used [57] as a basis for the presentation.

II.3.1 Homogeneous material

Maxwell's equations in time-domain, for linear, isotropic and homogenous materials, that describe the spatio-temporal evolution of electromagnetic waves on a domain $\Omega \subset \mathbb{R}^3$, over a given time interval $[0,\ T]$, $T \geq 0$, are defined, for all $(x, t) \in \Omega \times [0,\ T]$ by:

$$\nabla \times \mathbf{E} \ = \ -\frac{\partial \mathbf{B}}{\partial t},$$

$$\begin{aligned} \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}, \\ \nabla \cdot \mathbf{D} &= \rho, \\ \nabla \cdot \mathbf{B} &= 0. \end{aligned} \tag{II.16}$$

with $\mathbf{E}$ and $\mathbf{B}$ the electric field and the magnetic induction, which are vectors of $\mathbb{R}^3$. We also introduce the electric displacement $\mathbf{D}$, the magnetic field $\mathbf{H}$, the density of free electric charges ρ and the free electric current density $\mathbf{J}$. All these quantities are dependent on position, $\mathbf{x} = (x, y, z)$, and time t.

To close the system II.16, relations between $(\mathbf{E}, \mathbf{B})$ and $(\mathbf{D}, \mathbf{H})$ are required. These constitutive relations are:

$$\mathbf{D} = \varepsilon *_{(t,x)} \mathbf{E} \qquad \text{and} \qquad \mathbf{B} = \mu *_{(t,x)} \mathbf{H}, \tag{II.17}$$

where, in the general case, ε and μ are tensors (3×3 matrices), depending on $\mathbf{x}$, t, $\mathbf{E}$ and $\mathbf{B}$. The symbol $*_{(t,x)}$ denotes the space and time convolution. Assumptions on materials implies particular properties on the two tensors ε and μ:

- If a material is linear, then ε and μ are independent of $\mathbf{E}$ and $\mathbf{B}$;
- If a material is isotropic, then ε and μ are diagonal matrices;
- If a material is homogenous, then ε and μ are constant.

It is customary to introduce ε_0 and μ_0 the vacuum permittivity and permeability, as well as ε_r and μ_r, the relative permittivity and permeability of the considered material. The constitutive relations are then written as:

$$\mathbf{D} = \varepsilon_0 \varepsilon_r \mathbf{E} \qquad \text{and} \qquad \mathbf{B} = \mu_0 \mu_r \mathbf{H}. \tag{II.18}$$

Maxwell's equations for linear, homogenous, isotropic, non-dispersive materials are then given by:

$$\begin{aligned} \nabla \times \mathbf{E} &= -\mu_0 \mu_r \frac{\partial \mathbf{H}}{\partial t}, \\ \nabla \times \mathbf{H} &= \varepsilon_0 \varepsilon_r \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J}. \end{aligned} \tag{II.19}$$

For numerical solving stability, the system II.19 is adimenssionated. The adimenssionated variables of $(\mathbf{E}, \mathbf{H}, \mathbf{J})$ are noted $\left(\widetilde{\mathbf{E}}, \widetilde{\mathbf{H}}, \widetilde{\mathbf{J}}\right)$, and are computed as:

$$\widetilde{\mathbf{H}} = Z_0 \mathbf{H}, \tag{II.20}$$

$$\widetilde{\mathbf{E}} = \mathbf{E}, \tag{II.21}$$

$$\widetilde{t} = c_0 \mathbf{t}, \tag{II.22}$$

$$\widetilde{\mathbf{J}} = Z_0 \mathbf{J}, \tag{II.23}$$

(II.24)

with $Z_0 = \sqrt{\dfrac{\mu_0}{\varepsilon_0}}$ the vacuum impedance and $c_0 = \dfrac{1}{\sqrt{\varepsilon_0 \mu_0}}$ the speed of light in vaccum. Hence, dropping the tilde for convenience, the normalized Maxwell's sytem is:

$$\begin{aligned} \nabla \times \mathbf{E} &= -\mu_r \frac{\partial \mathbf{H}}{\partial t}, \\ \nabla \times \mathbf{H} &= \varepsilon_r \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J}. \end{aligned} \tag{II.25}$$

II.3.2 Dispersive materials

The use of Fourier transform allows to replace the time variable t into the frequency one, ω; moving from the time-domain to the frequency-domain. Mathematically, it transforms the time convolution, for instance $\mathbf{D} = \varepsilon *_{(t)} \mathbf{E}$, into the product, $\widehat{D}(\omega) = \widehat{\varepsilon}(\omega)\, \widehat{E}(\omega)$.

In the case of a non-dispersive material, the constituve relations, given by Eq. II.17, become:

$$\mathbf{D} = \varepsilon \mathbf{E} \qquad \text{and} \qquad \mathbf{B} = \mu \mathbf{H}, \tag{II.26}$$

A material is said to be non-dispersive when its permittivity is constant in the frequency-domain. Conversely, a material is dispersive when its permittivity is dependent on ω in the frequency-domain.

For time-domain solvers (see section II.4.1 and II.4.2), an analytical formulation of the permittivity as a function of frequency is required. Various models for the permittivity are provided in section I.2.3.2 . The choice of a permittivity model is required for solving the time-domain Maxwell equations. This choice has a great influence on the solver performance, and for proprietary solver, such as Lumerical[1], it is usually kept as a business secret.

The permittivity model of the DIOGENeS software, is presented, as well as the modified Maxwell equations derived. In DIOGENeS, the model chosen is the Generalized Dispersive Model (GDM); it is a Padé approximation, and is given as a sum of zeroth order, first order and second order poles:

$$\varepsilon_{r,g}(\omega) = \varepsilon_\infty - \frac{\sigma}{\mathrm{i}\,\omega} - \sum_{l \in L_1} \frac{a_l}{\mathrm{i}\,\omega - b_l} - \sum_{l \in L_2} \frac{c_l - \mathrm{i}\,\omega d_l}{\omega^2 - e_l + \mathrm{i}\,\omega f_l}, \tag{II.27}$$

where ε_∞, σ, $(a_l)_{l \in L_1}$, $(b_l)_{l \in L_1}$, $(c_l)_{l \in L_2}$, $(d_l)_{l \in L_2}$, $(e_l)_{l \in L_2}$, $(f_l)_{l \in L_2}$ are in $\mathbb{R}$, and L_1, L_2 are non-overlapping sets of indices. Most of the standard dispersion models are included in this formulation (Drude, Drude-Lorentz, Sellmeir's law ...).

[1] `https://www.ansys.com/products/photonics`

The GMD-Maxwell's equations are then given by (see [58] for details):

$$
\begin{aligned}
&\frac{\partial \mathbf{H}}{\partial t} = -\nabla \mathbf{E}, \\
&\frac{\partial \mathbf{E}}{\partial t} = \frac{1}{\varepsilon_\infty}\left(\nabla \times \mathbf{H} - \boldsymbol{\mathcal{J}}_s - \boldsymbol{\mathcal{J}}_O - \sum_{l \in L_1 \cup L_2} \boldsymbol{\mathcal{J}}_l\right), \\
&\boldsymbol{\mathcal{J}}_0 = \left(\sigma + \sum_{l \in L_2} d_l\right)\mathbf{E}, \\
&\boldsymbol{\mathcal{J}}_l = a_l \mathbf{E} - b_l \mathbf{P} && \forall l \in L_1, \\
&\frac{\partial \mathbf{P}_l}{\partial t} = \boldsymbol{\mathcal{J}}_l && \forall l \in L_1, \\
&\frac{\partial \boldsymbol{\mathcal{J}}_l}{\partial t} = (c_l - d_l f_l)\mathbf{E} - f_l \boldsymbol{\mathcal{J}}_l - e_l \mathbf{P}_l, && \forall l \in L_2, \\
&\frac{\partial \mathbf{P}_l}{\partial t} = d_l \mathbf{E} + \boldsymbol{\mathcal{J}}_l && \forall l \in L_2.
\end{aligned}
\tag{II.28}
$$

II.3.3 Frequency-domain formulation

In this section, Maxwell's equations in frequency-domain, for linear, isotropic and dispersive materials, are presented. Assuming the EM field has an harmonic time dependence of the form $\exp\{-\mathrm{i}\,\omega t\}$, from which an arbitrary solution by Fourier superposition can be built, the equations for the amplitudes $\mathbf{E}(\omega, \mathbf{x})$, etc, are:

$$
\begin{aligned}
\nabla \cdot \varepsilon_0 \varepsilon \mathbf{E} &= \quad 0, && \text{(II.29)} \\
\nabla \cdot \mathbf{H} &= \quad 0, && \text{(II.30)} \\
\nabla \times \mathbf{H} &= -i\omega \varepsilon_0 \varepsilon \mathbf{E}, && \text{(II.31)} \\
\nabla \times \mathbf{E} &= \quad i\omega \mu_0 \mathbf{H}, && \text{(II.32)}
\end{aligned}
$$

The zero-divergence equations (Eq. II.29 and Eq. II.30) are not independent, but are obtained by taking the divergences in Eq. II.31 and Eq. II.32. These equations are called the frequency-domain Maxwell's equations, and they can be solved with the Rigorous Coupling Wave Analysis (RCWA) numerical method [59] or the Finite Element Method (FEM) [60].

II.3.4 Polarization

Polarization is a property of transverse waves which specifies the geometrical orientation of the oscillations. In a transverse wave, the direction of the oscillation is perpendicular to the direction of propagation. In our work, we will mainly focus on two polarization, named Transverse Electric (TE) and Transverse Magnetic (TM). Basically the electric fields has only an E_x component in TM, and an E_y component in TE, similarly for the magnetic field. To be absolutely clear, Fig. II.1 illustrates the TM polarization and Fig II.2 the TE polarization.

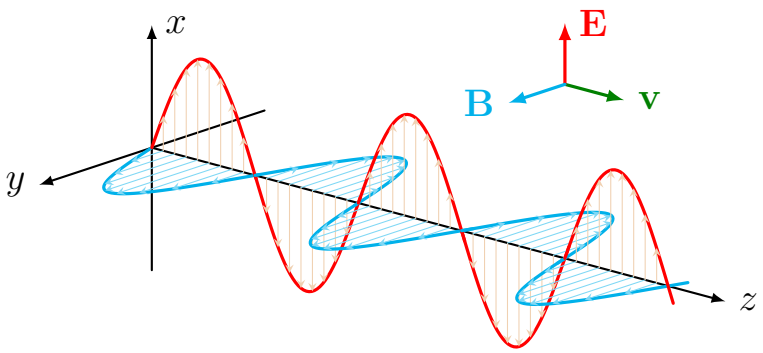


Figure II.1: A schematic view of TM electric and magnetic fields, with propagation axis along the z dimension.

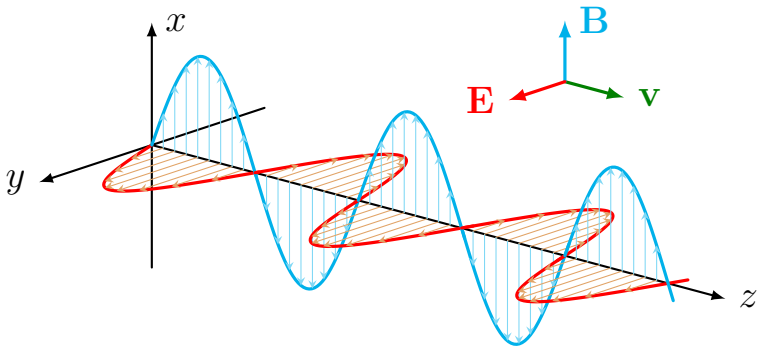


Figure II.2: A schematic view of TE electric and magnetic fields, with propagation axis along the z dimension.

II.4 Numerical methods

Three numerical methods for solving Maxwell's equations are presented in this section. First, the well-known Finite Difference Time-Domain method; then, the Discontinuous Galerkin Time Domain (DGTD) method; finally, the frequency-domain numerical method RCWA is detailed.

Since the reader might be unfamiliar with the DGTD method, and since a RCWA solver was implemented in this thesis, both methods are extensively presented, while the more common FDTD method is only briefly presented.

II.4.1 FDTD method

The Finite Difference Time-Domain (FDTD) method is undoubtedly one of the most popular method for time-domain computational nanophotonics. For instance, the reference optical solver, providing a solution of the Time-Domain Maxwell's Equations (see Eq. II.16), is Lumerical [2], which is based on the FDTD method. In STMicroelectronics, the Lumerical software is used on a daily basis by optical engineers to study and improve the desing of CMOS sensors.

The method is based on two Taylor expansions for both spatial and temporal derivatives. The popular version presented by K.S. Yee in 1966 [11] relies on a particular discretization of the computational space. The spatial element used is now referred to as Yee cells (see Fig. II.3). In the original Yee scheme, the spatial derivatives are discretized using second-order central differences while time integration is achieved with a second-order leap-frog scheme. As of today, FDTD represents an easy to use and efficient method to solve electromagnetic problems, combining simple implementation and high computational efficiency.

The FDTD method has a well-known limitation: a smooth discretization of curved geometries is impossible due to the fixed cartesian grid imposed by the Yee algorithm. This approximation leads to the staircasing effect (see Fig. II.4), which is an important source of inaccuracy [61]. To overcome this pitfall, one can either use a finer refinement of the grid, which leads to a rise of the computational cost, or exploit one of the numerous possible modifications of the FD method that have been proposed for tackling the staircasing effect [62].

The convergence of a FDTD simulation relies on the refinement of the cartesian mesh used. This mesh can be either uniform, or locally refined, One must remark, as shown in Fig. II.5, that a cartesian mesh cannot be strictly locally refined. A refinement in a smaller cube always propagates in the six directions of the refined cube.

A common way to build a well-refined cartesian mesh for a given structure is known, using the Lumerical terminology as the **meshfactor**. Basically, it represents a number of mesh dots per source wavelength, locally adapted to the permittivity of all materials. For instance, in a material of dielectric constant ε_r, supposing that the source wavelength is $\lambda = 900$ nm and the meshfactor is set at 10, then the mesh step is $\frac{900}{\sqrt{\varepsilon_r}10}$. The mesh step for air would be 90 nm, since the air dielectric constant is approximately 1.

In the following, all the FDTD simulations are characterized by a meshfactor value, indicating the refinement of its corresponding mesh. A higher meshfactor value implies

[2] `https://www.ansys.com/products/photonics`

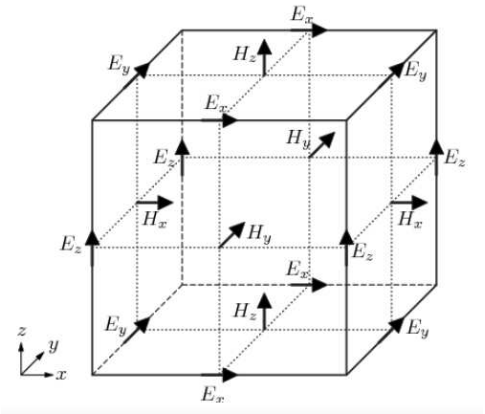


Figure II.3: A schematic view of a Yee cell. With a cartesian mesh, the Yee cell is a cuboid. The electric fields values are taken on the edges, while the magnetic fields is positioned on the faces of the cuboid.

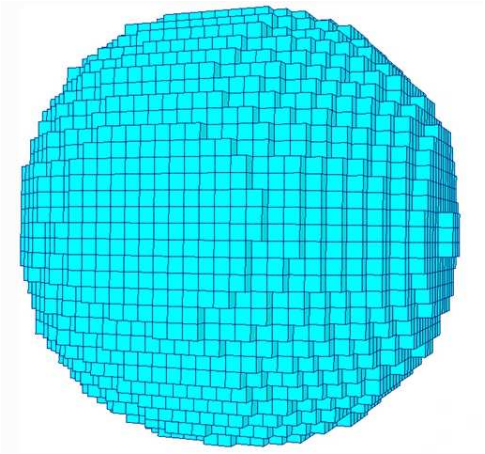


Figure II.4: Illustration of the staircasing effect with the approximation of a 3D sphere with a cartesian mesh. Figure taken from [63].

a more accurate simulation.

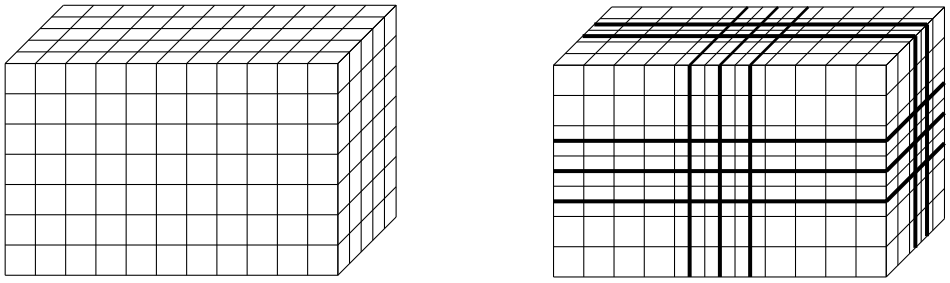


Figure II.5: Uniform mesh (left), and locally refined mesh (right) of a cuboid. The local refinement is performed on a smaller cuboid in the center of the original cuboid. The refinement is propagated up to the boundaries.

II.4.2 DGTD method

In this section, we present the DGTD method formulated on a conforming tetrahedral mesh and assuming an uniform interpolation order across all the elements of the mesh, which is essentially the outcome of the Ph.D. thesis of J. Viquerat [58]. This section is highly inspired from the Thesis of A. Gobé [64]. The DGTD method is also refered as a mixed between the Finite Volume method, using the concept of fluxes, and the FEM, where the quantities of interest, namely the electric and the magnetic fields, are approximed on tetrahedron as polynomials. The degree of these polynomials is refered as the interpolation order.

In the following, the start from the adimenssionated Maxwell's equations given by Eq. II.25.

II.4.2.1 Weak formulation

Let $\Omega \in \mathbb{R}^3$ be a bounded convex domain, and $\mathbf{n}$ the unitary outward normal to its boundary $\partial\Omega$. Let Ω_h be a discretization of Ω relying on a quasi-uniform triangulation $\mathcal{T}_h$, $(\exists \delta, \forall T_i \in \mathcal{T}_h, \forall k \in \mathcal{V}_i, h_k \leq \delta h_i$, where h_i is the size of the element T_i). This triangulation verifies $\mathcal{T}_h = \bigcup_{i=1}^{N} T_i$, where $N \in \mathbb{N}*$ is the number of mesh elements and $(T_i)_{i\in[[1,N]]}$ is the set of elements of $\mathcal{T}_h$. We denote by $a_{ik} := T_i \bigcap T_k$ the internal faces between the adjacent cells T_i and T_k.

One can now write the weak formulation of the problem of Eq. II.19, **locally**, in a cell T_i. Taking the L^2 scalar product of each term with a vector test function ψ, we obtain the following variational problem:

Find $(\mathbf{E}, \mathbf{H}) \in H_0(\mathbf{curl}, \Omega_h) \times H(\mathbf{curl}, \Omega_h)$ such that, $\forall \boldsymbol{\psi} \in H(\mathbf{curl}, \Omega_h)$,

$$\begin{aligned}
&\int_{T_i} \mu_r \frac{\partial \mathbf{H}}{\partial t} \cdot \boldsymbol{\psi} + \int_{T_i} \nabla \times \mathbf{E} \cdot \boldsymbol{\psi} = \mathbf{0}, \\
&\int_{T_i} \varepsilon_r \frac{\partial \mathbf{E}}{\partial t} \cdot \boldsymbol{\psi} - \int_{T_i} \nabla \times \mathbf{H} \cdot \boldsymbol{\psi} = -\int_{T_i} \mathbf{J} \cdot \boldsymbol{\psi}.
\end{aligned}$$

Then, using classical vectorial calculus and Green formulas yields, for all i in $[[1, N]]$:

$$\begin{aligned}
&\int_{T_i} \mu_r \frac{\partial \mathbf{H}}{\partial t} \cdot \boldsymbol{\psi} + \int_{T_i} \mathbf{E} \cdot \nabla \times \boldsymbol{\psi} = \int_{\partial T_i} (\boldsymbol{\psi} \times \mathbf{E}) \cdot \mathbf{n}_i, \\
&\int_{T_i} \varepsilon_r \frac{\partial \mathbf{E}}{\partial t} \cdot \boldsymbol{\psi} - \int_{T_i} \mathbf{H} \cdot \nabla \times \boldsymbol{\psi} = -\int_{T_i} \mathbf{J} \cdot \boldsymbol{\psi} - \int_{\partial T_i} (\boldsymbol{\psi} \times \mathbf{H}) \cdot \mathbf{n}_i.
\end{aligned}$$

By using the properties of the mixed product:

$$(\boldsymbol{\psi} \times E) \cdot n_i = (E \times n_i) \cdot \boldsymbol{\psi}, \tag{II.33}$$

one gets the local weak formulation of the Maxwell equations for the DGTD method:

$$\begin{aligned}
&\int_{T_i} \mu_r \frac{\partial \mathbf{H}}{\partial t} \cdot \boldsymbol{\psi} + \int_{T_i} \mathbf{E} \cdot \nabla \times \boldsymbol{\psi} = \int_{\partial T_i} (\mathbf{E} \times \mathbf{n}_i) \cdot \boldsymbol{\psi}, \\
&\int_{T_i} \varepsilon_r \frac{\partial \mathbf{E}}{\partial t} \cdot \boldsymbol{\psi} - \int_{T_i} \mathbf{H} \cdot \nabla \times \boldsymbol{\psi} = -\int_{T_i} \mathbf{J} \cdot \boldsymbol{\psi} - \int_{\partial T_i} (\mathbf{H} \times \mathbf{n}_i) \cdot \boldsymbol{\psi}.
\end{aligned} \tag{II.34}$$

II.4.2.2 Discretization in space

We start by defining the approximation space $\mathbf{V}_h$ such as:

$$\mathbf{V}_h \;=\; \left\{ v \in \left(\mathbf{L}^2\left(\Omega\right)\right)^3, \quad v_{|T_i} \in \left(\mathbb{P}_p\left(T_i\right)\right)^3, \quad \forall T_i \in \mathcal{T}_h \right\}, \tag{II.35}$$

where, for $p \in \mathbb{N}$, $\mathbb{P}_p\left(T_i\right)$ is the space of all polynomial functions of degree p on the cell T_i. We denote by $(\mathbf{E}_h,\ \mathbf{H}_h, \mathbf{J}_h)$, the semi-discrete fields seeked in the space $\mathbf{V}_h$, and their restriction on T_i as $\left(\mathbf{E}_{h|T_i},\ \mathbf{H}_{h|T_i}, \mathbf{J}_{h|T_i}\right)$. We also define the set of scalar basis functions $(\phi_{ik})_{1\leq k\leq d_i}$ for each T_i, with d_i the number of degree of freedom (d.o.f.) per dimension. Additionally, we associate the three vectors $\boldsymbol{\phi}_{ik}^v$ to each scalar basis function, such that:

$$\boldsymbol{\phi}_{ik}^1 = \begin{bmatrix} \phi_{ik} \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\phi}_{ik}^2 = \begin{bmatrix} 0 \\ \phi_{ik} \\ 0 \end{bmatrix}, \quad \boldsymbol{\phi}_{ik}^3 = \begin{bmatrix} 0 \\ 0 \\ \phi_{ik} \end{bmatrix}. \tag{II.36}$$

For a 3D system, we notice that $\mathbf{E}_i$ and $\mathbf{H}_i$ are vectors fields of $\mathbb{R}^3$:

$$\mathbf{E}_i = \begin{bmatrix} E_i^x \\ E_i^y \\ E_i^z \end{bmatrix}, \quad \mathbf{H}_i = \begin{bmatrix} H_i^x \\ H_i^y \\ H_i^z \end{bmatrix}, \tag{II.37}$$

which can both be locally expanded on the chosen set of basis functions:

$$\mathbf{E}_i^v = \sum_{j=1}^{d_i} E_{ij}^v \phi_{ij}, \quad \mathbf{H}_i^v = \sum_{j=1}^{d_i} E_{ij}^v \phi_{ij}, \quad v \in \{x, y, z\}. \tag{II.38}$$

For practical purpose, we define the six vectors of d_i components:

$$\bar{\mathbf{E}}_i^v = \begin{bmatrix} E_{i1}^v \\ \vdots \\ E_{id_i}^v \end{bmatrix}, \quad \bar{\mathbf{H}}_i^v = \begin{bmatrix} H_{i1}^v \\ \vdots \\ H_{id_i}^v \end{bmatrix}, \quad v \in \{x, y, z\}, \tag{II.39}$$

as well as the following $3d_i$ components vector:

$$\bar{\mathbf{E}}_i^v = \begin{bmatrix} \left(E_{ij}^x\right)_{1\leq j\leq d_i} \\ \left(E_{ij}^y\right)_{1\leq j\leq d_i} \\ \left(E_{ij}^z\right)_{1\leq j\leq d_i} \end{bmatrix}, \quad \bar{\mathbf{H}}_i^v = \begin{bmatrix} \left(H_{ij}^x\right)_{1\leq j\leq d_i} \\ \left(H_{ij}^y\right)_{1\leq j\leq d_i} \\ \left(H_{ij}^z\right)_{1\leq j\leq d_i} \end{bmatrix}. \tag{II.40}$$

Those vectors will be used to write the matrix-vector semi-discrete variational formulation of our system, analogously to system II.34. However, before that, the boundary terms in this formulation require some particular treatment before going further with the spatial discretization process.

II.4.2.3 Numerical fluxes

Given that the test functions and the unknowns can now be discontinuous at the interfaces a_{il} (between the cells T_i and T_l), it is important to notice that the surface integrals, define as:

$$\int_{a_{il}} (\mathbf{E}_h \times \mathbf{n}_{il}) \cdot \boldsymbol{\psi}, \quad \text{and} \quad \int_{a_{il}} (\mathbf{H} \times \mathbf{n}_{il}) \cdot \boldsymbol{\psi}, \tag{II.41}$$

are not clearly defined. Indeed, the fields $\mathbf{E}_h$ and $\mathbf{H}_h$ can relate to either the field value in T_i ($\mathbf{E}_{h,i}$ and $\mathbf{H}_{h,i}$) or T_l ($\mathbf{E}_{h,l}$ and $\mathbf{H}_{h,l}$). We need to define the numerical fluxes (or traces), which will allow us to recover a proper definition of the surface integrals. This will permit reconnecting the field values between neighboring cells.

In the current work, we present two very common flux choices. The first one is the centered flux, which is defined as:

$$\mathbf{E}_v = \frac{\mathbf{E}_{v,i} + \mathbf{E}_{v,l}}{2}, \quad \text{and} \quad \mathbf{H}_v = \frac{\mathbf{H}_{v,i} + \mathbf{H}_{v,l}}{2}. \tag{II.42}$$

This choice of flux leads to a non-dissipative DGTD scheme if combined with a Leap-Frog time integration and leads to L^2 spatial convergence of order p (in h^p) [65]. The second numerical flux considered is the upwind flux, which is defined as:

$$\mathbf{E}_v = \frac{1}{Y_i + Y_l} \left(\{Y\mathbf{E}\}_{il} + \alpha \mathbf{n} \times [[\mathbf{H}]]_{il} \right), \quad \text{and} \quad \mathbf{H}_v = \frac{1}{Z_i + Z_l} \left(\{Z\mathbf{H}\}_{il} + \alpha \mathbf{n} \times [[\mathbf{E}]]_{il} \right) \tag{II.43}$$

where $\{A\}_{il} = A_i + A_l$ is twice the mean value of A at the interface, $[[A]]_{il} = A_l - A_i$ is the jump of A at the interface, and $\alpha \in [0, 1]$ is a tunable parameter that allows to vary between the centered flux of Eq. II.42 (when $\alpha = 0$), to a fully upwind flux ($\alpha = 1$). The jump term of the upwind flux introduces dissipation in the DG scheme. This dissipation can help dampening non physical modes when instabilities occur [66]. It also leads to a better L^2 spatial convergence (as order $p+1$ (h^{p+1}) against order p (h^p) for the centered flux).

II.4.2.4 DG matrices

In this part, we present the Finite Elemente (FE) matrices which will allow us to write the matrix-vector form of the system II.34. We choose the test functions $\boldsymbol{\psi}$ such that they are the $3d_i$ vectors $\boldsymbol{\phi}^v_{ik}$, which constitutes the Galerkin choice:

$$\int_{t_i} \mu_r \frac{\partial \mathbf{H}_i}{\partial t} \cdot \boldsymbol{\phi}^v_{ik} + \int_{T_i} \mathbf{E}_i \cdot \nabla \times \boldsymbol{\phi}^v_{ik} = \sum_{l \in \mathcal{V}_i} \int_{a_{il}} (\mathbf{E}_* \times \mathbf{n}_{il}) \cdot \boldsymbol{\phi}^v_{ik},$$

$$\int_{t_i} \mu_r \frac{\partial \mathbf{E}_i}{\partial t} \cdot \boldsymbol{\phi}^v_{ik} - \int_{T_i} \mathbf{H}_i \cdot \nabla \times \boldsymbol{\phi}^v_{ik} = -\sum_{l \in \mathcal{V}_i} \int_{a_{il}} (\mathbf{H}_* \times \mathbf{n}_{il}) \cdot \boldsymbol{\phi}^v_{ik} - \int_{T_i} \mathbf{J}_i \cdot \boldsymbol{\phi}^v_{ik}. \tag{II.44}$$

$$\tag{II.45}$$

Mass Matrix We start by considering the time-derivative term of the $\mathbf{E}$ evolutionary equation of system II.44. The x component is, $\forall kin[[1, d_i]]$:

$$\int_{T_i} \varepsilon_r \frac{\partial \mathbf{E}_i}{\partial t} \cdot \boldsymbol{\phi}^x_{ik} = \int_{T_i} \varepsilon_r \frac{\partial \mathbf{E}^x_i}{\partial t} \cdot \phi^x_{ik}$$

$$
\begin{aligned}
&= \sum_{j=1}^{d_i} \frac{\partial}{\partial t} E_{ij}^x \int_{T_i} \varepsilon_r \phi_{ij} \phi_{ik} \\
&= \left(\mathbb{M}_i^{\varepsilon_r} \frac{\partial \bar{\mathbf{E}}_i^x}{\partial t} \right)_k
\end{aligned}
\tag{II.46}
$$

where $\mathbb{M}_i^{\varepsilon_r}$ is the mass matrix, of dimension $d_i \times d_i$:

$$
\mathbb{M}_{jk}^{\varepsilon_r} = \int_{T_i} \varepsilon_r \phi_{ij} \phi_{ik}, \tag{II.47}
$$

with $(j,k) \in [[1, d_i]]^2$.

Stiffness matrices We focus now on the curl integral of the **E** evolutionary equation of system II.44. The x component is, $\forall kin[[1, d_i]]$:

$$
\begin{aligned}
\int_{T_i} \mathbf{H}_i \cdot \nabla \times \boldsymbol{\phi}_{ik}^x &= \int_{T_i} \left(H_i^y \frac{\partial \phi_{ik}}{\partial z} - H_i^z \frac{\partial \phi_{ik}}{\partial y} \right) \\
&= \int_{T_i} \sum_{j=1}^{d_i} \left(H_{ij}^y \phi_{ij} \frac{\partial \phi_{ik}}{\partial z} - H_{ij}^z \frac{\partial \phi_{ik}}{\partial y} \right) \\
&= \sum_{j=1}^{d_i} H_{ij}^y \int_{T_i} \phi_{ij} \frac{\partial \phi_{ik}}{\partial z} - \sum_{j=1}^{d_i} H_{ij}^z \int_{T_i} \phi_{ij} \frac{\partial \phi_{ik}}{\partial y} \\
&= \left(\mathbb{K}_i^z \bar{\mathbf{H}}_i^y - \mathbb{K}_i^y \bar{\mathbf{H}}_i^z \right)_k \\
&= - \left(\mathbb{K}_i \times \bar{\mathbf{H}}_i \right)_k^x .
\end{aligned}
\tag{II.48}
$$

with the three stiffness matrices defined as:

$$
\left(\mathbb{K}_i^v \right)_{jk} = \int_{T_i} \phi_{ij} \frac{\phi_{ik}}{\partial v} \quad \text{for} \quad v \in \{x, y, z\}, \tag{II.49}
$$

with $(j,k) \in [[1, d_i]]^2$. We can also define the global $3d_i \times d_i$ stiffness matrix that will be used in the matrix-vector form of the final system:

$$
\bar{\mathbb{K}}_i = \begin{bmatrix} \mathbb{K}_i^x \\ \mathbb{K}_i^y \\ \mathbb{K}_i^z \end{bmatrix} . \tag{II.50}
$$

Flux matrices Finally, we consider the right handside of system II.44 that contains the flux contribution. Here we are using the centered flux Eq. II.42, but the upwind case Eq. II.43, as well as the generalization to other fluxes is straightforward. We also note that in the conforming case, the field expansion is defined on a_{il} over the basis functions

of cells T_i or T_l is equivalent. We proceed as previously, focusing on the x component of the flux of the $\mathbf{E}$ evolutionary equation of system II.44 :

$$\begin{aligned}\int_{a_{il}} (\ \mathbf{H}_x \times \mathbf{n}_{il}) \cdot \boldsymbol{\phi}_{ik}^x &= \int_{a_{il}} (H_*^y n_{il}^z - H_*^z n_{il}^y)\, \phi_{ik} \\ &= \int_{a_{il}} \left(\frac{H_i^y + H_l^y}{2} n_{il}^z - \frac{H_i^z + H_l^z}{2} n_{il}^y \right) \phi_{ik} \\ &= \frac{1}{2} \sum_j^{d_i} \left(\{H^y\}_{il} n_{il}^z - \{H^z\}_{il} n_{il}^y \right) \int_{a_{il}} \phi_{ij} \phi_{ik} \\ &= \left(\mathbb{S}_{il} \left(\bar{\mathbf{H}}_* \times \mathbf{n}_{il} \right) \right)_k^x\end{aligned}$$

where the flux matrices are:

$$\left(\mathbb{S}_{il}\right)_{jk} = \int_{a_{il}} \phi_{ij} \phi_{ik}, \tag{II.51}$$

with $(j,k) \in [[1, d_i]]^2$.

Semi-discrete formulation Thanks to the definition of the elementary matrices, we can define both global mass and flux matrices, which will allow to write the semi-discrete formulation in a compact form:

$$\bar{\mathbb{M}}_i^u = \begin{bmatrix} \mathbb{M}_i^u & \mathbb{0}_{d_i \times d_i} & \mathbb{0}_{d_i \times d_i} \\ \mathbb{0}_{d_i \times d_i} & \mathbb{M}_i^u & \mathbb{0}_{d_i \times d_i} \\ \mathbb{0}_{d_i \times d_i} & \mathbb{0}_{d_i \times d_i} & \mathbb{M}_i^u \end{bmatrix}, \quad \bar{\mathbb{S}}_i^u = \begin{bmatrix} \mathbb{S}_i^u & \mathbb{0}_{d_i \times d_i} & \mathbb{0}_{d_i \times d_i} \\ \mathbb{0}_{d_i \times d_i} & \mathbb{S}_i^u & \mathbb{0}_{d_i \times d_i} \\ \mathbb{0}_{d_i \times d_i} & \mathbb{0}_{d_i \times d_i} & \mathbb{S}_i^u \end{bmatrix}. \tag{II.52}$$

which leads to the following matrix vector expression of the semi-discrete DG scheme for Maxwell's equations:

$$\bar{\mathbb{M}}_i^{\mu_r} \frac{\partial \bar{\mathbf{H}}}{\partial t} = -\bar{\mathbb{K}}_i \times \bar{\mathbf{E}}_i \ + \ \sum_{l \in \mathcal{V}_i} \bar{\mathbb{S}}_{il} \left(\bar{\mathbf{E}}_* \times \mathbf{n}_i l \right), \tag{II.53}$$

$$\bar{\mathbb{M}}_i^{\varepsilon_r} \frac{\partial \bar{\mathbf{E}}}{\partial t} = \ \bar{\mathbb{K}}_i \times \bar{\mathbf{H}}_i \ - \ \sum_{l \in \mathcal{V}_i} \bar{\mathbb{S}}_{il} \left(\bar{\mathbf{H}}_* \times \mathbf{n}_i l \right) - \bar{\mathbb{M}}_i \bar{\mathbf{J}}_i. \tag{II.54}$$

II.4.2.5 Elements mapping

One of the strength ot the DG method is that the FE matrices described in section II.4.2.4 are not stored for each element of $\mathcal{T}_h$, but are calculated only once for all on a reference element noted $\hat{T}$ and then mapped on the considered physical tetrahedron T_i. Let $\hat{T}$ be defined as follows in the $\boldsymbol{\xi} = (\xi, \eta, \zeta)$ coordinate system:

$$\hat{T} \ = \ \left\{ \ (\xi, \eta, \zeta) \in \mathbb{R}_+^3 \mid \xi + \eta + \zeta \leq 1 \right\}. \tag{II.55}$$

Then, we define the physical tetrahedron in the $\mathbf{x} = x, y, z$ coordinate system as the image of $\hat{T}$ by a mapping $\boldsymbol{\psi}_{T_i}$ (see Fig. II.6) defined as:

$$\boldsymbol{\psi}_{T_i} \ : \ \hat{T} \mapsto T_i, \text{ such that, } \forall \boldsymbol{\xi} \in \hat{T}, \ \ \mathbf{x} = \boldsymbol{\psi}_{T_i}(\boldsymbol{\xi}). \tag{II.56}$$

The vertices of $\hat{T}$ are noted (A_1, A_2, A_3, A_4), and the vertices of T_i are noted (v_1, v_2, v_3, v_4). In this case, $\boldsymbol{\psi}_{T_i}$ is a linear combination of ξ, η and ζ, defined as:

$$\boldsymbol{\psi}_{T_i}(\boldsymbol{\xi}) = v_1 + (v_2 - v_1)\xi + (v_3 - v_1)\eta + (v_4 - v_1)\zeta. \tag{II.57}$$

Let $(\phi_{ij})_{j=1\dots d_i}$ be the basis functions on T_i, and $(\hat{\phi}_j)_{j=1\dots d_i}$ defined by $\hat{\phi}_j = \phi_{ij} \circ \boldsymbol{\psi}_{T_i}$ on $\hat{T}$. Then, we can write the mass matrix on the element T_i as:

$$\begin{aligned}
(\mathbb{M}_i)_{jk} &= \int_{T_i} \phi_{ij}(\mathbf{x})\phi_{ik}(\mathbf{x})\mathrm{d}\mathbf{x} \\
(\mathbb{M}_i)_{jk} &= \int_{\hat{T}_i} \hat{\phi}_j(\boldsymbol{\xi})\hat{\phi}_k(\boldsymbol{\xi}) \left|\mathbf{J}_{\boldsymbol{\psi}_{T_i}}\right| \mathrm{d}\boldsymbol{\xi}
\end{aligned}$$

where $\mathbf{J}_{\boldsymbol{\psi}_{T_i}}$ is the jacobian matrix of the mapping $\boldsymbol{\xi}$, defined as:

$$\left(\mathbf{J}_{\boldsymbol{\psi}_{T_i}}\right)_{jl} = \left(\frac{\partial \mathbf{x}_j}{\partial \boldsymbol{\xi}_l}\right)_{jl} = \begin{bmatrix} (v_2 - v_1)_x & (v_3 - v_1)_x & (v_4 - v_1)_x \\ (v_2 - v_1)_y & (v_3 - v_1)_y & (v_4 - v_1)_y \\ (v_2 - v_1)_z & (v_3 - v_1)_z & (v_4 - v_1)_z \end{bmatrix}. \tag{II.58}$$

In this case, the determinant $|\mathbf{J}_{\boldsymbol{\psi}_{T_i}}|$ is constant and only depends on (v_1, v_2, v_3, v_4). Hence, the mass matrix for each physical tetrahedron of $\mathcal{T}_h$ is a simple multiplication of the mass matrix calculated on the reference tetrahedron:

$$(\mathbb{M}_i)_{jk} = \left|\mathbf{J}_{\boldsymbol{\psi}_{t_i}}\right| \left(\hat{\mathbb{M}}\right)_{jk}. \tag{II.59}$$

A similar situation occurs for stiffness and flux matrices by using a change of variables (see [67] for additional details).

$$\begin{aligned}
(\mathbb{K}_i^v)_{jk} &= \sum_{m=1}^{3} \left[\left|\mathbf{J}_{\boldsymbol{\psi}_{t_i}}\right| \mathbf{J}_{\boldsymbol{\psi}_{t_i}}^{-1}\right]_{vm} \left(\hat{\mathbb{K}}^m\right)_{jk}. \\
(\mathbb{S}_i^v)_{jk} &= \left|\mathbf{J}_{\boldsymbol{\psi}_{t_i}}\right| \left|\mathbf{J}_{\boldsymbol{\psi}_{t_i}}^{-1} \hat{\mathbf{n}}\right| \left(\hat{\mathbb{S}}\right)_{jk}.
\end{aligned}$$

II.4.2.6 Polynomial expansion basis

In this part, we discuss the choice of the basis function $(\hat{\phi}_j)_{j=1\dots d_i}$. Over all the possible polynomial bases available, Lagrange polynomials are quite a common choice. They can be defined by a set of interpolation nodes distributed across the cell. The Lagrange interpolants L_i are defined by the following property:

$$L_i(\mathbf{x}_j) = \delta_{ij}, \quad \forall (i,j) \in [[1, d_i]]^2. \tag{II.60}$$

There must be the same number of nodes $\mathbf{x}_j$ and polynomials in order to define a complete basis. In a tetrahedron the number of Lagrande nodes needed to have a polynomial order p is equal to:

$$n(p) = \frac{(p+1)(p+2)(p+3)}{6}, \tag{II.61}$$

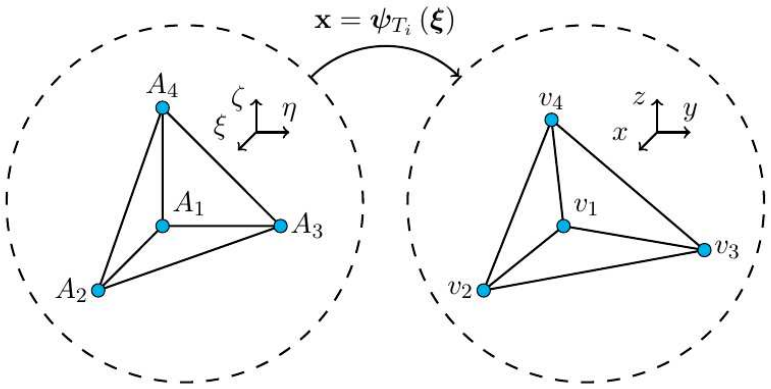


Figure II.6: Linear mapping from the reference element $\hat{T}$ to the physical element T_i. In this figure, one has $\boldsymbol{\varepsilon} = (\varepsilon, \zeta, \eta)$.

and for each of its faces, it is:

$$s(p) = \frac{(p+1)(p+2)}{2}. \tag{II.62}$$

Lagrange polynomials with equispaced nodes are chosen as it allows a simple integration of the elementary matrices on the reference element since the node positions are known exactly.

II.4.2.7 Time discretization

Previously, the semi-discrete scheme, Eq. II.54, was obtained by discretizing the spatial derivatives of Maxwell's equations. Similarly, the time derivatives need to be considered for the time discretization. Then, we consider the reduced problem of the 1D Maxwell's equations:

$$\begin{aligned} \mu_r \frac{\partial H_y}{\partial t} &= \frac{\partial E_z}{\partial x}, \\ \varepsilon_r \frac{\partial E_z}{\partial t} &= \frac{\partial H_y}{\partial x} + j(t). \end{aligned}$$

By dropping the spatial subscripts of the unknowns, the semi-discrete formulation associated to this system becomes:

$$\begin{aligned} \mathbb{M}_i^{\mu_r} \frac{\partial H_i}{\partial t} &= \mathbb{K}_i E_i + [E_*]_{x_{i-1}}^{x_i}, \\ \mathbb{M}_i^{\varepsilon_r} \frac{\partial E_i}{\partial t} &= \mathbb{K}_i H_i + [H_*]_{x_{i-1}}^{x_i} - \mathbb{M}_i J_i, \end{aligned} \tag{II.63}$$

$$\tag{II.64}$$

with,

$$E_* = \frac{1}{Y_i + Y_l}(\{YE\}_{il} + \alpha[[H]]_{il}), \qquad H_* = \frac{1}{Z_i + Z_l}(\{ZH\}_{il} + \alpha[[E]]_{il}). \tag{II.65}$$

In the previous equalities, subsection l designates $i-1$ or $i+1$. When trying to go from system II.63 to **??**, a non-symmetric $\mathbb{A}$ matrix is obtained in the general case:

$$\mathbb{A} = \begin{bmatrix} \mathbb{A}_{\alpha,H} & \mathbb{A}_H \\ \mathbb{A}_E & \mathbb{A}_{\alpha,E} \end{bmatrix}. \tag{II.66}$$

The off-diagonal blocks represent both the centered part of the flow and the stiffness part, while the diagonal blocks represent the upwind contribution. Therefore, $\mathbb{A}$ is strictly anti-diagonal in the case of centered flows. Furthermore, the following properties are obtained in the case of a homogeneous medium ($Y_i = Y_l = Y$):

- $\mathbb{A}_{\alpha,H}$ and $\mathbb{A}_{\alpha,E}$ are equal, *e.g.* $\mathbb{A}_{\alpha,H} = \mathbb{A}_{\alpha,E} := \mathbb{A}_\alpha$;
- $\mathbb{A}_H$ and $\mathbb{A}_E$ are multiples of each other, *e.g.* $\mathbb{A}_H = \frac{\varepsilon_r}{\mu_r}\mathbb{A}_E$.

To begin with, we consider the simple case of vacuum ($\varepsilon_r = \mu_r = 1$). In this case, one has a symmetric $\mathbb{A}$ matrix and thus a diagonalizable system. Thus, one can just retain the corresponding formulation in the diagonalized basis, for the study of the time-stepping schemes. This has the effect of reducing to a system of ODEs of the form:

$$\frac{\partial \phi}{\partial t}(t) = \lambda\phi(t) + b(t) := f(t, \phi(t)), \tag{II.67}$$

for each λ eigenvalue of $\mathbb{A}$. A number of time-integration techniques from the ODE community are appropriate to solve Eq. II.67.

We can classify time-stepping methods into two main categories: the explicit time integration techniques, and the implicit methods. For the first category, the time state $\phi(t + \Delta t)$ is computed explicitly from $\phi(t)$. For the second category, the time-updated solution is obtained by solving an implicit expression of the form $g(\phi(t), \phi(t + \Delta t)) = 0$. It leads to the resolution of a linear system of equations at each time-step. As expected, the implicit method is more expensive than the explicit technique. But, implicit methods are generally unconditionally stable. This means that any choice of Δt leads to a stable algorithm. For explicit methods, a numerical criterion on the time-step called the Courant–Friedrichs–Lewy (CFL) condition, must be respected. Otherwise, the resulting algorithm will be unstable and blow up. In conclusion, explicit time-stepping generally requires much more time-steps but each time-step is significantly less numerically costly. Hereafter, solutions are investigated on intervals of the form $[0, T]$ with $T > 0$, discretized in time-steps of length Δt. In order to refer to the discrete approximation of $\phi(t_n)$, with $t_n = n\Delta t$, we used the notation ϕ^n. A family of DG methods, called space-time DG methods (see [68, 69]) is designed to deal with time derivatives similarly to space. The main disadvantage of these methods is that they usually lead to an implicit scheme.

In the rest of this section, we present the second and fourth-order Runge-Kutta (RK) schemes. RK time schemes are a class of multi-stage algorithms based on the multiple evaluations of the right side of Eq. II.67 to evolve the system in time. Suppose that Eq. II.67 is integrated between t and $t + \Delta t$:

$$\phi(t + \Delta t) = \phi(t) \int_t^{t+\Delta t} f(u, \phi(u))\mathrm{d}u. \tag{II.68}$$

We can approximate Eq. II.68 using a quadrature formula with s terms:

$$\phi(t+\Delta t) \backsimeq \phi(t) + \Delta t \sum_{j=1}^{s} \beta_j f\left(t+\delta_j \Delta t, \phi(t+\delta_j \Delta t)\right), \tag{II.69}$$

where $(\beta_j)_{j\in[[1,s]]}$ and $(\delta_j)_{j\in[[1,s]]}$ are constants depending of the choice of the quadrature formula. In order to evaluate the $\phi(t+\delta_j\Delta t)$ values, RK methods use a so-called prediction/correction technique. It builds on the previous guesses to calculate the next one. The n^{th} time-step with an s-stage RK algorithm is written as follow:

$$\begin{aligned}
\phi_1 &= f(t_n, \phi^n), \\
\phi_k &= f(t_n + \delta_k \Delta t,\ \phi^n + \Delta t \sum_{j=1}^{s} \alpha_{j,k}\phi_j) \text{ for } k \in [[2,s]], \\
\phi^{n+1} &= \phi^n + \Delta t \sum_{j=1}^{s} \beta_j \phi_j,
\end{aligned} \tag{II.70}$$

where we suppose that $\phi_0 = \phi^n$. The system II.70 is implicit in the general case. The summation in each intermediate stage extends to the maximum number of stages. Explicit RK schemes can be obtained if one has $\alpha_{j,k} = 0,\ \forall k \geq j$. A second order RK scheme is given by:

$$\begin{aligned}
\phi_1 &= f(t_n, \phi^n), \\
\phi_2 &= f(t_n \Delta t, \phi^n + \Delta t \phi_1), \\
\phi^{n+1} &= \phi^n + \frac{\Delta t}{2}(\phi_1 + \phi_2).
\end{aligned}$$

Similarly, the most classical version of an explicit fourth-order RK algorithm is given by:

$$\begin{aligned}
\phi_1 &= f(t_n, \phi^n), \\
\phi_2 &= f(t_n \frac{\Delta t}{2}, \phi^n + \frac{\Delta t}{2}\phi_1), \\
\phi_3 &= f(t_n \frac{\Delta t}{2}, \phi^n + \frac{\Delta t}{2}\phi_2), \\
\phi_4 &= f(t_n \Delta t, \phi^n + \Delta t \phi_3), \\
\phi^{n+1} &= \phi^n + \frac{\Delta t}{6}(\phi_1 + 2\phi_2 + 2\phi_3 + \phi_4).
\end{aligned}$$

The main drawback of explicit methods is the CFL condition. This time-step restriction is computed from an energy-based stability study. In DIOGENeS, theoretical results from [65] are used. Consequently, for a space discretization with polynomial order p, the time-step is chosen as follows:

$$\Delta t_p = c_p \min_{T_i \in \mathcal{T}_h} \frac{V_{T_i}}{A_{T_i}}, \tag{II.71}$$

where V_{T_i} is the volume, A_{T_i} is the area of the boundary the cell T_i, and c_p is an order-dependent constant. The maximal acceptable value for c_p, can be determined on a basic test case.

II.4.2.8 Boundary condition

Each computational setup needs to be closed by boundary conditions, which depends on the physics of the problem. In principle, the physical domain might be infinite; in practice, the computed domain must be finite. Moreover, the computed domain is a compact of $\mathbb{R}^3$, and usually a cube, or a rectangular parallelepiped is used. Then the boundary conditions are equations applied to the unknowns on the boundary of the computed domain. There are four common boundary conditions, the perfect electric conductor (PEC), the perfect magnetic conductor (PMC), the absorbing boundary condition (ABC) and the periodic boundary condition (PBC).

In DGTD methods, the boundary conditions are imposed by adding a layer of cells called 'ghost cells'. This layer is outside the computational domain (see Fig. II.7). Then, specific values of the fields are set in these ghost cells. Thus, the behavior of the solution on the boundary is controlled, without any special treatment, via numerical fluxes. Hereafter, the field inside the ghost cells are noted $\mathbf{E}_{gc}$ and $\mathbf{H}_{gc}$, while the fields in the boundary cells are denoted $\mathbf{E}_{bc}$ and $\mathbf{H}_{bc}$.

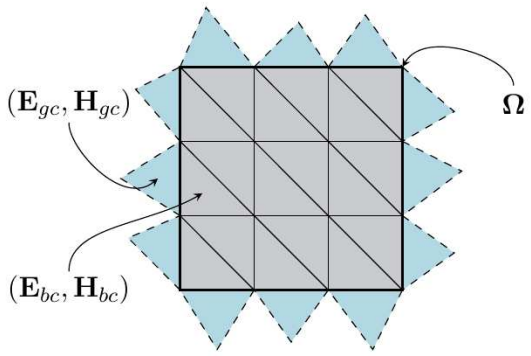


Figure II.7: Ghost cells layer on the computational domain boundary. Boundary conditions are naturally handled via numerical fluxes and ghost cells.

Perfect electric conductor condition PEC is an idealized material which exhibits infinite electrical conductivity. A zero tangential electric field, and a zero normal magnetic field, are enforced by setting the field values in the ghost cells on a PEC boundary as:

$$\mathbf{E}_{gc} = -\mathbf{E}_{bc} \quad \text{and} \quad \mathbf{H}_{gc} = \mathbf{H}_{bc}. \tag{II.72}$$

Perfect magnetic conductor condition PMC represents the reciprocal of the PEC. Most of the time, it is used to impose symmetry planes (see below). A zero tangential magnetic field, and a zero normal electric field, are enforced by setting the field values in the ghost cells on a PMC boundary as:

$$\mathbf{E}_{gc} = \mathbf{E}_{bc} \quad \text{and} \quad \mathbf{H}_{gc} = -\mathbf{H}_{bc}. \tag{II.73}$$

Absorbing boundary condition ABC allows to partially absorb fields radiating out from the physical domain. There exist several forms of these conditions. The first-order

Silver-Müller boundary conditions is often used (see [67]):

$$\mathbf{n} \times (\mathbf{E} + Z(\mathbf{n} \times \mathbf{H})) = \mathbf{0},$$
$$\mathbf{n} \times (\mathbf{H} - Y(\mathbf{n} \times \mathbf{E})) = \mathbf{0},$$

with $\mathbf{n}$ the exterior normal derivative. Imposing the Silver-Müller boundary condition is similar to setting the incoming flux to zero on the boundary. For this reason, its expression depends on the upwind factor α:

$$Y_{gc} = Y_{bc},$$
$$\mathbf{E}_{gc} = \mathbf{0} + \frac{1-\alpha}{Y_{bc}} \mathbf{n} \times \mathbf{H}_{bc},$$
$$\mathbf{H}_{gc} = \mathbf{0} - \frac{1-\alpha}{Z_{bc}} \mathbf{n} \times \mathbf{E}_{bc}.$$

These conditions perfectly absorb normally-incident plane waves, when imposed on the boundary. However, when waves are incident at increasing angles, its performance decreases.

Periodic boundary condition PBC can be used in order to simulate infinite mono or bi-directional arrays, considering one elementary pattern. Cells from a periodic boundary face are matched with cells on the opposite boundary face of the domain. This way, every cell has a neighbor which is well-defined, and standard fluxes can be applied. A periodic mesh is a prerequisite to use PBC. In other words, opposite faces in the periodic direction must match.

When a simulation domain is symmetrical in two directions, only a quarter of the whole simulation domain can be simulated in order to retrieve the electric and the magnetic field on the whole volume. This is done by applying PEC and PMC to the symmetrical directions. Since these conditions actually depend on the source polarization, let's describe them with an example.

Consider a simulation volume exhibiting a symmetry in both x and y directions, with periodic condition in x and y, and absorbing condition in both z_{min} and z_{max}. One could simulate the whole domain, without taking advantage of the symmetries, by applying the following boundary conditions, for all source polarizations:

- PBC on x_{min}, x_{max}, y_{min} and y_{max} boundaries;
- ABC on z_{min} and z_{max} boundaries.

Then, taking advantage of the domain symmetries in both x and y axis, so simulating only a quarter of the domain but retrieving the solution on the whole volume, can be performed by applying the following boundary conditions, for TE source polarization:

- PMC on x_{min} and x_{max} boundaries;
- PEC on y_{min} and y_{max} boundaries;

- ABC on z_{min} and z_{max} boundaries;

and for TM source polarization:

- PEC on x_{min} and x_{max} boundaries;
- PMC on y_{min} and y_{max} boundaries;
- ABC on z_{min} and z_{max} boundaries.

The reduction of the simulation volume by a factor of 4 allows, *a priori*, to reduce the simulation time by a factor of 4.

Finally, one must remark that the PEC are, within Lumerical software, denoted as, "antisymmetric condition", while the PMC are denoted as "symmetric condition". And the ABC conditions are denoted as PML, for Perfect Match Layer.

II.4.2.9 Perfectly matched layers

A novel numerical concept was developed in 1994 by Bérenger in order to overcome the limitations of ABC. The objective is to absorb the waves radiated from a system. To do so, Bérenger defined an artificial volume surrounding the physical domain in which a damping should occur progressively. Those artificial volumes are known as perfectly matched layers (PML). Outgoing waves propagate in the physical domain toward the PML, and cross the interface. No reflection occurs, and the wave is progressively damped by the artificial medium while it continues to propagate in the PML. At one point, the wave will encounter the boundary of the computational domain, which is usually PEC or ABC. In both cases, the remainder of the wave will be totally (PEC) or partially (ABC) reflected toward the domain. In practice, the wave will travel a second time over the PML length, causing more damping. In general, when it re-enters again in the physical domain, the amplitude of the wave is attenuated by several orders of magnitude. For this reason, the error induced by PMLs is supposed to be small enough not to lose the benefits of high-order methods. ABCs can be seen as a "geometric" condition (i.e. it only becomes more efficient with a larger distance from the source), while PMLs take advantage of the high-order spatial discretization to allow higher levels of damping.

PMLs have evolved since Bérenger's implementation, and several variations are now available [70]. One can cite the uniaxial PML (UPML) [58], and the complex frequency-shifted PML (CFS-PML), which are in use for Maxwell's equations. The last one is used in DIOGENeS.

II.4.2.10 Illumination sources

Plane waves The most simple kind of source are plane waves. Indeed, they are commonly used in numerical electromagnetics to determine the fundamental properties of a physical system even if these waves correspond to an asymptotic physical configuration (i.e. any radiating source propagating on a sufficiently large distance should look like a plane wave). Resonances and modes of a physical system can be excited depending on the spectral profile of the source. Monochromatic plane waves are the most basic kind of

time dependence. In the temporal domain, a monochromatic wave, with frequency ω_0, is formulated as:

$$\mathbf{E}(t) = \mathbf{E}_0 \sin(\omega_0 t). \tag{II.74}$$

One can notice that the Fourier transform of II.74 is proportional to the Dirac function δ_{ω_0}. This type of source can be useful, but running a whole time-domain simulation only to obtain the response of the system at one frequency may not be worth it, and a frequency-domain method, such as RCWA, may be more suited for this type of source. In order to obtain a broader frequency spectrum, one can use:

$$\mathbf{E}(t) = \mathbf{E}_0 \sin(\omega_0(t - t_0)) \exp\left(-\frac{(t-t_0)^2}{2\sigma^2}\right), \tag{II.75}$$

which is a Gaussian function of width σ centered around t_0, and modulated by a sine function. Its Fourier transform has the expression:

$$\widehat{\mathbf{E}}(\omega) = \mathbf{E}_0 \mathrm{i}\, \sigma \sqrt{\frac{\pi}{2}} \exp\left(\mathrm{i}\, \omega t_0\right) \left(\exp\left(-\frac{\sigma^2(\omega-\omega_0)^2}{2}\right) - \exp\left(-\frac{\sigma^2(\omega+\omega_0)^2}{2}\right)\right), \tag{II.76}$$

This means that such a pulse traveling through a structure will excite it on a wideband of wavelength, and not just on a single one.

TF/SF formulation Different possibilities are available in order to impose the plane waves inside the physical domain. A first simple one is to use the ghost cells on the ABCs. This solution is no longer viable when PMLs are used (imposing fields in the PMLs will directly damp them). A possible alternative is to define an additional artificial contour inside the physical domain and between PMLs and the scatterer, on which the field could be imposed directly. For periodic domains, the artificial contour can be replaced with infinite artificial surfaces.

First, we consider the splitting of the electric field in two parts:

$$\mathbf{E}_{tot}(\mathbf{x}, t) = \mathbf{E}_{inc}(\mathbf{x}, t) + \mathbf{E}_{sca}(\mathbf{x}, t), \tag{II.77}$$

where $\mathbf{E}_{tot}$ is the total field, $\mathbf{E}_{inc}$ is the incident field, and $\mathbf{E}_{sca}$ is the scattered field. In our case, the incident field is a plane wave and it is known since it is imposed analytically. We consider now a splitting of the computational domain in two regions: the first one which in which the total field is computed, while in the second region the scattered field is computed. We call the interface between these two regions the total field/scattered field (TF/SF) interface. The DGTD formulation II.54 is still valid in both regions, and no modification is required, except for the computation of the fluxes between both regions. Consider a TF/SF interface between two cells, such that the local cell T_i is in the total field region and its neighbor cell T_l is the scattered field region. The upwind flux calculated for cell T_i is:

$$\mathbf{E}_{*,tot} \;=\; \frac{1}{Y_i + Y_l}\left(\{Y\mathbf{E}_{tot}\}_{il} + \alpha \mathbf{n} \times [[\,\mathbf{H}_{tot}]]_{il}\right), \tag{II.78}$$

with $\{Y\mathbf{E}_{tot}\}_{il} = Y_i\mathbf{E}_{i,tot} + Y_l\mathbf{E}_{l,tot}$ and $[[\mathbf{H}_{tot}]]_{il} = \mathbf{H}_{l,tot} - \mathbf{H}_{i,tot}$. However, the field values corresponding to cell T_l are not $\mathbf{E}_{l,tot}$ and $\mathbf{H}_{l,tot}$, but $\mathbf{E}_{l,sca}$ and $\mathbf{H}_{l,sca}$. Hence, the flux

formulation must be modified to account for this difference. By considering Eq. II.77, one easily shows that the right flux can be calculated as follows:

$$\mathbf{E}_{*,tot} = \frac{1}{Y_i + Y_l}(\{Y\mathbf{E}\}_{il} + \alpha\mathbf{n} \times [[\mathbf{H}]]_{il}) + \frac{1}{Y_i + Y_l}(Y_l\mathbf{E}_{inc} + \alpha\mathbf{n} \times \mathbf{H}_{inc}). \quad \text{(II.79)}$$

Symmetrically, the upwind flux for cell T_l in the scattered field region is:

$$\mathbf{E}_{*,sca} = \mathbf{E}_* - \mathbf{E}_{*,inc}. \quad \text{(II.80)}$$

II.4.2.11 DIOGENeS software suite

This DGTD method is implemented in the object-oriented framework of the DIOGENeS[3] computational nanophotonics software suite, which is programmed in Fortran 2008. DIOGENeS (DIscOntinuous GalErkin Nanoscale Solvers) is a software suite, which is dedicated to the numerical modeling of nanoscale wave-matter interactions in 3D. The initial (and current) version of this software concentrates on light-matter interactions with nanometer scale structures for applications to nanophotonics and nanoplasmonics. DIOGENeS relies on a two layer architecture. The core of the suite is a library of generic software components (data structures and algorithms) for the implementation of high order DGTD (Discontinuous Galerkin Time-Domain) and HDGFD (Hybridizable Discontinuous Galerkin Frequency-Domain) schemes formulated on unstructured tetrahedral and hybrid structured/unstructured (cubic/tetrahedral) meshes. This library is used to develop dedicated simulation software for time-domain and frequency-domain problems relevant to nanophotonics and nanoplasmonics, considering various material models. The parallelization of these fullwave solvers relies on coarse grain SPMD (Single Program Multiple Data) strategy, which is implemented with the MPI message passing standard. In addition, DIOGENeS has recently evolved with the inclusion of components dedicated to geometrical modeling (GFactory) and numerical optimization (Optim). The architecture of DIOGENeS is sketched in Fig. II.8.

The DGTD method being presented, we focus, in the next section, on the third numerical method used for solving Maxwell equations in this thesis: the RCWA method.

II.4.3 RCWA

In the following sections, the modal numerical method named Rigorous Coupling Waves Analysis (RCWA), also referred to as the Fourier Modal Method (FMM), is described. It was first introduced in [13] and then further developed by the use of the scattering matrix formalism in [71] and [59]. The following section is based mainly on [59] and [72]

Firstly, the layers definitions of the simulated structure are recalled. Secondly, the source definition is made precise. Thirdly, the computation of the Fourier transform of each layer permittivity is explained. Fourthly, the eigenmodes problem of each layer is described, followed by the scattering matrix that reassembles the solution on the whole structure. Finally the electric and magnetic fields computation is explained.

[3] `https://diogenes.inria.fr`

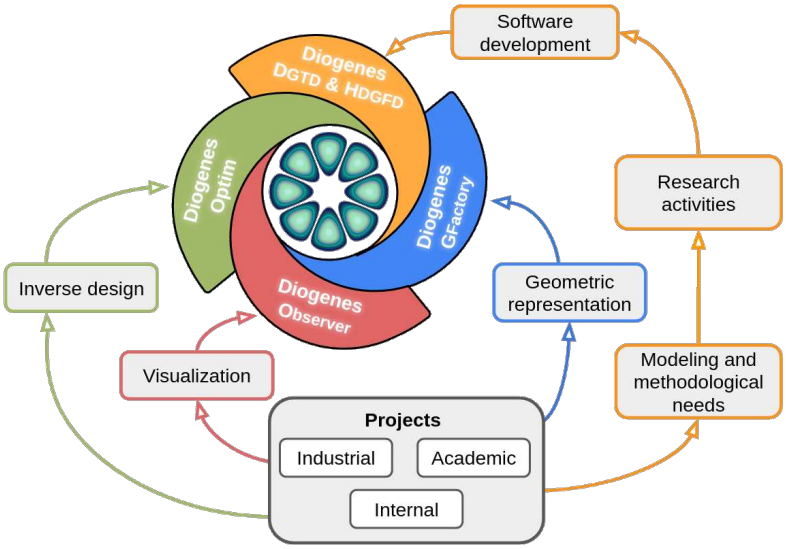


Figure II.8: Components of the DIOGENeS software suite, schema extracted from `https://diogenes.inria.fr`.

II.4.3.1 Geometric definitions and Fourier domain

For the rest of this chapter, the light is assumed to pass from top to bottom. Thus, the light is at normal incidence when its propagation vector is parallel to the z-axis.

In RCWA, the coordinate system is oriented such that the z-axis is normal to the layers of the structure, and the structure is assumed to be periodic in the xy-plane, with a rectangular basis, defined by two vectors of $\mathbb{R}^2$, $\mathbf{I}_1$ and $\mathbf{I}_2$. The xy-basis of the structure being rectangular, $\mathbf{I}_1$ and $\mathbf{I}_2$ are perpendicular. We assume the xy-coordinate to be chosen such that $\mathbf{I}_1$ is collinear to the x axis, and $\mathbf{I}_2$ is collinear to the y axis. By convention, we assume, in the following, that the origin of the xy-plane is exactly at the center of the structure xy-basis. Finally, the xy-basis of the simulation is a centered rectangle of side a_x, a_y, defined by:

$$\begin{aligned} a_x &:= ||\mathbf{I}_1||, \qquad x_{min} := -\frac{a_x}{2}, \qquad x_{max} := \frac{a_x}{2}, \\ a_y &:= ||\mathbf{I}_2||, \qquad y_{min} := -\frac{a_y}{2}, \qquad y_{max} := \frac{a_y}{2}. \end{aligned} \tag{II.81}$$

For simplicity, we will lump the transverse coordinates in the xy-plane into a vector $\mathbf{r}$.

The main requirement for the RCWA method is that the structure is defined into z-layers of constant permittivity along the z axis. Namely the structure is cut into layers along the z axis and the permittivity within each layer varies only on the transverse coordinate, *i.e.* we have $\varepsilon(\mathbf{r}, z) = \varepsilon(\mathbf{r})$ within each layer.

In the following, this layer definition requires specific notations: each layer is indexed by i, with thickness d_i, extending from z_i to $z_i + d_i$, with layer 1 extending from $z_1 \in \mathbb{R}$ to $z_1 + d_1$. The infinite half-space, $\{(x, y, z) \in \mathbb{R}^3 \mid z < z_1\}$, under the structure, is denoted layer 0. And the infinite half-space on top of the structure, $\{(x, y, z) \in \mathbb{R}^3 \mid z > z_{M-1} + d_{M-1}\}$ (assuming there is $M - 1$ layer in the structure) is denoted layer M.

To determine the reciprocal lattice of the Fourier domain, namely the reciprocal space, or the plane wave space, we first define the real space primitive lattice vector matrix whose columns are $\mathbf{I}_1$ and $\mathbf{I}_2$:

$$L_r \;=\; [\mathbf{I}_1\ \mathbf{I}_2] \;=\; \begin{bmatrix} I_{1x} & I_{2x} \\ I_{1y} & I_{2y} \end{bmatrix}. \tag{II.82}$$

Then, the reciprocal lattice is defined by the columns of:

$$L_k \;=\; 2\pi L_r^{-T}, \tag{II.83}$$

where L_r^{-T} denotes the transpose of the inverse matrix of L_r. Since we assumed that the basis is rectangular, we have:

$$L_r \;=\; \begin{bmatrix} a_x & 0 \\ 0 & a_y \end{bmatrix} \quad \text{and} \quad L_k \;=\; \begin{bmatrix} \frac{2\pi}{a_x} & 0 \\ 0 & \frac{2\pi}{a_y} \end{bmatrix}. \tag{II.84}$$

II.4.3.2 Source

Typical problems require solving for transmission, reflection, or absorption spectra of a structure. In these cases, incident radiation from layer 0 is assumed to be a plane wave propagating in the negative z direction. The incident wavevector is $\mathbf{k}_0$, with an in-plane component in the reciprocal space $\mathbf{k}$. For a plane wave of wavelength λ, with incident angle θ and ϕ (see Fig. II.9), we have:

$$\mathbf{k} = \begin{bmatrix} k_0 + k_{par,x} \\ k_0 + k_{par,y} \end{bmatrix}. \tag{II.85}$$

where k_0, $k_{par,x}$ and $k_{par,y}$ are defined as:

$$\begin{aligned} k_0 \quad &:= \frac{2\pi}{\lambda}, \\ k_{par,x} &= k_0 * \sin(\theta) * \sin(\phi), \\ k_{par,y} &= k_0 * \sin(\theta) * \cos(\phi). \end{aligned}$$

II.4.3.3 Units and conventions

In this section, Maxwell's equations in time-harmonic form, solved by the RCWA numerical method, are recalled, as well as the units chosen for both the electric and the magnetic field.

We will adopt a derivation and notion similar to those used in [72]. The starting point is Maxwell's equations in time-harmonic form, assuming an $\exp(-i\omega t)$ time dependence:

$$\begin{aligned} \nabla \cdot \varepsilon_0 \varepsilon \mathbf{E} &= \quad 0, \\ \nabla \cdot \mathbf{H} &= \quad 0, \\ \nabla \times \mathbf{H} &= -i\omega\varepsilon_0\varepsilon\mathbf{E}, \\ \nabla \times \mathbf{E} &= \quad i\omega\mu_0\mathbf{H}, \end{aligned} \tag{II.86}$$

For simplicity, we assumed that materials are linear and nonmagnetic. These assumptions are satisfied for most calculations for nanophotonics. From here on after, we will use Lorentz-Heaviside units, so that the speed of light c and vacuum impedance $Z_0 := \sqrt{\frac{\mu_0}{\varepsilon_0}}$ are both unity (making c, μ_0 and ε_0 drop out). These units are effectively the same as starting with SI units and scaling with:

$$\begin{aligned} \sqrt{\mu_0 \varepsilon_0}\, \omega_{SI} &\to \omega, \\ \sqrt{\frac{\mu_0}{\varepsilon_0}} \mathbf{H}_{SI} &\to \mathbf{H}, \\ \mathbf{E}_{SI} &\to \mathbf{E}. \end{aligned} \tag{II.87}$$

This change of units brings the electric and magnetic fields onto the same scale and the temporal and spatial frequency scales onto the same scale, providing better numerical conditioning in the implementation, and simplifying notation. In these new units, Maxwell's equations becomes

$$\begin{aligned} \nabla \times \mathbf{H} &= -i\omega\varepsilon\mathbf{E}, \\ \nabla \times \mathbf{E} &= \quad i\omega\mathbf{H}. \end{aligned} \tag{II.88}$$

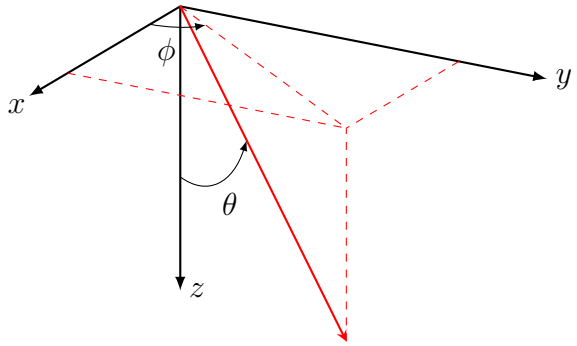


Figure II.9: Definition of the incident light angles: θ and ϕ. The red vector represents the light propagation vector.

II.4.3.4 Fourier transforms

In this section, the Fourier decomposition of the fields, and the Fourier transform of the permittivity, involved in the RCWA numerical method, are described.

The next step is to take the spatial Fourier transform in the xy-plane. Because of the periodicity and separability of the z-axis, the fields have the form:

$$\begin{aligned} \mathbf{E}(\mathbf{r}, z) &= \sum_{\mathbf{G}} \mathbf{E}_{\mathbf{G}}(z) \exp\left(\mathrm{i}\ \ (\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}\right), \\ \mathbf{H}(\mathbf{r}, z) &= \sum_{\mathbf{G}} \mathbf{H}_{\mathbf{G}}(z) \exp\left(\mathrm{i}\ \ (\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}\right). \end{aligned} \tag{II.89}$$

where $\mathbf{k}$ is the source wave vector in-plane component (see Eq. II.85) and $\mathbf{G}$ is a reciprocal lattice vector: $L_k^{-1}\mathbf{G} \in \mathbb{Z}^2$. The plane wave truncation, namely the truncation of the infinite summation of Eq. II.89, leads to the introduction of, in principle, the only approximation parameter required for RCWA. Assuming that a fixed set of discrete $\mathbf{G}$ has been chosen as well as an ordering (the same of all layers), we denote by $\mathbf{h}(z)$, the vector

$$[\mathbf{H}_{\mathbf{G}_1}(z),\ \mathbf{H}_{\mathbf{G}_2}(z),\ \dots\ ]^T, \tag{II.90}$$

and similarly for $\mathbf{e}(z)$:

$$[\mathbf{E}_{\mathbf{G}_1}(z),\ \mathbf{E}_{\mathbf{G}_2}(z),\ \dots\ ]^T. \tag{II.91}$$

The Fourier transform of the in-plane dielectric function, or permittivity, noted ε, is:

$$\varepsilon_{\mathbf{G}} \ = \ \frac{1}{|L_r|}\int_{cell} \varepsilon(\mathbf{r}) \exp\left(-\mathrm{i}\ \mathbf{G}\cdot\mathbf{r}\right) \mathrm{d}\mathbf{r}, \tag{II.92}$$

where the integral is over one unit cell of the real space lattice, a centered rectangle, defined by Eq. II.81. In general ε can be a tensor, but we assume in the following that the z-axis is separable for simplicity; *i.e.* that it is of the form:

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} & \varepsilon_{xy} & 0 \\ \varepsilon_{yx} & \varepsilon_{yy} & 0 \\ 0 & 0 & \varepsilon_z \end{bmatrix}. \tag{II.93}$$

In this case, each component can be Fourier transformed separately, and we obtain five sets of coefficients: $\widehat{\varepsilon}_{\mathbf{G},xx}$, $\widehat{\varepsilon}_{\mathbf{G},xy}$, $\widehat{\varepsilon}_{\mathbf{G},yx}$, $\widehat{\varepsilon}_{\mathbf{G},yy}$, $\widehat{\varepsilon}_{\mathbf{G},z}$. Using the same ordering of $\mathbf{G}$ as for Eq. II.90, we can form the block Toeplitz matrix $\widehat{\varepsilon}_{xx}$ whose (m,n)-th element is defined by:

$$\widehat{\varepsilon}_{xx,mn} = \varepsilon_{(\mathbf{G}_m - \mathbf{G}_n),xx}. \tag{II.94}$$

That is, the (m,n) entry of $\widehat{\varepsilon}_{xx}$ is the Fourier coefficient corresponding to the reciprocal lattice vector $\mathbf{G}_m - \mathbf{G}_n$. The matrices $\widehat{\varepsilon}_{xx}$, $\widehat{\varepsilon}_{xy}$, $\widehat{\varepsilon}_{yx}$, $\widehat{\varepsilon}_{yy}$ and $\widehat{\varepsilon}_z$ are defined analogously. Generally, the hat symbol ($\widehat{\ }$) is used to refer to square matrix operators acting on the reciprocal space, on $\mathbf{G}$. Using these definitions, we can Fourier transform Maxwell's equations (Eq. II.88 for each field component):

$$\begin{aligned}
\mathrm{i}\ \hat{k}_y h_z(z) - h_y'(z) &= -\mathrm{i}\ \omega d_x(z), \\
h_x'(z) - \mathrm{i}\ \hat{k}_x h_z(z) &= -\mathrm{i}\ \omega d_y(z), \\
\mathrm{i}\ \hat{k}_x h_y(z) - \mathrm{i}\ \hat{k}_y h_x(z) &= -\mathrm{i}\ \omega \widehat{\varepsilon}_z e_z(z), \\
\mathrm{i}\ \hat{k}_y e_z(z) - e_y'(z) &= -\mathrm{i}\ \omega h_x(z), \\
e_x'(z) - \mathrm{i}\ \hat{k}_x e_z(z) &= -\mathrm{i}\ \omega h_y(z), \\
\mathrm{i}\ \hat{k}_x e_y(z) - \mathrm{i}\ \hat{k}_y e_x(z) &= -\mathrm{i}\ \omega e_z(z),
\end{aligned} \tag{II.95}$$

where primes denote differentiation with respect to z, and $\hat{k}_x$ is a diagonal matrix with entries $(k_x + G_{1x},\ k_x + G_{2x},\ \dots\ )$ and analogously for $\hat{k}_y$. The first two equations of II.95 contain d_x and d_y, which are the Fourier coefficients of the displacement field $\mathbf{D}$. To obtain a closed set of equations, we need to relate the displacement field, d_x and d_y, to

the electric field e_x and e_y. This turns out to be subtle due to the need to apply the proper Fourier factorization rules, taking into account account discontinuities in both ε and E. For now, we assume that there exists a matrix $\mathcal{E}$ such that,

$$\begin{bmatrix} -d_y(z) \\ d_x(z) \end{bmatrix} = \mathcal{E} \begin{bmatrix} -e_y(z) \\ e_x(z) \end{bmatrix}. \tag{II.96}$$

The definition of $\mathcal{E}$ is made precise next to the description of our implementation of RCWA, in section II.5.3.

Starting from the system of equations II.95, eliminating the z components using the third equations of II.95, the fourth and fith equations become:

$$-\hat{k}_y\hat{\varepsilon}_z^{-1}\hat{k}_x h_y(z) \;+\; \hat{k}_y\hat{\varepsilon}_z^{-1}\hat{k}_y h_x(z) \;+\; \mathrm{i}\,\omega e_y'(z) \;=\; \omega^2 h_x(z), \tag{II.97}$$

$$-\mathrm{i}\,\omega e_x'(z) \;+\; \hat{k}_x\hat{\varepsilon}_z^{-1}\hat{k}_x h_y(z) \;-\; \hat{k}_x\hat{\varepsilon}_z^{-1}\hat{k}_y h_x(z) \;=\; \omega^2 h_y(z), \tag{II.98}$$

$$\tag{II.99}$$

or in matrix form,

$$\left(\omega^2 I - \mathcal{K}\right) \begin{bmatrix} h_x(z) \\ h_y(z) \end{bmatrix} \;=\; -\mathrm{i}\,\omega \begin{bmatrix} -e_y'(z) \\ e_x'(z) \end{bmatrix}, \tag{II.100}$$

where I is the identity matrix of proper dimension, and where $\mathcal{K}$ is:

$$\mathcal{K} = \begin{bmatrix} \hat{k}_y\hat{\varepsilon}_z^{-1}\hat{k}_y & -\hat{k}_y\hat{\varepsilon}_z^{-1}\hat{k}_x \\ -\hat{k}_x\hat{\varepsilon}_z^{-1}\hat{k}_y & \hat{k}_x\hat{\varepsilon}_z^{-1}\hat{k}_x \end{bmatrix}. \tag{II.101}$$

Similarly, eliminating the z components with the sixth equation of II.95, the first and second equation of II.95 become:

$$\mathrm{i}\,\omega h_x'(z) \;+\; \hat{k}_x\hat{k}_x e_y(z) \;-\; \hat{k}_x\hat{k}_y e_x(z) \;=\; \omega^2 d_y(z), \tag{II.102}$$

$$-\,\hat{k}_y\hat{k}_x e_y(z) \;+\; \hat{k}_y\hat{k}_y e_x(z) \;-\; \mathrm{i}\,\omega h_y'(z) \;=\; \omega^2 d_x(z), \tag{II.103}$$

which can be written as:

$$\left(\omega^2\mathcal{E} - K\right) \begin{bmatrix} -e_y(z) \\ e_x(z) \end{bmatrix} \;=\; -\mathrm{i}\,\omega \begin{bmatrix} h_x'(z) \\ h_y'(z) \end{bmatrix}, \tag{II.104}$$

where

$$K = \begin{bmatrix} \hat{k}_x\hat{k}_x & \hat{k}_x\hat{k}_y \\ \hat{k}_y\hat{k}_x & \hat{k}_y\hat{k}_y \end{bmatrix}. \tag{II.105}$$

Finally, by applying the Fourier transform, the original Maxwell's equations of Eq. II.88 are reduced to Eq. II.100 and II.104.

II.4.3.5 Layer eigenmodes

The main idea behind the RCWA is to expand the fields within a layer into eigenmodes which have a simple $\exp(\mathrm{i}\, qz)$ dependence for some complex number q. In this section, the eigenvalue problem associated with each layer is described, *i.e.* from Eq. II.100 and II.104, we wish to define an eigenvalue problem. The eigenvalues are also referred as modes, or eigenmodes.

We assume the form of the magnetic field eigenmode is:

$$\mathbf{H}(z) = \sum_{\mathbf{G}} \left[\phi_{\mathbf{G},x}\mathbf{x} + \phi_{\mathbf{G},y}\mathbf{y} - \frac{(k_x + G_x)\phi_{\mathbf{G},x} + (k_y + G_y)\phi_{\mathbf{G},y}}{q}\mathbf{z} \right] \exp\left(\mathrm{i}\,(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r} + \mathrm{i}\, qz\right), \tag{II.106}$$

where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are the Cartesian unit vectors and $\phi_{\mathbf{G},x}$ and $\phi_{\mathbf{G},y}$ are expansion coefficients. These coefficients may be written as vectors $\phi_x = [\phi_{\mathbf{G}_1,x}, \phi_{\mathbf{G}_2,x}, \ldots]^T$, and analogously for ϕ_y. We then have:

$$h(z) = \left[\phi_x\mathbf{x} + \phi_y\mathbf{y} - \frac{\hat{k}_x\phi_x + \hat{k}_y\phi_y}{q}\mathbf{z} \right] \exp\left(\mathrm{i}\, qz\right), \tag{II.107}$$

where $h(z)$ is a column vector whose elements correspond to $\mathbf{G}$ vectors. With this, Eq. II.100 and II.104 become:

$$\begin{aligned} \left(\omega^2 I - \mathcal{K}\right) \begin{bmatrix} \phi_x \\ \phi_y \end{bmatrix} &= \omega q \begin{bmatrix} -e_y \\ e_x \end{bmatrix}, \\ \left(\omega^2 \mathcal{E} - K\right) \begin{bmatrix} -e_y \\ e_x \end{bmatrix} &= \omega q \begin{bmatrix} \phi_x \\ \phi_y \end{bmatrix}, \end{aligned}$$

where the z dependence is dropped on e_x and e_y to represent a fixed mode with $\exp(\mathrm{i}\, qz)$ variation. Eliminating the electric field and using the fact that $\mathcal{K}K = 0$, we finally obtain the eigenvalue problem:

$$\left(\mathcal{E}\left(\omega^2 - \mathcal{K}\right) - K\right)\boldsymbol{\phi} = \boldsymbol{\phi} q^2, \qquad \text{for} \quad \boldsymbol{\phi} = \begin{bmatrix} \phi_x \\ \phi_y \end{bmatrix}, \tag{II.108}$$

where q^2 is the diagonal matrix whose diagonal elements are the eigenvalues q_n^2. The columns of the square matrix $\boldsymbol{\phi}$ are $[\phi_{x,n}, \phi_{y,n}]^T$, the Fourier coefficients of the eigenmodes. Eq. II.108 is an asymmetric matrix eigenproblem, which may be complex if ε has an imaginary part. A useful orthogonality property, exploited in the next section, can be obtained by multiplying through by $\omega^2 - \mathcal{K}$ and using $\mathcal{K}K = 0$. Then Eq. II.108 becomes:

$$\left((\omega^2 - \mathcal{K})\mathcal{E}\left(\omega^2 - \mathcal{K}\right) - \omega^2 K\right)\boldsymbol{\phi} = (\omega^2 - \mathcal{K})\boldsymbol{\phi} q^2, \tag{II.109}$$

which has the form of a generalized symmetric eigenproblem. It follows that the eigenvectors $\boldsymbol{\phi}_n$, $\boldsymbol{\phi}_{n'}$, corresponding to the eigenvalues q_n, $q_{n'}$, satisfy the orthogonality relationship:

$$\phi_n^T(\omega^2 - \mathcal{K})\phi_{n'} = \delta_{nn'}. \tag{II.110}$$

The matrix $\omega^2 - \mathcal{K}$ being not positive definite, it is not easier to solve the generalized symmetric problem than the asymmetric problem.

II.4.3.6 Field recovery

In this section, the computation of the electric and magnetic field in real space, from eigenmode basis, is determined by solving Eq. II.109, is shown.

The transverse magnetic field, in layer i, is represented as:

$$\begin{bmatrix} h_x(z) \\ h_y(z) \end{bmatrix} = \sum_n \begin{bmatrix} \phi_{x,n} \\ \phi_{y,n} \end{bmatrix} \left(a_n \exp(\mathrm{i}\, q_n z) + b_n \exp(\mathrm{i}\, q_n (d_i - z)) \right), \tag{II.111}$$

where n indexes the eigenmodes, a_n is the coefficient of a forward propagating wave (towards negative z) at $z = z_i + d_i$, and b_n is the coefficient of a backward propagating wave at $z = z_i$. For $q = \pm\sqrt{q^2}$, there are two choices depending on the sign chosen. For numerical stability, the sign is chosen such that $\mathrm{Im} q \geq 0$, so that the defined coefficient is the maximum amplitude of each wave in the layer. Let now define a diagonal matrix operator $f(z)$ with entries:

$$f(z)_{nn} = \exp(\mathrm{i}\, q_n z), \tag{II.112}$$

which represents the modal phase accumulation operator. Let us also define transverse field component vectors in the Fourier basis:

$$h_t(z) = [\ h_x(z), h_y(z)]^T \quad \text{and} \quad e_t(z) = [-e_y(z), e_x(z)]^T, \tag{II.113}$$

and the diagonal matrix $\hat{q}$ such that $\hat{q}_{nn} = q_n$, as well as the mode amplitude vectors for forward and backward waves:

$$a = [a_1,\ a_2,\ \ldots\ ]^T \quad \text{and} \quad b = [b_1,\ b_2,\ \ldots\ ]^T. \tag{II.114}$$

With these definitions, the mode amplitudes are related to the physical fields with the following equation:

$$\begin{aligned} \begin{bmatrix} e_t(z) \\ h_t(z) \end{bmatrix} &= \begin{bmatrix} (\omega^2 - \mathcal{K})\, \boldsymbol{\phi} \hat{q}^{-1} & -(\omega^2 - \mathcal{K})\, \boldsymbol{\phi} \hat{q}^{-1} \\ \boldsymbol{\phi} & \boldsymbol{\phi} \end{bmatrix} \begin{bmatrix} f(z)a \\ f(d-z)b \end{bmatrix} \\ &= M \begin{bmatrix} f(z)a \\ f(d-z)b \end{bmatrix}. \end{aligned} \tag{II.115}$$

Using the orthogonality relationship of Eq. II.110, which in matrix form is $\boldsymbol{\phi}^T(\omega^2 - \mathcal{K})\boldsymbol{\phi} = 1$, it can be verified that the inverse of M is:

$$M^{-1} = \frac{1}{2} \begin{bmatrix} \hat{q}\boldsymbol{\phi}^T & \boldsymbol{\phi}^T(\omega^2 - \mathcal{K}) \\ -\hat{q}\boldsymbol{\phi}^T & \boldsymbol{\phi}^T(\omega^2 - \mathcal{K}) \end{bmatrix}, \tag{II.116}$$

Then, the procedure to retrieve the z fields component is described in [73].

The fields can now be computed at any position in the real space, given the forward and backward propagation coefficient, a_n and b_n.

The computation of the propagation coefficient, a_n and b_n, is performed through the use of the scattering matrix. The scattering matrix for the structure is constructed from the solutions to the q eigenvalue problem. The details of each of these steps is not repeated here since all implementation details are available in [59].

The RCWA method being presented, we focus, in the next section, on the actual implementation of our 2D and 3D RCWA solvers. During this thesis, RCWA is the only numerical that was actually implemented from scratch. Both DGTD and FDTD have been used though already available software, Lumerical for FDTD and DIOGENeS for DGTD.

II.5 RCWA implementation

In this section, the RCWA implementation of the 2D and 3D solver are described. This section also constitutes a minimal documentation for future users of this software.

II.5.1 Convergence inputs

The RCWA method relies on two approximations: the geometrical definition of the device, that is detailed in section II.5.2, and the plane wave truncation, detailed in the following paragraphs.

The proper choice of the plane wave truncation, namely the finite set of reciprocal vectors $\mathbf{G}$, of Eq. II.89, has been the object of numerous discussions in the literature [72]. In our implementation, we focus only on the standard squared plane waves truncation, even if the circular truncation, where all the $\mathbf{G}$ vectors within a circular region around the origin in reciprocal space are used, is promising.

In practice, the user defines a variable: half_{npw} so that the chosen wave vectors are the sets:

$$\begin{aligned} G_x &\in \{\, k_0 n, \mid \forall n \in [[-\text{half}_{npw}, \text{half}_{npw}]] \}\,, \\ G_y &\in \{\, k_0 n, \mid \forall n \in [[-\text{half}_{npw}, \text{half}_{npw}]] \}\,, \end{aligned} \tag{II.117}$$

in the 3D case and where $[[a, b]]$ denotes all the (positive or negative) integers between a and b. Geometrically, one could expect that Eq. II.117 is equivalent to a square truncation. In 2D, only the plane set of plane wave G_x is considered.

The eigenvalue problem II.108 is solved in using the eig [4] Matlab function. One can see directly that the size of the eigenvalue problem scales as N, the number of $\mathbf{G}$. Thus, in 3D, the storage requirements for an entire simulation scale as $O(MN^2)$, where M is the number of layers. And since the solution to the eigenvalue problems usually requires on the order of $O(N^3)$ operations (see [72]), the total run time of a single simulation scales as $O(MN^3)$.

Since the plane wave truncation, the time complexity and the memory requirements are described, we focus in the next section on the definition of the geometry for a simulation.

[4] `https://www.mathworks.com/help/matlab/ref/eig.html`

II.5.2 Geometry and visualization

As mentionned previously, the RCWA method supposes that the simulation region is partitioned in z-layers, whose permittivity is homogeneous in the z direction. This constitutes the main limitation of RCWA of simulating complex geometrical structures, since a staircasing effect must be introduced in the z dimension of such an object. And, as shown in section II.5.1, the numerical cost of a simulation is proportional to the number of layers. So layer discretization comes with a high numerical cost.

Given a layered structure in the z dimension, the structure, or more precisely the permittivity ε, can vary in each layer, in the x (respectively x and y) direction for a 2D structure (respectively a 3D structure). For the following, we assume that the period in the x (respectively y) dimension is a_x (resp. a_y), and that the structure is centered, *i.e* that the structure is defined on the interval $[-\frac{a_u}{2},\ \frac{a_u}{2}]$, where the u denotes either x or y.

In 2D, for each layer, the structure is defined by a 1D step function. For instance, given a layer with the following properties, assuming $a_x = 1$ μm:

- Minimum is at $z_{min} = 0$ μm.
- Maximum is at $z_{max} = 1$ μm.
- Permittivity jumps are at $x = -1$, $x = -0.5$, $x = 0.5$ and $x = 1$ μm with corresponding values:
  - $\varepsilon(x) = 10 + \mathrm{i} \qquad \forall x \in [-1,\ -0.5]$,
  - $\varepsilon(x) = 2 \qquad \forall x \in [-0.5,\ 0.5]$,
  - $\varepsilon(x) = 1 \qquad \forall x \in [0.5,\ 1]$.

Then the permittivity, and thus all the materials within these layers, can be directly described by the following 2D array:

$$A = \begin{bmatrix} -1 & 10+\mathrm{i} \\ -0.5 & 2 \\ 0.5 & 1 \\ 1 & \cdot \end{bmatrix}, \tag{II.118}$$

where for every line i, the permittivity value of $A(i,2)$ lies in the $[A(i,1),\ A(i+1,1)]$ interval. The $\cdot$ symbol in the last line denotes an unused value. For simplicity, we used abstract materials and we avoid mentioning the wavelength dependence of the permittivity in this example.

This representation of materials within a layer as a 1D step function will be used in the computation of the Fourier transform of the permittivity, in section II.5.3.

In 3D, the geometry description is more complex than in 2D. Since the permittivity in the z dimension is assumed constant in each layer, ε, in each layer, is a 2D function on the simulation basis square, noted D, defined as:

$$D := \left[-\frac{a_x}{2}, \frac{a_x}{2}\right] \times \left[-\frac{a_y}{2}, \frac{a_y}{2}\right]. \tag{II.119}$$

In our implementation, we rely on three 2D formal primitives: the polygons, the disk and the ellipsoid. We claim that combining these three primitives would allow the user to build the various structures needed for simulating CMOS sensors. In Figs. II.10 and II.11 are shown two example structures. Fig. II.10 illustrates the use of the three primitives functions:

- *Polygon2D* that adds a polygon to the structure,
- *Disk_2D* that adds a disk to the structure and
- *Ellipse_2D* that adds an ellipsoid.

On the left of Fig. II.10, a top view of the layer is shown. Fig. II.11 illustrates the usage of three helper functions that perform 2D usual transformation: a rotation with the *Rot2D* function, a translation with the *Translation2D* function and an axial symmetry with the *Symmetry2D* function. More advanced features, such as multiples including polygons, and examples of z layered structures or complex gratings, are illustrated in the Appendix B.

Two rules must be respected in the generation of such 2D layers:

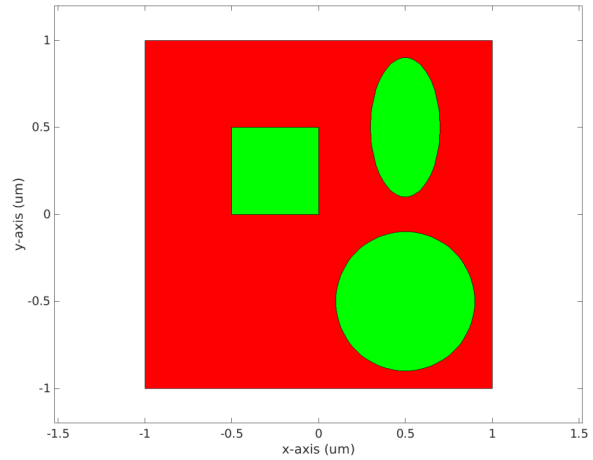
- Polygons must be strictly included in each other.
- Disk and ellipsoid must be empty, *i.e.* they must not include other polygons, other disks or other ellipsoids.

Following these rules allow us to automatically compute the 2D layer structure as a strict union of polygons and ellipsoids. This technical feature is not detailed here, but it is of high importance for the computation of the Fourier transform of such a 2D layer, that is described in section II.5.3. Basically the solver takes as input a list of polygons that can be strictly included in each other, and provides, for the Fourier computation, a list of polygons in strict union that can have holes. To achieve such geometrical operations, we extensively use the matlab polyshape object[5]. A check is then performed on the disks and ellipsoids to ensure their boundaries do not cross another object (polygon, disk or ellipsoid) boundary.

Finally, we must notice that a disk is a particular type of ellipsoid. We distinguished them here only for convenience.

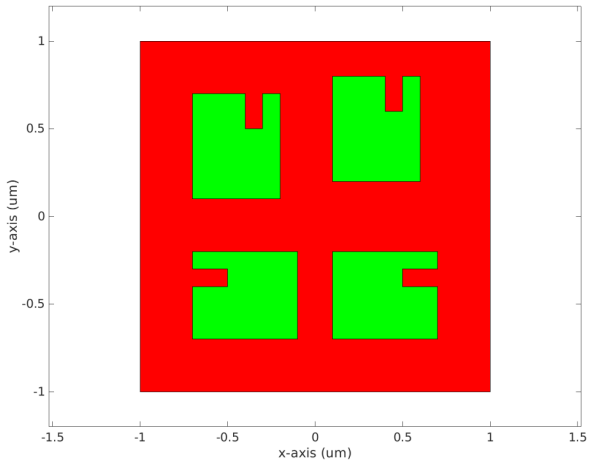
The geometry module of our RCWA solver being described, we focus in the next section on the computation of the Fourier transform of the permittivity in each layer.

[5] `https://www.mathworks.com/help/matlab/ref/polyshape.html`




```
Array_of_circle = [];
Array_of_Polygon = [];

% [x , y] Polygone corrdinates.
p_coor = [-0.5,        0;
             0,        0;
             0,      0.5;
          -0.5,      0.5];

Array_of_Polygon = [ Polygon2D(p_coor,    ...
                               epsSiO2)];

% Disk
D_radius =  0.4;
D_center = [0.5, -0.5];
Array_of_circle =  [ Disk_2D(D_radius,    ...
                             D_center,    ...
                             epsSiO2)];

% Ellipsoid
E_X_Ray = 0.4;
E_Y_Ray = 0.2;
E_center = [0.5, 0.5];
E_theta = pi/2;
Array_of_circle =  [Array_of_circle,      ...
                    Ellipse_2D(E_X_Ray,   ...
                               E_Y_Ray,   ...
                               E_center,  ...
                               E_theta,   ...
                               epsSiO2)   ...
                   ];

% Creating layer
Layers(end + 1) = [layer_2D(d,                    ...
                            eps_square,           ...
                            Array_of_Polygon,     ...
                            Array_of_circle,      ...
                            current_z-d,          ...
                            current_z)            ...
                   ];
current_z = current_z - d;
```

Figure II.10: Illustrating the generation of a 2D layer for the 3D RCWA solver. An ellipse, a disk and a square of SiO_2 (green) are added in a substrate of Si (red).




```
Array_of_circle = [];
Array_of_Polygon = [];

% [x , y] Polygones corrdinates.
p1 = [  -0.7,  0.1;
        -0.2,  0.1;
        -0.2,  0.7;
        -0.3,  0.7;
        -0.3,  0.5;
        -0.4,  0.5;
        -0.4,  0.7;
        -0.7,  0.7];
p2 = Rot2D(p1_coor, pi/2, [0, 0]);
p3 = Translation2D(p1_coor, [0.8, 0.1]);
p4 = Symmetry2D(p1_coor, [-1, -1], [1, 1]);

% Adding 4 polygones
Array_of_Polygon = [ Polygon2D(p1, epsSiO2), ...
                     Polygon2D(p2, epsSiO2), ...
                     Polygon2D(p3, epsSiO2), ...
                     Polygon2D(p4, epsSiO2)];

% Creating layer
Layers = [layer_2D(d,                     ...
                   eps_square,            ...
                   Array_of_Polygon,      ...
                   Array_of_circle,       ...
                   current_z-d,           ...
                   current_z)];
current_z = current_z - d;
```

Figure II.11: Illustrating the generation of a 2D layer for the 3D RCWA solver. A complex polygon is added (top left), as well as its translation (top right), rotation (bottom left) and axial symmetry (bottom right).

II.5.3 Fourier transform computation

As explained in section II.4.3, the Fourier computation of the permittivity function $\varepsilon_{\mathbf{G}}$ must be computed for each layer. Recalling Eq II.92, We have:

$$\varepsilon_{\mathbf{G}} \;=\; \frac{1}{|L_r|}\int_{cell} \varepsilon(\mathbf{r}) \exp\left(-\mathrm{i}\ \mathbf{G}\cdot\mathbf{r}\right) \mathrm{d}\mathbf{r}.$$

In a 2D structure, where each layer is of one dimension, the analytical computation of $\varepsilon_{\mathbf{G}}$ is equivalent to computing the Fourier transform of a 1D step function (see section II.5.2). There are no specific difficulties, and the formula is available in appendix C. Fast Fourier Transform [74] (FFT) can also be used. Since this is only a 1D Fourier transform, the time difference between analytical and FFT is negligible.

In 3D structures, the computation of $\varepsilon_{\mathbf{G}}$ is more subtle. Similarly to the 2D case, an FFT can be used, which would introduce a x-y mesh and so geometrical errors. The FFT is available, and preferred when the number of primitives in the layer is high (≥ 50), leading to a high computational cost of the analytical Fourier transform.

For the analytical computation of $\varepsilon_{\mathbf{G}}$, we first focus on the Fourier transform of the three primitives used: polygons, disk and ellipsoid. The Fourier computation of a polygon with holes is known and available in [75] or [76]. For convenience we provided the proof and our formula in appendix C. The computational cost of such formula scales linearly according to the number of dots on the polygons boundary. Then, the Fourier transform of a disk is the well-known Bessel function (see appendix C), this formula allows us to avoid approximating disks as polygons. Finally, the Fourier computation of an ellipsoid is performed with the Bessel function and the use of a linear transformation of the plane, to retrieve a disk from the original ellipsoid, and with the property II.13 of the Fourier transform.

Since the Fourier transform of all the three primitives is known, the problem of computing $\varepsilon_{\mathbf{G}}$ for each 2D layer consists in combining these primitives Fourier transform. This is where our transformation of multiple strictly included polygons into polygons in strict union comes into play (see section II.5.2). From a set of polygons in strict union, which can have holes, a simple loop is needed to compute the Fourier transform of the permittivity of the 2D layer. Disk and ellipsoid are then added simply with their Fourier transformed, scaled with the permittivity value of the polygon that contains them (details are available in appendix C).

To test our implementation of the analytical Fourier transform, we used the inverse Fourier transform, in order to compare ε and $\mathcal{F}^{-1}\left(\mathcal{F}_A\left(\varepsilon\right)\right)$. Such comparisons allow us to test both the primitive Fourier transform function, as well as our transformation of polygons strictly included, into polygons in strict union with holes. In Fig. II.12, such a test for the structure generated by table II.10 is shown.

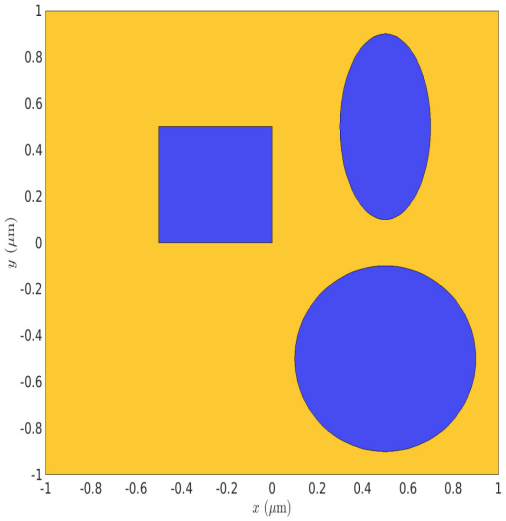
The Fourier transform computation being described, the next section illustrates the electric and magnetic field computation in our RCWA solver.

II.5.4 Field computation

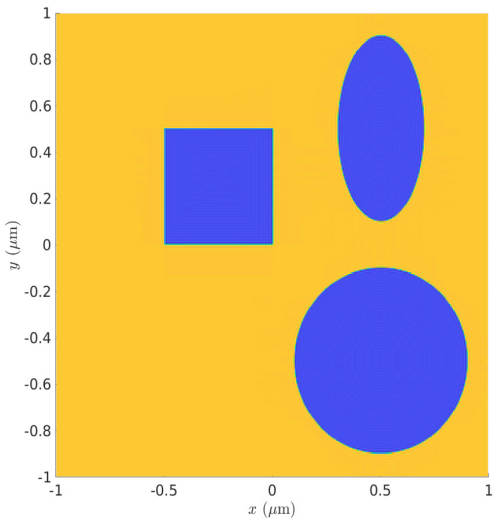
To keep the following description concise, we choose to not provide details of the fields computation of our RCWA software. Instead, one of the tests for field computation is provided. The structure is a simple 3D elliptical grating, shown in Fig. II.13. This structure was simulated with both our RCWA software and Lumerical[6]. Each field component, real and imaginary parts, are shown in Fig. II.14 and II.15. The exact juxtaposition of field values from our software and Lumerical confirms the validity of our implementation.

The RCWA solver implementation being described, the next section focuses on the benchmark, on structure of increasing complexity, of the three numerical methods presented previously.

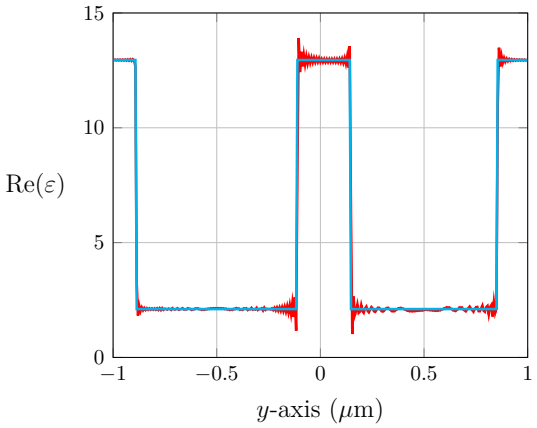
[6] `https://www.ansys.com/products/photonics`



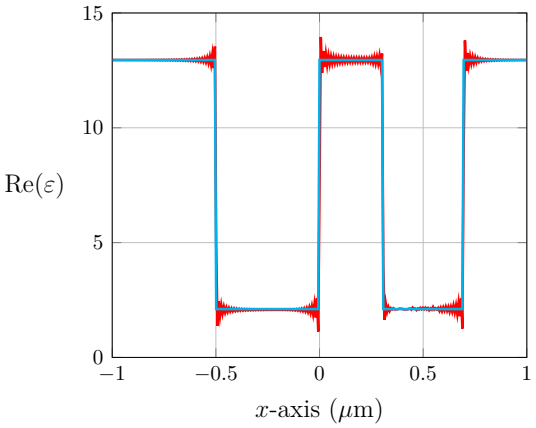
(a) Exact $\mathrm{Re}(\varepsilon)$.



(b) Inverse Fourier transform.



(c) Slice at $x = 0.4$ μm.



(d) Slice at $y = 0.4$ μm.

Figure II.12: Test of the analytical Fourier transform by comparing the real part of the exact permittivity value, ε, (shown in Fig. II.12a) and the real part of $\mathcal{F}^{-1}\left(\mathcal{F}_A\left(\varepsilon\right)\right)$ (shown in Fig II.12b). Figures on the second row are 1D cut (at $x = 0.4$ μm for Fig. II.12c and at $y = 0.4$ μm for Fig. II.12d) of the two above functions: the blue curve is the exact permittivity and the red curve is the result of $\mathcal{F}^{-1}\left(\mathcal{F}_A\left(\varepsilon\right)\right)$. One clearly sees the well-known Gibbs phenomenon, or the L^2 convergence of the Fourier transform, with the oscillation of the red curves.

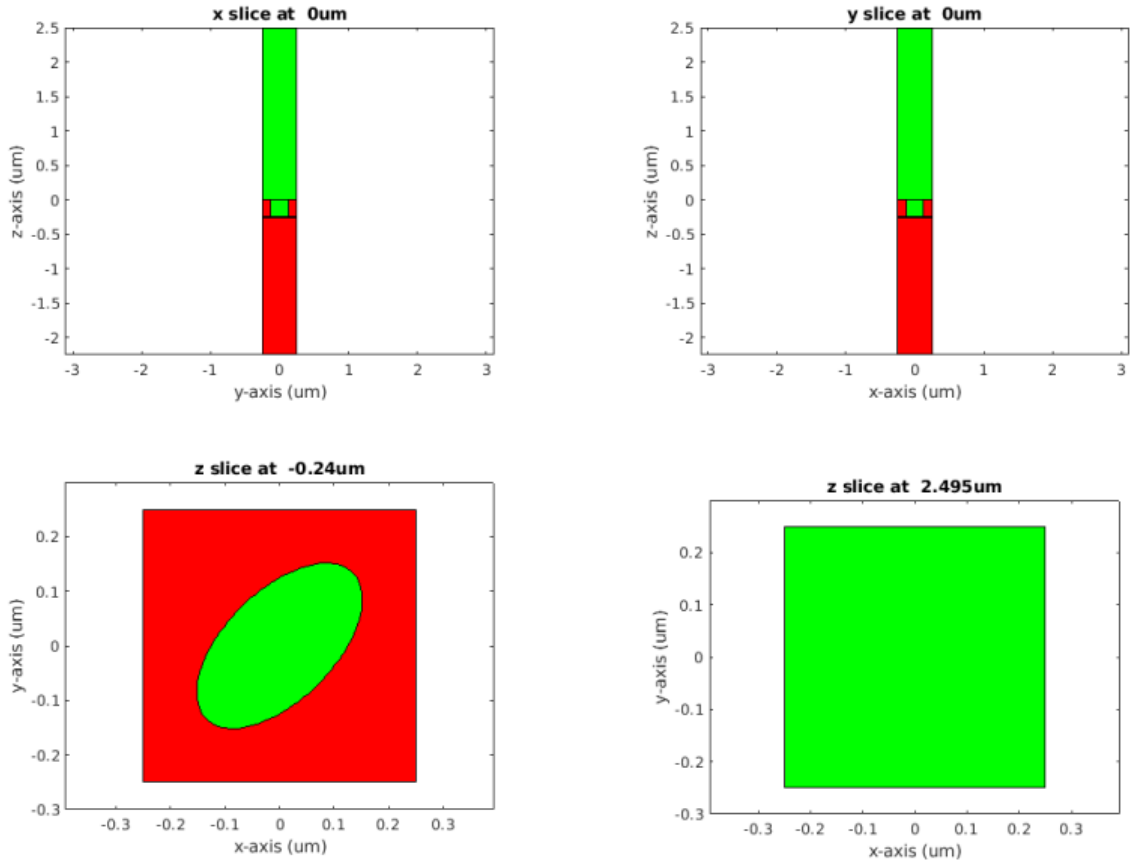
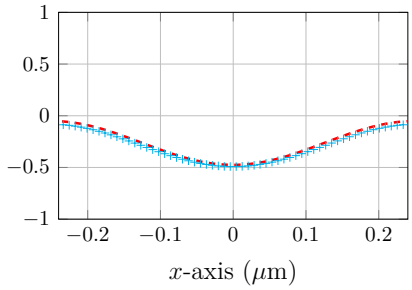
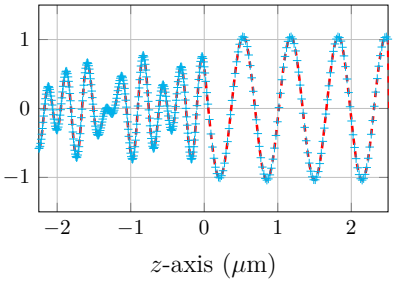


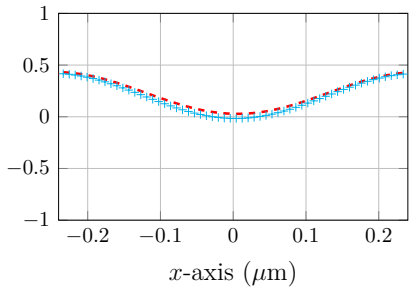
Figure II.13: Example structure of an ellipsoid grating. Green material is SiO_2 and red material is Si. The electric and magnetic fields component shown in Fig. II.14 and II.15 are computed on this structure. This structure is analogous to the simple silicon slab described in section II.6.1. The ellipsoid has an angle of $\frac{\pi}{4}$, a x radius of 0.2 and a y radius of 0.1.



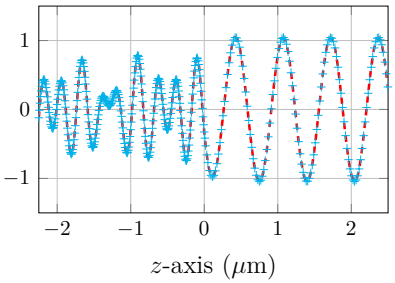
(a) Re (E_y)



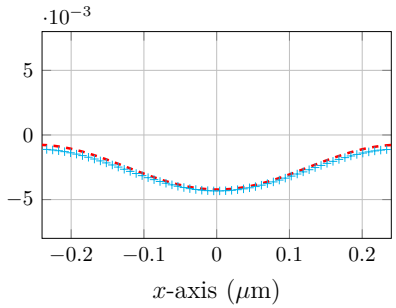
(b) Re (E_y)



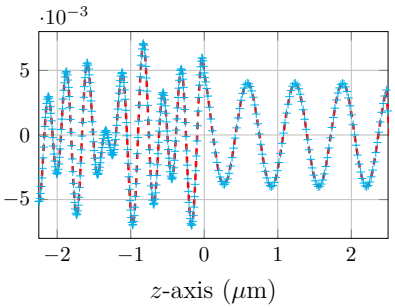
(c) Im (E_y)



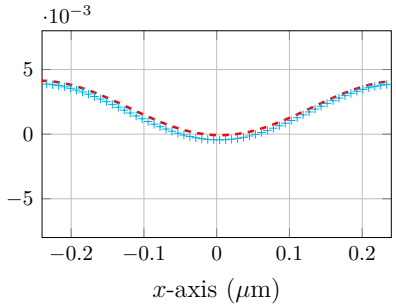
(d) Im (E_y)



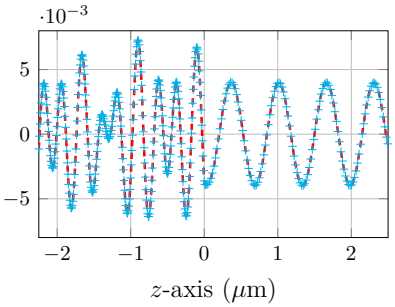
(e) Re (H_x)



(f) Re (H_x)

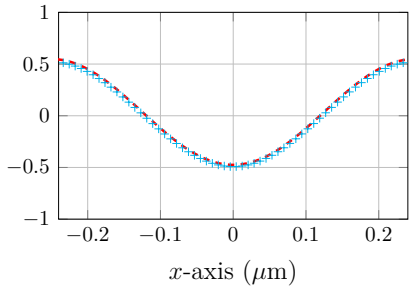


(g) Im (H_x)

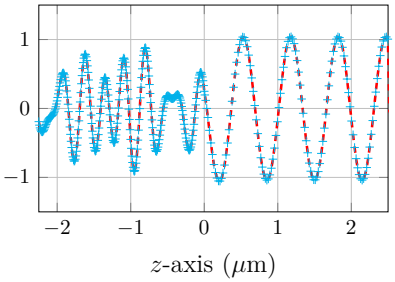


(h) Im (H_x)

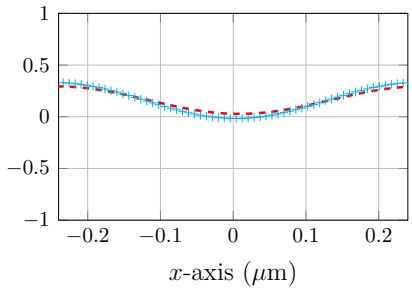
Figure II.14: Comparison of electric and magnetic fields components, at $\lambda = 940$ nm, computed with both our in-house RCWA solver and the Lumerical solver, on the structure desbribed by Fig. II.13. Dashed lines are RCWA results and blue crosses are FDTD results. 1D slices are extracted from the 3D domain: left column plots are at fixed $y = 0$ and $z = -1.968$ μm, right column plots at$x = -0.176$ and $y = 0$ μm. TE polarization is considered.



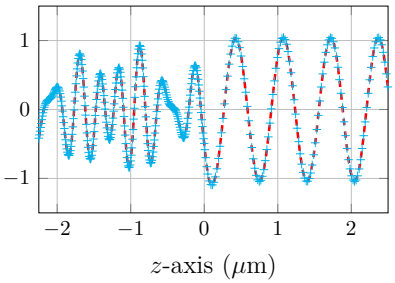
(a) Re (E_x)



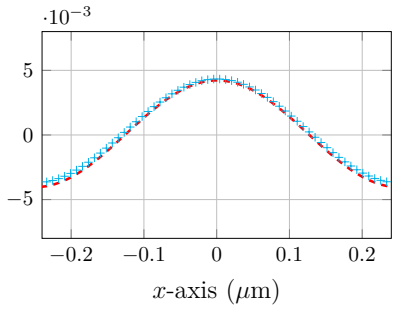
(b) Re (E_x)



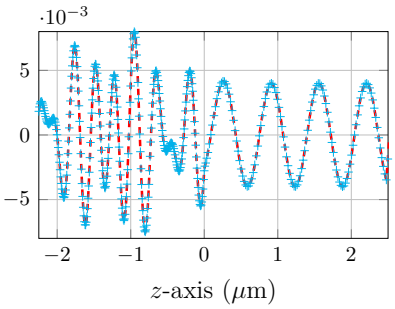
(c) Im (E_x)



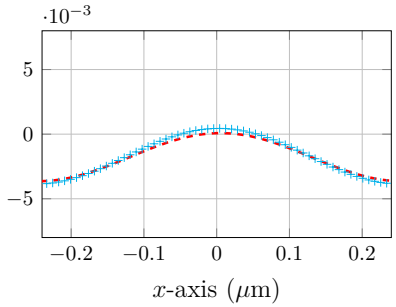
(d) Im (E_x)



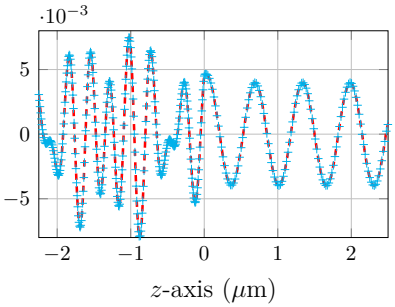
(e) Re (H_y)



(f) Re (H_y)



(g) Im (H_y)



(h) Im (H_y)

Figure II.15: Comparison of electric and magnetic fields components, at $\lambda = 940$ nm, computed with both our in-house RCWA solver and the Lumerical solver, on the structure desbribed by Fig. II.13. Dashed lines are RCWA results and blue crosses are FDTD results. 1D slices are extracted from the 3D domain: left column plots are at fixed $y = 0$ and $z = -1.968$ μm, right column plots at $x = -0.176$ and $y = 0$ μm. TM polarization is considered.

II.6 Benchmark on various structures

In this section, the numerical study of CMOS pixels is performed, as well as a benchmark of the three numerical methods presented in section II.4. First, the three numerical methods are compared in section II.6.1 on a simple silicon slab, and then a pixel-like geometry, with and without nanostructuration, is considered, comparing DGTD and FDTD methods in section II.6.2, and FDTD and RCWA methods in section II.6.3.

For the DGTD method, the DIOGENeS[7] sofware was used. For the FDTD method, Lumerical[8] was used and for the RCWA numerical method, our in-house Matlab solver, briefly described in section II.5, was used.

These benchmarks will provide useful informations on the pros and cons of each method for the specific geometry of CMOS pixels.

II.6.1 Simple structures

Three simple 3D cases are first considered that can be seen as a simplification of a nanostructured pixel, where both the DTI and the bottom metallic reflector are not taken into account. Thus they consist of a simple Oxide Silicon interface, which is either planar or patterned.

The simplicity of the geometry helps us to *a priori* guarantee the convergence to identical results. So these structures are well-suited for a first benchmark and comparison of the numerical methods.

II.6.1.1 Geometry

In this section, the exact dimensions of the three simple cases considered are described, as well as the material permittivity used. These cases are simply an SiO_2-Si interface, which is either planar, or structured with a square grating, or a pyramidal grating. Each case is named according to the interface type: planar, square or pyramidal.

On Fig. II.16, the three geometries are visualized. They share a common simulation basis, which is centered in the x and y-axis, of size 0.5 μm in both axis. The total size in z is 5 μm. Periodic boundaries are applied in x and y, while PML are used in the z direction. Light is coming from top, on the wavelength range [850, 1250] nm. The figures of merit are both the reflection, measured on top of structure, and the transmission, measured above the SiO_2-Si interface (see Fig. II.16). The three cases are distinguished according to the SiO_2-Si interface: the planar case is not structured, while the rectangular one is patterned with a centered square of size 250 nm, and the pyramidal case is patterned with a centered pyramid of depth 250 nm and basis square of 350 nm. The permittivity of Silicon and SiO_2 are obtained from Palik [53].

The geometry of the considered structure is now clarified. Next, we focus on the comparison of the absorption spectrum obtained with either FDTD, DGTD, or RCWA numerical methods.

[7] `https://diogenes.inria.fr/`

[8] `https://www.ansys.com/products/photonics`

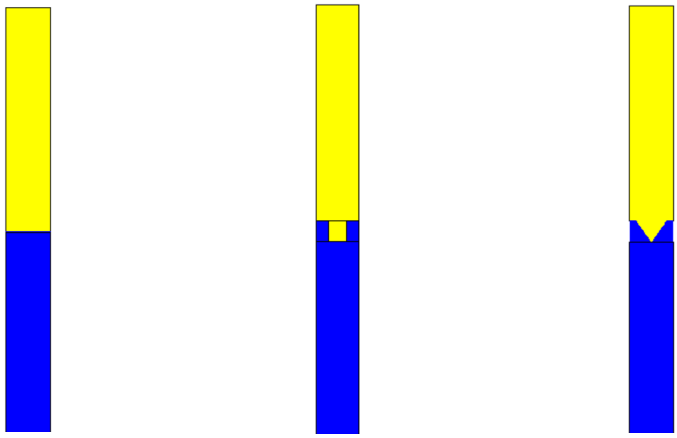


Figure II.16: Illustration of the simple structures definition: left figure is planar case, second is square case and on the right is the pyramidal case. Yellow material is SiO_2, and blue material is Silicon.

II.6.1.2 Results

In this section, the numerical results of the three simple Silicon slab structures, for the three numerical methods, RCWA, FDTD and DGTD, are shown. To keep this work concise, the convergence study is not shown here. For the DGTD method, a mesh with 22659 cells and a P_2 interpolation degree are used. For the FDTD methods, a meshfactor (see section II.4.1) of 22 is used. For both time-domain solvers, the energy threshold is set to 0.1%. The energy threshold is a criteria that limits the physical simulation time. It represents the fraction of the remaining energy inside the domain, to the energy introduced into the domain. For the RCWA, a plane wave truncation of 41, and an approximation of the pyramid in 40 layers, are used.

In Fig. II.17 (respectively II.18), the resulting reflection (respectively transmission) spectrum, for all three structures, on all three numerical methods, for both TE and TM polarizations, are shown. The planar structure being equivalent in TE and TM, only one plot is provided. The superposition of the curves confirms the validity of the implementation of each software used and our ability to use effectively these softwares. The only small discrepancy can be seen on the reflection spectrum of the pyramidal case, and it is discussed in the next section II.6.1.3.

The corresponding simulation times are available in Table II.1. One must notice that the simulation time of the RCWA solver, being a frequency-domain one, cannot be strictly compared to the simulation time of the time-domain method, such as FDTD or DGTD. Indeed, to obtain the curve shown in Fig. II.17 for instance, 100 RCWA simulations were required, while a single time-domain simulation provides the full spectrum response. Comparing the simulation time of FDTD and DGTD, one must notice the speed of the finite difference method. Despite the expectancy, the pyramidal case, enhancing the well-known staircasing effect, is faster when simulated with the FDTD. This will also be discussed in the next section II.6.1.3. For the RCWA solver, the cost of approximating the inverted pyramid with 40 layers is clear (see Table II.1), leading to a high computational cost.

Case	Method	Wall time TE	Wall time TM
Planar	RCWA	<1 mn	-
-	DGTD-P_2	13 mn	-
-	FDTD	7 mn	-
Rectangular	RCWA	<1 mn	<1 mn
-	DGTD-P_2	8 mn	10 mn
-	FDTD	7 mn	7 mn
Pyramidal	RCWA	12 mn	12 mn
-	DGTD-P_2	8 mn	10 mn
-	FDTD	12 mn	14 mn

Table II.1: Simulation time for the three numerical methods, on the three simple cases described in section II.6.1. One clearly sees the cost of the z layer approximation for the RCWA, for the pyramidal case.

The main results being presented, a focus on the pyramidal case is shown in the next section.

II.6.1.3 Discussion on pyramidal case

The pyramidal case is singular for two reasons:

- Firstly, it enhances the staircasing effects, imposing to FDTD a finer mesh, and to RCWA a multiple z-layers structure, leading to higher computational cost (see Table II.1).

- Secondly, it introduces a peak, at the bottom of the pyramid, that is respected by Delaunay meshes by introducing ill-shape tetrahedra (tetrahedra with a small angle at one of their vertices). These ill-shape tetrahedra introduce, through the CFL condition (see II.71) a smaller time step, leading to a higher computational cost for the DGTD method. In order to control this phenomenon, an apex, namely a truncation of the pyramid peak, is introduced in the geometry for DGTD simulations. Some examples of Delaunay meshes with a different apex are shown in Fig. II.19. The corresponding time-step are available in Table II.2, as well as the wall time for the DGTD-P_3 simulation using these meshes. One remarks easily the influence of the apex on the timestep, and thus on the total simulation time. The smaller the apex is, the higher the time step and the longer the simulations.

On Fig. II.20, the reflection and transmission spectra for the DGTD-P_3 simulation with various values of the apex are compared to the reference FDTD results. One can clearly notice the influence of the apex on the figure of merit.

The discrepancy of Fig. II.17 and II.18 are explained by the apex definition being different between the three numerical methods.

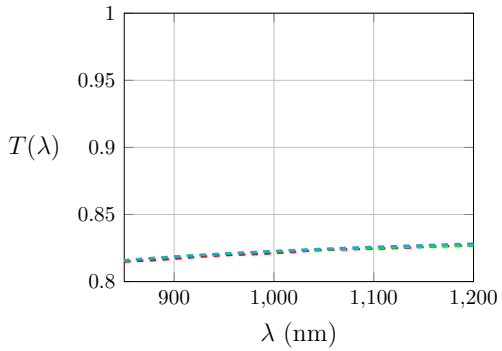
Apex size	Method	Δt (sec)	Wall time TE	Wall time TM
2 nm	DGTD-P_3	3.90×10^{-18}	1 h 18 mn	1 h 35 mn
8 nm	DGTD-P_3	1.34×10^{-17}	24 mn	34 mn
20 nm	DGTD-P_3	1.33×10^{-17}	27 mn	28 mn
12 nm	FDTD	5.73×10^{-18}	12 mn	14 mn
8 nm	DGTD-P_2	1.36×10^{-17}	8 mn	10 mn

Table II.2: Simulation time and time step (Δt) according to the apex definition (see Fig. II.19). All DGTD times refer to the 0.1% energy threshold. FDTD time corresponds to the reference FDTD results with a meshfactor of 22. The increase of Δt according to the apex is clear, leading to shorter simulation time.

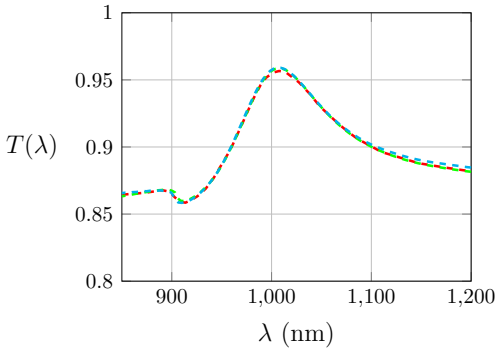
II.6.1.4 Conclusion on simple structures

In this section, our ability to converge to identical results, with the three optical solvers and for three simple structures, has been demonstrated. Comparing the time, the reference Lumerical software is faster on those three simple cases.

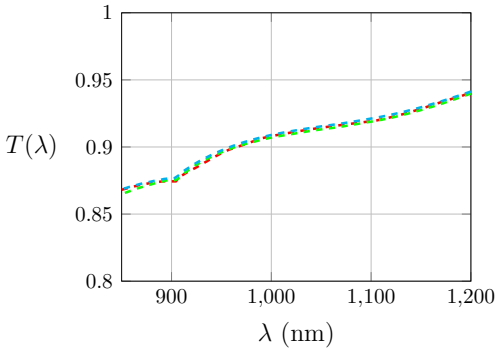
In this study, the importance of respecting the geometry, in particular for the pyramidal case, was emphasized, and we did not, on purpose, discuss the importance of using the identical permittivity value. The RCWA, being a frequency-domain solver, uses directly the exact value provided by literature, while FDTD and DGTD, being time-domain solvers, use a fit of the permittivity on the wavelength range of interest [850, 1200] nm. Those fits can induce a different spectrum response if not precise enough. The best way to avoid permittivity fit error is to compare the reflection and transmission for the simplest structure, namely the planar case. In our work, the exact agreement of the spectrum response for the planar case is confirming that the fits used for both FDTD and DGTD are precise enough. To keep the current work concise, these fits are not shown here.



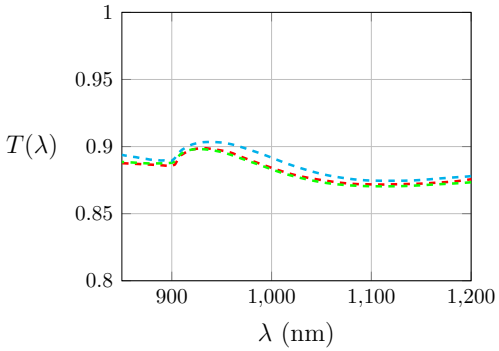
(a) Planar



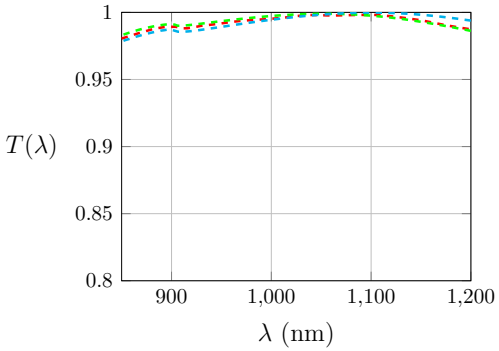
(b) Rectangular TE



(c) Rectangular TM

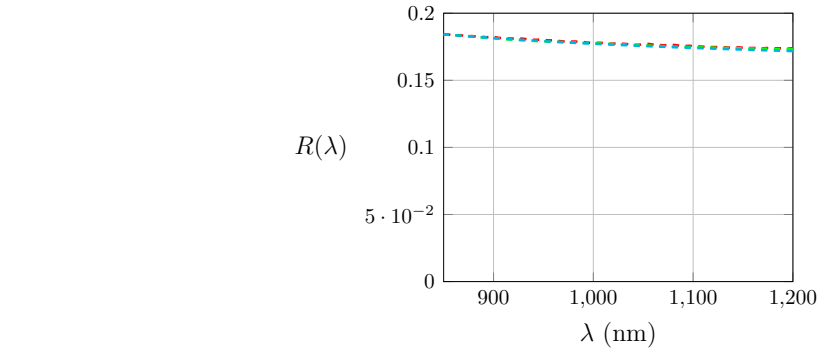


(d) Pyramidal TE

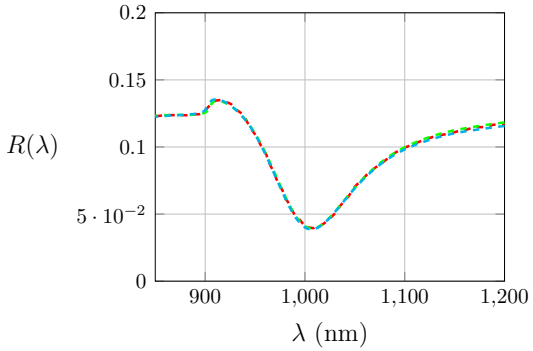


(e) Pyramidal TM

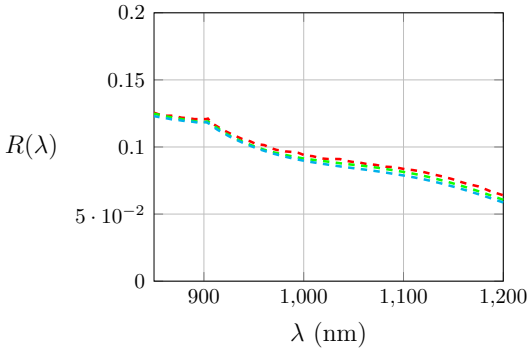
Figure II.17: Transmission spectrum for all three numerical methods, on all three simple cases. Dashed blue is DGTD, dashed green is FDTD and dashed red is RCWA. The y axis is zoomed on the interval $[0.8,\ 1]$. Results match perfectly, except for the pyramidal case, which is discussed in section II.6.1.3.



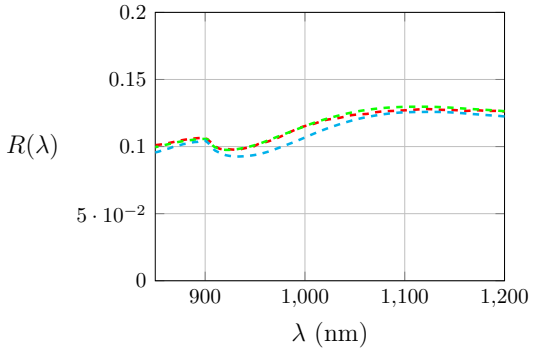
(a) Planar



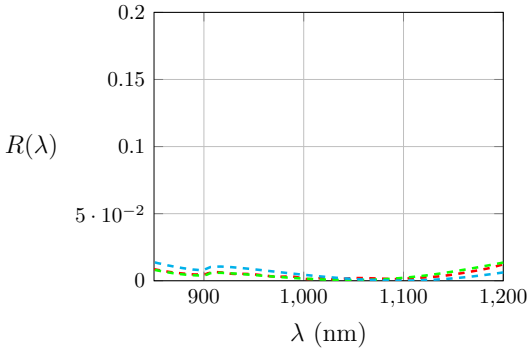
(b) Rectangular TE



(c) Rectangular TM



(d) Pyramidal TE



(e) Pyramidal TM

Figure II.18: Reflection spectrum for all three numerical methods, on all three simple cases. Dashed blue is DGTD, dashed green is FDTD and dashed red is RCWA. The y axis is zoomed on the interval $[0,\ 0.2]$. Results match perfectly, except for the pyramidal case, which is discussed in section II.6.1.3.

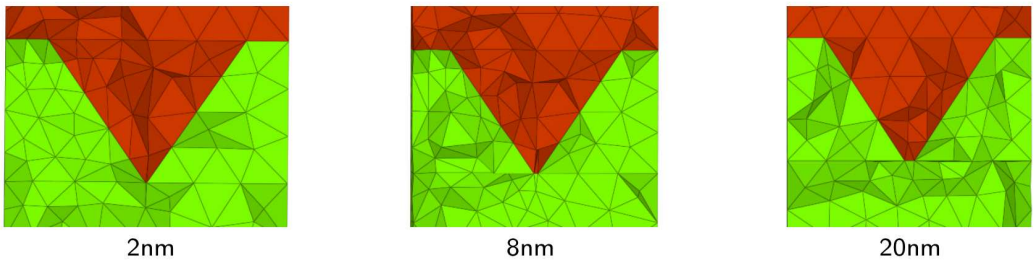


Figure II.19: Definition of the apex for the Delaunay mesh used with the DGTD method, of the pyramidal simple case described in section II.6.1.1. The pyramid head is truncated to avoid ill-shaped tetrahedra (tetrahedra with small angle). The FDTD, using a cartesian mesh, or the RCWA, approximating the z geometry variations into layers, implicitly introduces an apex whose size is relative to either the mesh cell size for FDTD, or the layers approximation for RCWA.

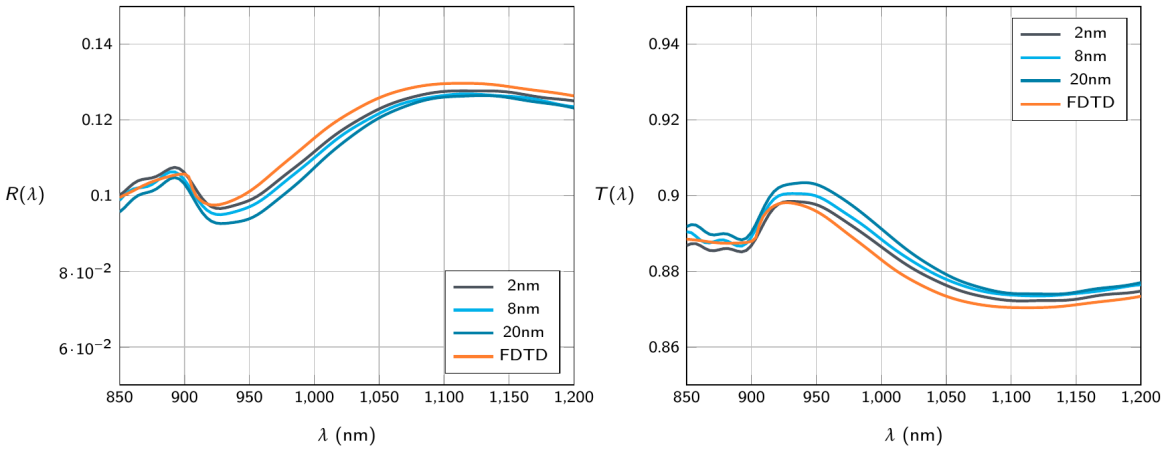


Figure II.20: Reflection (left) and transmission (right) spectra obtained with DGTD-P_3 for the pyramidal case (defined in section II.6.1.1), for various values of the apex sizes (see Fig. II.19). The lower the apex, the closer the DGTD result is to the reference FDTD results. In order to emphasize this effect, the y-axis is zoomed in: the differences observed are of order 10^{-3}.

II.6.2 FDTD and DGTD on pixels

This section is devoted to a numerical study of the model pixel geometries introduced in section II.6.2.1. Firstly, we consider an inner Si volume with a flat top surface and perform a detailed assessment of the DGTD method. Secondly, we consider a pixel with a nanostructuring of one-dimensional gratings of the top surface of the inner Si volume. In both cases, the DGTD results are compared with reference results provided by the FDTD method.

These pixel geometries can be seen as a simple silicon slab, where two complex geometrical elements have been added: firstly, Deep Isolation Trenches are added, as well as a Tungstene shield covering the exterior Silicon volume. Secondly, a bottom metallic reflector is added. These structures do not include lens, nor nanostructuration. Benchmarks on nanostructured pixels are performed in section II.6.3 and II.6.4.

II.6.2.1 Geometry

The geometrical characteristics of the model pixels are the following:

- Size in X: $\sim$6 μm
- Size in Y: $\sim$6 μm
- Size in Z: $\sim$5 μm + 0.7 μm (TF/SF air) + PML
- PBC in XoZ and YoZ boundaries
- PEC in XoY for origin in Z (Z = 0)

The exact dimensions are confidential and thus not specified. The Si, W, SiO_2 material permittivity used are taken from Palik [53].

The square pixel and the octogonal pixel are identical, except for the DTI shapes, as visible on Fig. II.21 and II.22.

II.6.2.2 FDTD results

The first set of results obtained with the FDTD method are shown in Fig. II.23 in the form of spectra of the volumic absorption in the internal Si volume. These results have been obtained in the following conditions:

- Mesh resolution: $\lambda/18$, full geometry
- PEC conditions at the bottom side
- Simulation size in grid points: 209 x 84 x 390
- Physical time: 2×10^{-12} sec (2000 fs)
- System with 10 processors
- Wall clock time: 13 h 16 mn (square pixel) and 16 h 15 mn (octogonal pixel)

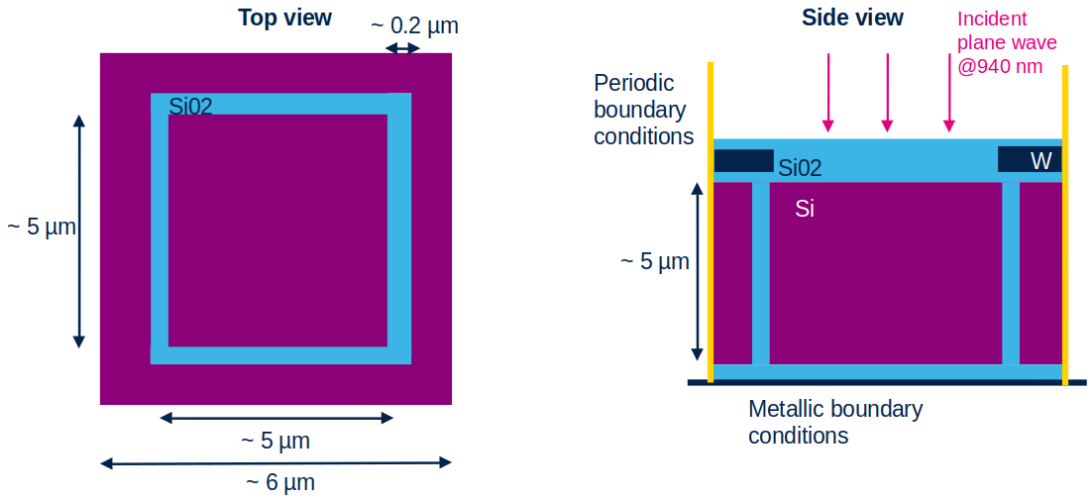


Figure II.21: Sketched of the model square pixel. Purple is Si material, blue is SiO_2, black is tungsten and white is Air. The exact dimensions of the simulated pixel are confidential. The sketch is not at scale and for illustration purposes only.

We shall make here two additional remarks: (1) the full geometry is considered, i.e., the symmetry of the problem is not exploited and, (2) the physical simulation time has been voluntarily set to a large value for ensuring the convergence in time. We note that the differences between the results for the two shapes, i.e., square versus octogonal pixel, are relatively minimal. This has motivated our choice to focus on the square pixel geometry in the sequel.

II.6.2.3 FDTD results for the square pixel

We now present the results of a numerical convergence in space assessment of the FDTD method for the square pixel geometry, see Fig. II.24. These results have been obtained in the following conditions:

- 1/4 of the full geometry with PEC/PMC conditions on the lateral sides
- Cu layer (200 nm) at the bottom side
- Physical time: 2×10^{-12} sec (2000 fs)
- System with 15 processors

Details of the characteristics of FDTD meshes, time steps and CPU times are summarized in Tab. II.3. Note that these simulations have been performed with 1/4 of the full geometry by taking into account the symmetry of the problem. Moreover, the PEC condition at the bottom boundary has been replaced by a layer of dispersive Cu. We can conclude from these results that a FDTD mesh with a characteristic length $\lambda/22$ yields a sufficiently converged result.

Finally, we select the FDTD mesh with a characteristic length $\lambda/22$ and perform simulations for different physical simulation times in order to assess the convergence in time, see Fig. II.25. These results have been obtained in the following conditions:

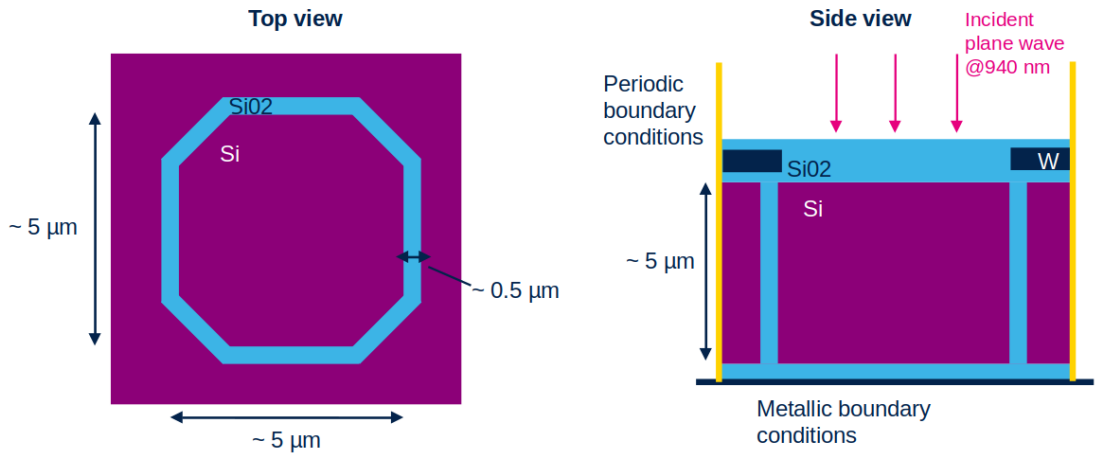


Figure II.22: Sketched of the model octogonal pixel. Purple is Si material, blue is SiO_2, black is tungsten and white is Air. The exact dimensions of the simulated pixel are confidential. The sketch is not at scale and for illustration purposes only.

Mesh resolution	Size in gridpoint	# time steps	Wall time	Δt (sec)
$\lambda/14$	54 x 88 x 81	137,108	57 mn	1.458×10^{-17}
$\lambda/18$	69 x 92 x 97	145,921	1 h 53 mn	1.371×10^{-17}
$\lambda/22$	85 x 99 x 114	158,858	2 h 57 mn	1.259×10^{-17}
$\lambda/26$	99 x 103 x 131	170,443	3 h 29 mn	1.173×10^{-17}
$\lambda/30$	114 x 107 x 147	182,900	6 h 01 mn	1.093×10^{-17}
$\lambda/34$	129 x 111 x 163	196,063	7 h 30 mn	1.020×10^{-17}

Table II.3: FDTD results with semi-infinite Cu layer at the bottom. Volumic absorption in the internal Si volume. Numerical convergence versus the average number of grid points per wavelength.

- 1/4 of the full geometry with PEC/PMC conditions on the lateral sides
- Cu layer (200 nm) at the bottom side

It is clear from these results that a physical simulation time between 500 fs and 700 fs is acceptable. In the sequel, we fix the physical simulation time to 600 fs.

Before moving to the presentation of results obtained with the DGTD method, we illustrate one of the issues faced with the structured (Cartesian) mesh used in the FDTD method. Indeed, even if a non-uniform discretization is possible with this type of mesh, ensuring a perfect alignment of mesh lines with material interfaces is a difficult task as depicted in Fig. II.26. Even if the mismatch between mesh lines and material interfaces can be minimized, we have observed that FDTD results are sensible to the mesh at material interfaces.

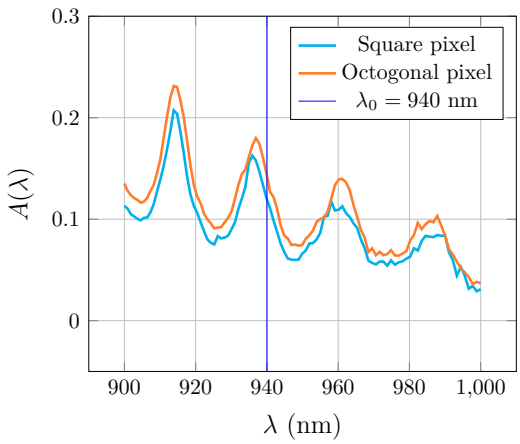


Figure II.23: FDTD results with PEC condition at the bottom boundary. Volumic absorption in the internal Si volume.

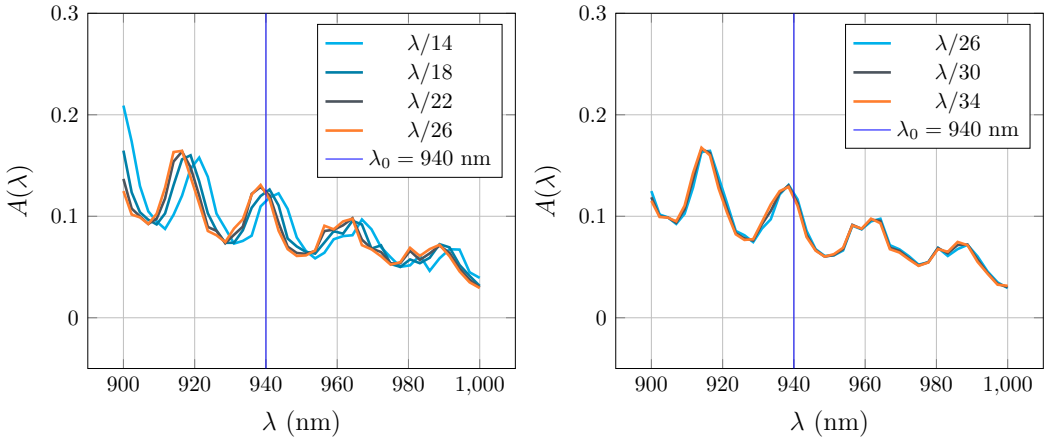


Figure II.24: FDTD results with semi-infinite Cu layer at the bottom. Volumic absorption in the internal Si volume. Numerical convergence versus the average number of grid points per wavelength.

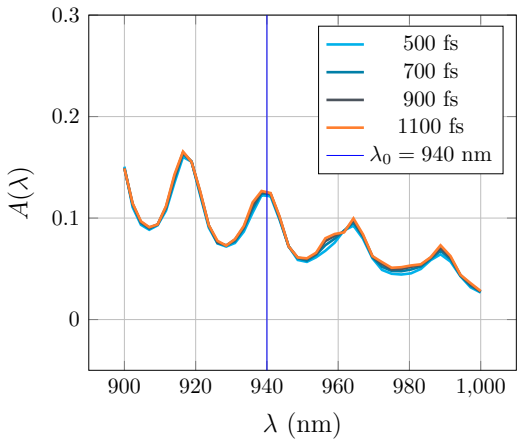
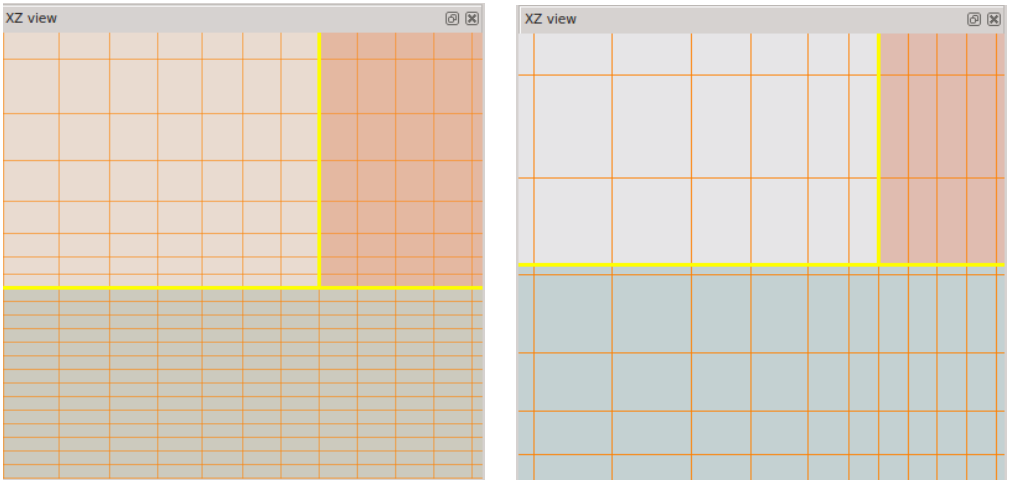


Figure II.25: FDTD results with semi-infinite Cu layer at the bottom. Volumic absorption in the internal Si volume. Numerical convergence versus the physical simulation time. Mesh with $\lambda/22$.



Conformal mesh Non-conformal mesh

Figure II.26: FDTD results with semi-infinite Cu layer at the bottom. Influence of the type of mesh at material interfaces.

Mesh	Cu layer (nm)	# cells	h_m (nm)	h_M (nm)	Ratio
M1	200.0	97,629	41.6	344.9	8.3
M2	200.0	175,876	38.5	336.4	8.7
M3	200.0	314.174	35.2	344.5	9.8

Table II.4: Characteristics of meshes for the DGTD simulations.

Mesh	$[h_1 , h_2]$ (nm)	$[h_1 , h_2]$ (nm)	$[h_1 , h_2]$ (nm)	$[h_1 , h_2]$ (nm)
M1	[41.6 , 117.5]	[117.5 , 193.3]	[193.3 , 269.1]	[269.1 , 344.9]
	5,247	54,212	37,078	1,090
M2	[38.5 , 113.0]	[113.0 , 187.5]	[187.5 , 261.9]	[261.9 , 336.4]
	29,591	122,608	23,051	626
M3	[35.2 , 112.6]	[112.6 , 190.0]	[190.0 , 267.4]	[267.4 , 344.5]
	195,596	98,914	19,176	487

Table II.5: Characteristics of meshes for the DGTD simulations.

II.6.2.4 Comparison between FDTD and DGTD

We now switch to a numerical convergence study with the DGTD method. We consider the following setting:

- 1/4 of the full geometry with PEC/PMC conditions on the lateral sides
- Top PML: 600 nm
- Bottom PML: 200 nm
- Dispersive Cu layer instead of PEC wall (bottom side)
- Physical time: 600 fs

Several unstructured tetrahedral meshes have been constructed whose characteristics are summarized in Tab. II.4. We recall that in the case of the DGTD method one has access to two parameters for assessing the convergence in space: the usual discretization parameter, which is here denoted by h, and the interpolation degree p, which is used for approximating the components of the electromagnetic field within each mesh cell (tetrahedron). In Tab. II.5, we give the minimum and maximum values of h for each mesh.

Results are presented in Figs. II.27 and II.28 and compared with the reference FDTD result. Overall, the selected combinations of mesh resolution and interpolation degree all yield results that are in line with the reference FDTD result. Finally, the convergence in time is depicted in Fig. II.29 for mesh M1 and using the DGTD-$\mathbb{P}_5$ method.

II.6.2.5 Conclusion on square pixel comparision between FDTD and DGTD

Finally, these comparisons between the FDTD and DGTD numerical methods on the flat square pixels demonstrate the ability to reproduce identical absorption spectrum, as

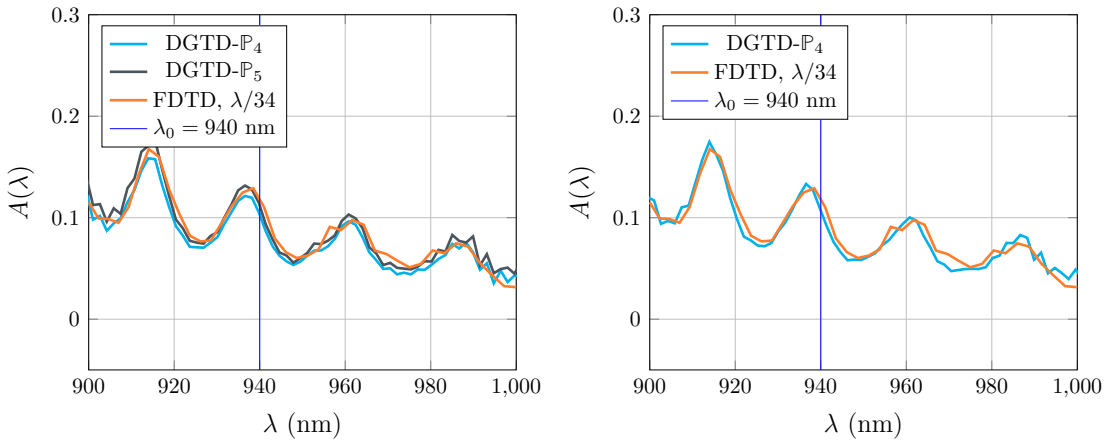


Figure II.27: DGTD results with semi-infinite Cu layer at the bottom. Volumic absorption in the internal Si volume. Left: mesh M1 - Right: mesh M2.

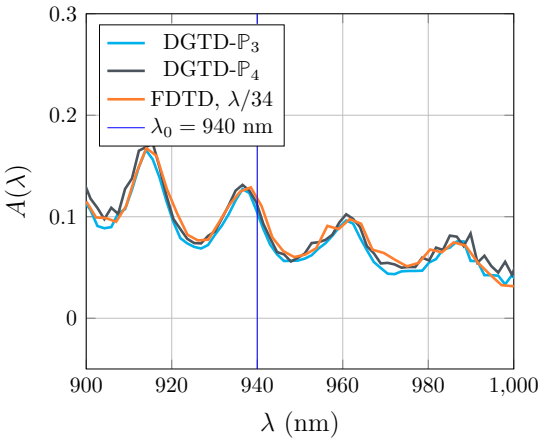


Figure II.28: DGTD results with semi-infinite Cu layer at the bottom. Volumic absorption in the internal Si volume. Mesh M3.

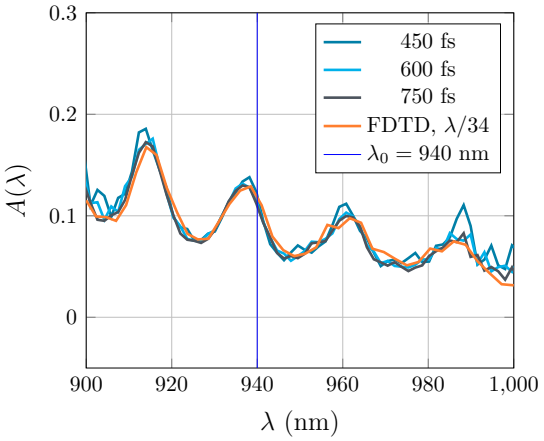


Figure II.29: DGTD results with semi-infinite Cu layer at the bottom. Volumic absorption in the internal Si volume. Numerical convergence versus the physical simulation time.

shown in Fig. II.28. Since this structure has a flat top surface, the light goes back and forth, fully reflected on the bottom metallic reflector and partially reflected on the top Si-SiO_2 and SiO_2-Air interfaces. These multiple reflections explain the four resonances that are visible in Fig. II.28. The physical phenomena is thus captured by both numerical methods. A more detailed discussion about simulation time is given in the final conclusion of this chapter.

II.6.3 FDTD and DGTD with a nanostructuring of the square pixel

II.6.3.1 Comparison between FDTD et DGTD

In fig. II.30 we summarize the results of a numerical convergence study with the FDTD method. In these plots, "Meshfactor" is the number of points per wavelength. Simulations have been performed on 1/4 of the full geometry with PEC/PMC conditions on the lateral sides. The maximum physical time is set to 1000 fs. Results obtained with the DGTD method are shown in Fig. II.31. Two meshes have been used: M1 with 174,637 cells (h_m = 32.4 nm and h_M = 284.6 nm) and M1 with 343,736 cells (h_m = 32.6 nm and h_M = 225.6 nm). Form these results, we can note the sensibility of the FDTD method to the alignment of the mesh with material interfaces. This alignment is improved as the mesh is refined but it is also clear that numerical convergence is hard to achieve. On the contrary, the DGTD method relies on a mesh that is perfectly positioned on material interfaces. As a consequence, DGTD results are less erratic and, for a given mesh, numerical convergence can be achieved by increasing the interpolation degree as demonstrated here with the results for mesh M1.

It must be noted that, for the square pixel with 1D grating, the absorption spectrum obtained with both DGTD and FDTD, do not match exactly, as visible in Fig. II.31.

II.6.3.2 Parametric study

The objective of this section is to conduct a preliminary assessment of the impact of nanostructuring on light absorption in the active Si layer. The results of this study will serve as a guide for the numerical optimization study, which will be realized in the chapter III. Here we only consider a nanostructuring of the top surface of the active Si layer based on a 1D grating with a rectangular sections. Several partial views of geometrical models (tetrahedral meshes) for some selected grating parameters are shown in Fig. II.32.

The first configuration that we consider is characterized by the following setting:

- Configuration C1
  - 1/4 of the full geometry with PEC/PMC conditions on the lateral sides
  - Top PML: 600 nm
  - Bottom PML: 200 nm
  - Dispersive Cu layer instead of PEC wall (bottom side)
  - Grating width: 600 nm
  - Minimum space between rods: 1000 nm
  - Physical time: 600 fs
  - DGTD-$\mathbb{P}_4$ method

Three tetrahedral meshes have been constructed whose characteristics are summarized in Tab. II.6. They correspond to three choices of the grating depth. Results are depicted in Fig. II.33. The plot also includes the FDTD result for the flat structure. For this

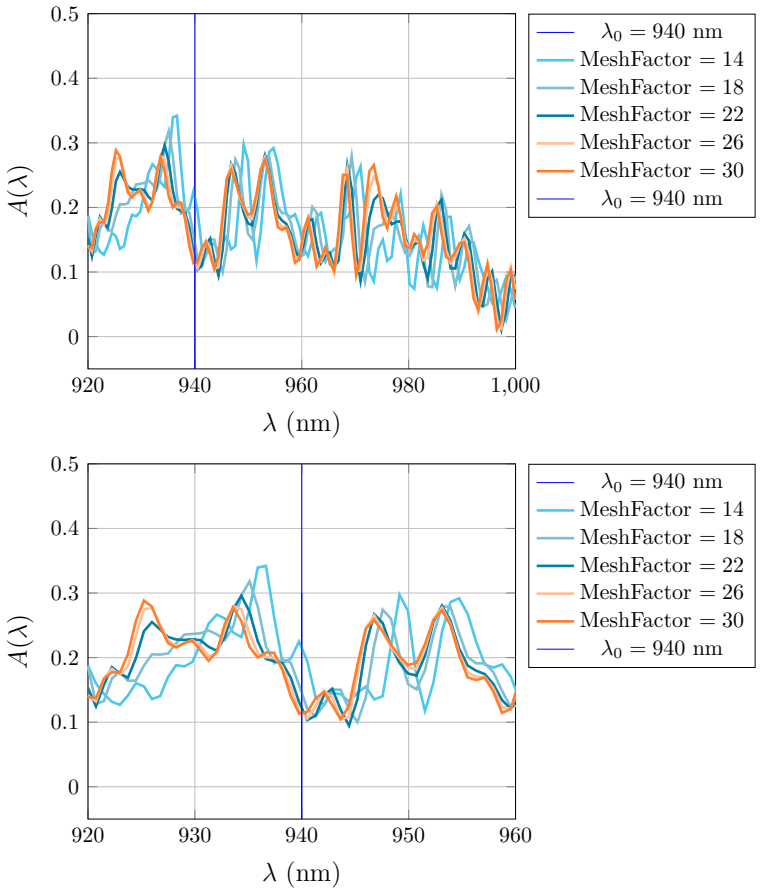


Figure II.30: Square pixel with a 1D grating. Volumic absorption in the internal Si volume. FDTD results (bottom figure is a zoom).

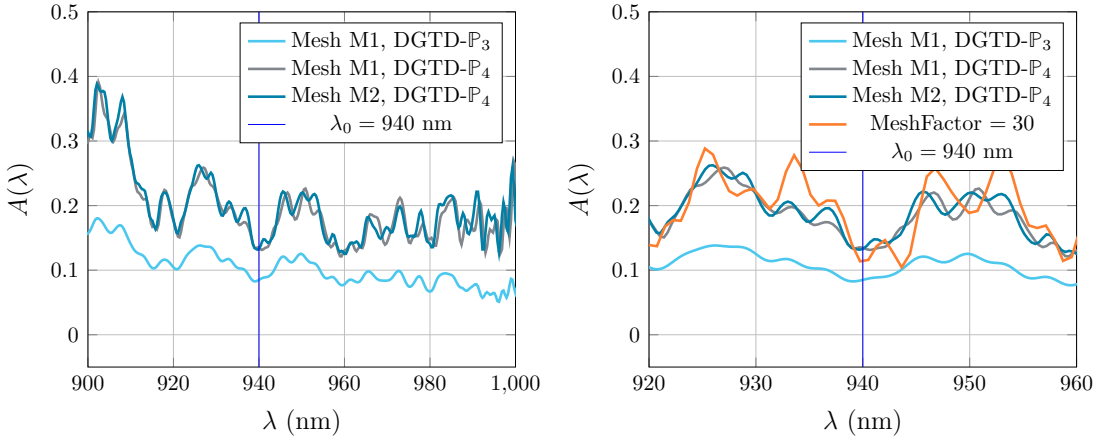


Figure II.31: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating. Volumic absorption in the internal Si volume. Numerical convergence versus mesh and interpolation order.

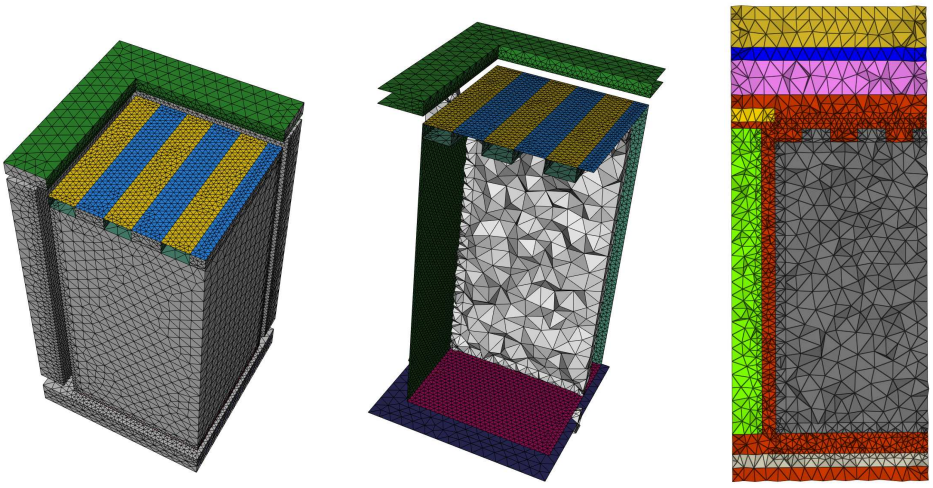


Figure II.32: Views of meshes of the square pixel with a 1D grating. 1D grating of active Si volume along O_x using rods with rectangular section.

Mesh	Depth (nm)	# cells	h_m (nm)	h_M (nm)	Ratio
M1	15.0	99,810	15.0	349.4	23.3
M2	30.0	99,721	30.0	368.7	12.3
M3	60.0	99,915	42.2	368.4	8.8

Table II.6: Characteristics of meshes of the square pixel with a 1D grating. Configuration C1.

first configuration, the selected grating parameters do not notably affect the absorption profile as compared to the reference FDTD result. The second configuration that we consider is characterized by the following setting:

- Configuration C2
  - 1/4 of the full geometry with PEC/PMC conditions on the lateral sides
  - Top PML: 600 nm
  - Bottom PML: 200 nm
  - Dispersive Cu layer instead of PEC wall (bottom side)
  - Width grating: 400 nm
  - Minimum space between rods: 800 nm
  - Physical time: 600 fs
  - DGTD-$\mathbb{P}_4$ method

Three tetrahedral meshes have been constructed whose characteristics are summarized in Tab. II.7. As previously, these meshes correspond to three choices of the grating depth. Results are depicted in Fig. II.34. For this second configuration, we note that

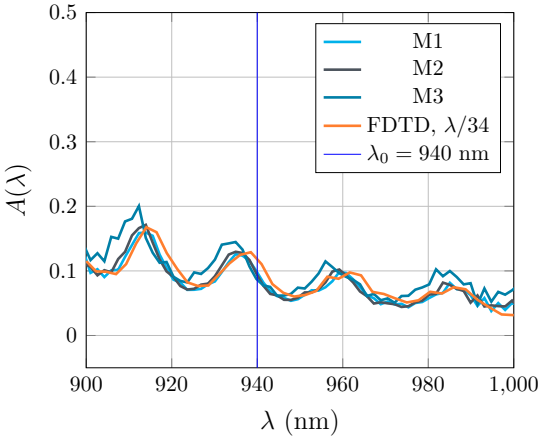


Figure II.33: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C1. Volumic absorption in the internal Si volume. Numerical convergence versus mesh resolution.

Mesh	Depth (nm)	# cells	h_m (nm)	h_M (nm)	Ratio
M1	15.0	100,517	15.0	345.4	23.1
M2	30.0	100,528	30.0	353.1	11.8
M3	60.0	101,377	33.3	344.9	10.4

Table II.7: Characteristics of meshes of the square pixel with a 1D grating. Configuration C2.

as the depth of the grating is increased, while the values selected for the other grating parameters are kept fixed, the absorption spectrum is affected in the lower and upper part of the wavelength window considered here. The third configuration that we consider is characterized by the following setting:

- Configuration C3
  - 1/4 of the full geometry with PEC/PMC conditions on the lateral sides
  - Top PML: 600 nm
  - Bottom PML: 200 nm
  - Dispersive Cu layer instead of PEC wall (bottom side)
  - Width grating: 400 nm
  - Minimum space between rods: 400 nm
  - Physical time: 600 fs
  - DGTD-$\mathbb{P}_4$ method (unless otherwise stated)

Five tetrahedral meshes have been constructed whose characteristics are summarized in Tab. II.8. As previously, these meshes correspond to three choices of the grating depth, which is here taken in the range [15 nm , 180 nm]. Results for meshes M1 to M4 are

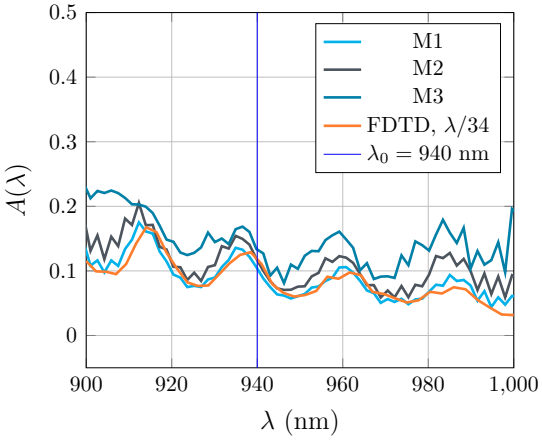


Figure II.34: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C2. Volumic absorption in the internal Si volume. Numerical convergence versus mesh resolution.

Mesh	Depth (nm)	# cells	h_m (nm)	h_M (nm)	Ratio
M1	15.0	101,388	15.0	402.7	26.9
M2	30.0	101,664	30.0	364.3	12.2
M3	60.0	102,550	32.3	377.5	11.5
M4	120.0	101,306	40.7	381.0	9.4
M5	180.0	101,046	39.1	360.3	9.2

Table II.8: Characteristics of meshes of the square pixel with a 1D grating. Configuration C3.

depicted in Fig. II.35. In addition, Fig. II.36 shows a comparison of results that have been obtained with the DGTD-$\mathbb{P}_4$ and DGTD-$\mathbb{P}_5$ methods for mesh M4. Similarly, Fig. II.37 shows a comparison of results that have been obtained with the DGTD-$\mathbb{P}_4$ and DGTD-$\mathbb{P}_5$ methods for mesh M5. With this third configuration, the calculated absorption spectra are always improved as compared to that of the flat structure. However, we also note that for the deeper gratings, numerical convergence in the lower part of the wavelength window is harder to achieve. The fourth configuration that we consider is characterized by the following setting:

- Configuration C4
  - 1/4 of the full geometry with PEC/PMC conditions on the lateral sides
  - Top PML: 600 nm
  - Bottom PML: 200 nm
  - Dispersive Cu layer instead of PEC wall (bottom side)
  - Width grating: 400 nm
  - Depth grating: 120 nm

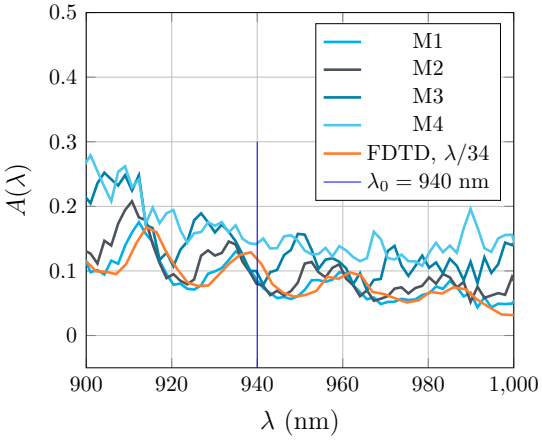


Figure II.35: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C3. Volumic absorption in the internal Si volume. Numerical convergence versus mesh resolution.

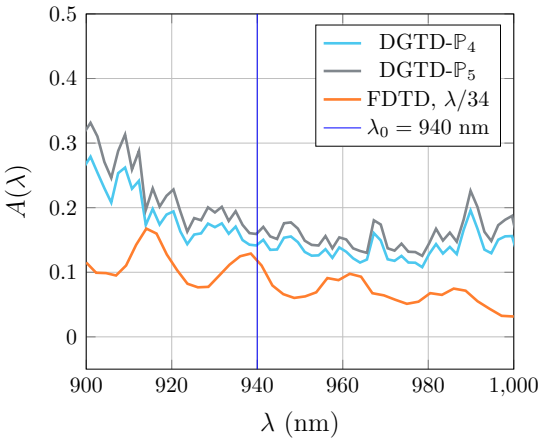


Figure II.36: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C3. Volumic absorption in the internal Si volume. Numerical convergence versus interpolation order with mesh M4.

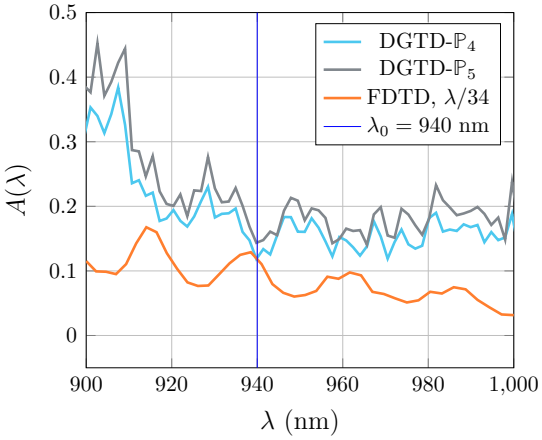


Figure II.37: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C3. Volumic absorption in the internal Si volume. Numerical convergence versus interpolation order with mesh M5.

Mesh	Δ (nm)	# cells	h_m (nm)	h_M (nm)	Ratio
M1	200.0	102,072	43.8	342.7	7.8
M2	100.0	103,715	34.4	334.8	9.8
M3	50.0	104,002	16.8	344.9	20.6

Table II.9: Characteristics of meshes of the square pixel with a 1D grating. Configuration C4.

- Minimum space between rods: Δ
- Physical time: 600 fs
- DGTD-$\mathbb{P}_4$ method (unless otherwise stated)

Three tetrahedral meshes have been constructed whose characteristics are summarized in Tab. II.9. This time, the grating parameter which is varied is the minimum space between rods. Results are presented in Figs. II.38 (mesh M1), II.39 (mesh M2) and II.40 (mesh M3). Finally, in Fig. II.41 we compare the results obtained with the three meshes, i.e., for the three values of the parameter Δ, when using the DGTD-$\mathbb{P}_5$ method. Clearly, all these grating definitions yield a notable improvement of the volumic absorption in the internal Si volume.

II.6.3.3 Conclusion on FDTD and DGTD with a nanostructuring of the square pixel

These comparisons of the absorption spectrum of a nanostructuring of the square pixel illustrate the difficulties that are faced when simulating highly resonant structures. The FDTD abd DGTD results are of similar amplitude, and exhibit similar resonances, as visible on Fig. II.31. However, we do not achieved a perfect absorption spectrum match. This discrepancy is explained by the main difference between the FDTD and the DGTD methods: a cartesian mesh is used by FDTD, and a conformal mesh by DGTD. This

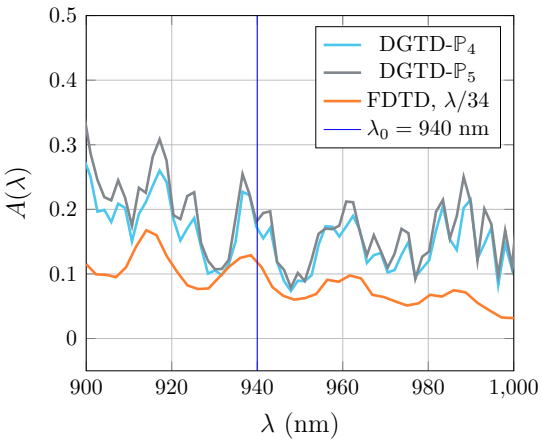


Figure II.38: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C4. Volumic absorption in the internal Si volume. Numerical convergence versus interpolation order with mesh M1.

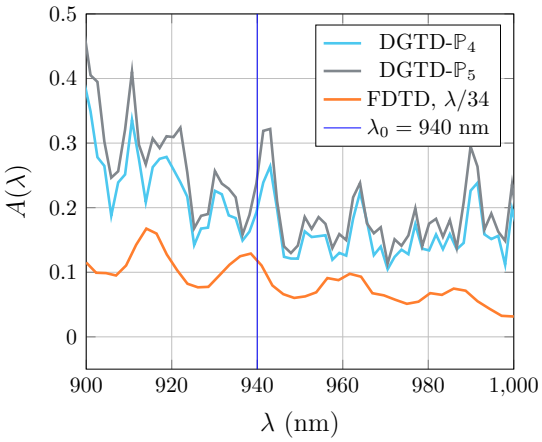


Figure II.39: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C4. Volumic absorption in the internal Si volume. Numerical convergence versus interpolation order with mesh M2.

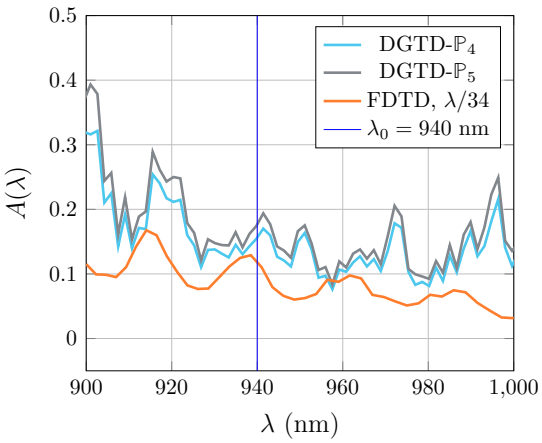


Figure II.40: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C4. Volumic absorption in the internal Si volume. Numerical convergence versus interpolation order with mesh M3.

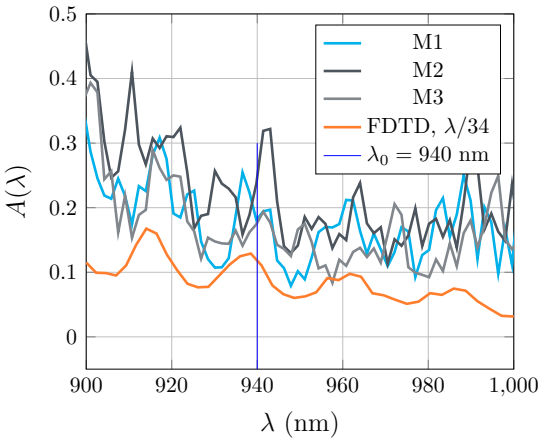


Figure II.41: DGTD results with semi-infinite Cu layer at the bottom. Square pixel with a 1D grating, configuration C4. Volumic absorption in the internal Si volume. DGTD-$\mathbb{P}_5$ method.

difference introduces small geometrical variations. These differences were not impacfull for the three simple structures (see section II.6.1) and for the flat square pixel (see section II.6.2), but for a nanostructured pixel, the light is diffracted on the grating, and then bounces back and forth on both the DTI and the bottom metallic reflector. This diffraction is highly dependent on the trenches size (see section III.1). Thus, a small geometrical error has a high impact on the resulting absorption spectrum, due to the multiple reflections on all interfaces.

A more detailed discussion about simulation time is given in the final conclusion of this chapter.

Finally, these DGTD simulations of a nanostructuring of the square pixel illustrate how important gratings are in order to improve the performance of CMOs imagers. A complete optimization of gratings parameters is performed on a 2D structure in the chapter III.

II.6.4 FDTD and RCWA, octogonal pixels

In this section, we compare our implementation of the RCWA method, described in section II.5, and the reference FDTD software, from Lumerical. The chosen structure is a nanostructured octogonal pixel, in order to focus on a test case particularly difficult for the FDTD methods, since the staircasing effect is enhanced in the x and y dimensions at the grating layer. We do not compare with the DGTD method in this case in order to not increase exponentially the total number of simulations since the previous study already brings us enough results. This study can be seen as the following of the previous one, since nanostructured octogonal pixels were not studied in section II.6.2.

The chosen geometry is a pixel with both octogonal DTI and octogonal grating, a Cu reflector layer at the bottom and a tungsten shield covering the exterior Si volume. Various slices on the geometry are available in Fig. II.42. In order to minimize the permittivity fit error, a short wavelength range ([935, 945] nm) is chosen.

The convergence study of the FDTD method is shown in Fig. II.43, and we note that a meshfactor of 28 is sufficient. The RCWA convergence study is shown in Fig. II.44. A plane wave truncation of 81 seems enough, even if spectral difference can still be seen between the 101 and the 81 plane wave truncation. The cost of a single RCWA simulation, for various plane waves truncation, is shown in Fig. II.45. The spectrum response of both FDTD and RCWA is shown in Fig. II.46. Despite an obvious discrepancy in the results from FDTD and RCWA in Fig. II.46, one must remark that the x-axis is zoomed in, only an interval of 10 nm is compared. This leads us to conclude that the results from FDTD and RCWA are actually in good agreement.

The simulation time required for a wider plane wave truncation is too important (≥ 50 hours) to obtain more converged RCWA results.

Finally, the results of these benchmarks are discussed in the conclusion.

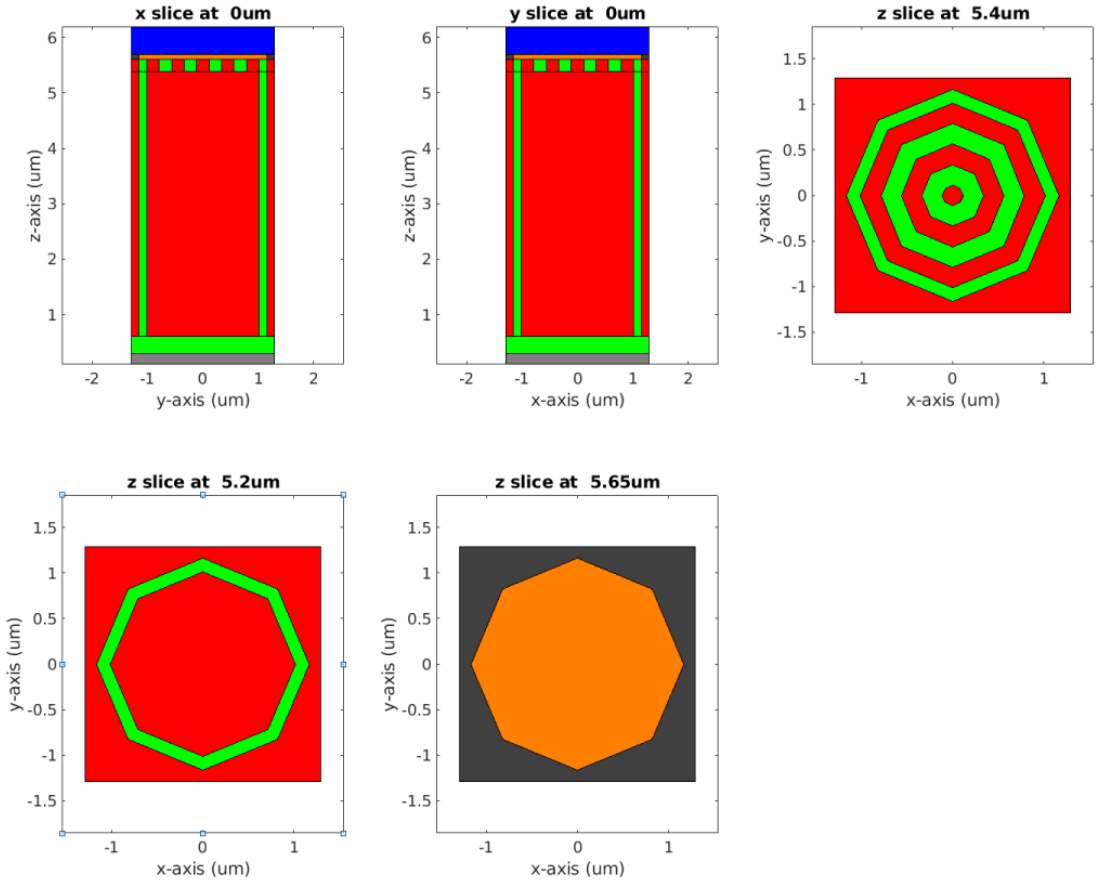


Figure II.42: Description of the nanostructured octogonal pixels. Blue material is air, green is SiO_2, brown is Cu, red is Si, black is Tungstene and orange is TaO_2. Exact dimensions are provided since this pixel is not confidential.

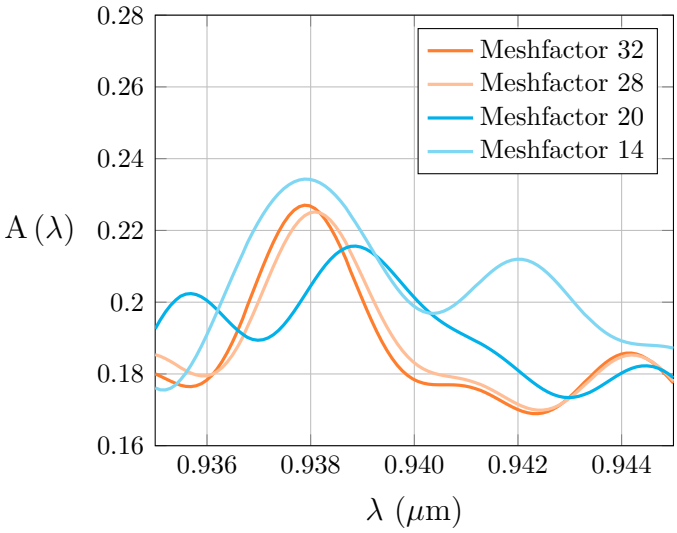
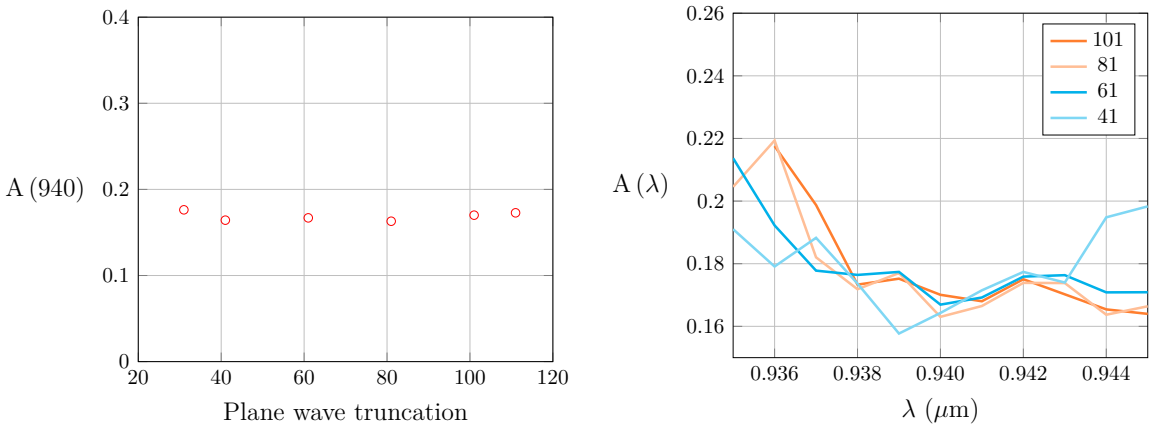


Figure II.43: FDTD convergence study on octogonal pixel described in Fig. II.42. The meshfactor refers to the number of mesh dots per wavelength, adapted to the material permittivity (see section II.4.1 for details).



(a) $\lambda = 940$ nm.

(b) Ten different wavelengths.

Figure II.44: RCWA convergences analysis of the two periods octogonal pixels. For Fig. II.44b, the legend indicates the plane waves truncation.

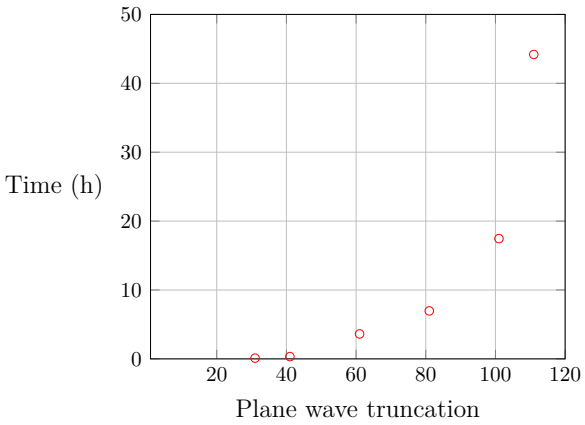


Figure II.45: CPU time of RCWA simulations of Fig. II.44a.

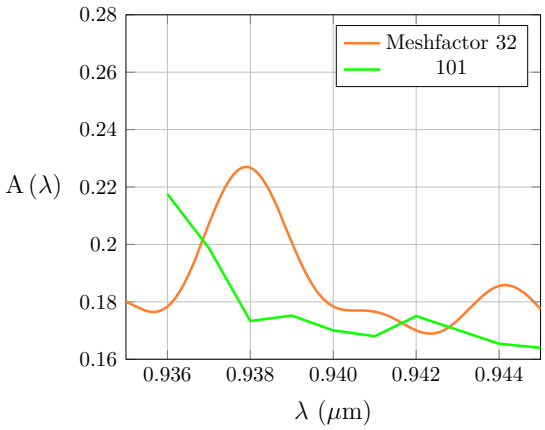


Figure II.46: Comparing FDTD and RCWA on octogonal pixel.

II.7 Conclusion

In this chapter, three numerical methods solving Maxwell's equations in time-domain (FDTD and DGTD) and in frequency-domain (RCWA) have been presented, and compared on various geometries of interest to our study.

The effect of structuration, enhancing light absorption, has been studied. The grating introduces a diffractive effect that is particularly sensitive to the exact geometry simulated: this has been seen with, first, the pyramidal grating, whose reflection and transmission spectra are sensitive to the apex definition (see section II.6.1). Secondly, the 1D grating on the square pixel, simulated with both FDTD and DGTD, exhibits absorption spectrum differences (see section II.6.3), mainly due to the non-conformal mesh used by the FDTD method. Thirdly, the comparison between FDTD and RCWA on octogonal pixels exhibits a similar discrepancy, since the absorption spectrum of RCWA is shifted by 2 nm and present amplitude discrepancies of 10%. These studies show the great importance of conformal numerical methods, since the optical response of nanostructured pixels greatly depends on the geometry considered, in particular with small geometrical variation, of the scale of a typical FDTD mesh steps.

In this study, we faced a common problem of benchmarking studies. The stake was to compare numerical methods, which are theoretical algorithms leading to the unique solution of a well-posed problem, but only software simulations are compared, meaning that all empirical benchmarks compare actual implementation of a given numerical method, and not the numerical methods by themself. This extra layer, the implementation, leads to a specific dimensionality of the benchmark problem, often not clarified by the authors of benchmarks studies: the quality of a solver implementation can drastically impact performance, as well as the tests on speeds performed, and the time allowed on code optimization. In our study, we compared the reference FDTD software, Lumerical developed for more than 20 years by a team of experienced developers, with a RCWA software written in two years by myself and my thesis tutor in STMicroelectronics, with a DGTD solver implemented by the INRIA team Atlantis since December 2015. How one could compare solvers speed, when the amount of time and effort to optimize these software is so unequally distributed ?

Another common problem when benchmarking numerical methods, lies in the definition of the test case considered. In our study, we choose a series of test cases of increasing complexity, starting with a simple nanostructured silicon slab to pixel-like structure. It is worth mentioning that exact pixels were not considered here as a lens must be added on top of the pixel structures studied. But an exact pixel structure is not given only by adding a lens: fabrication process cannot exactly etch DTIs with a constant width in the x and y dimension, nor with a perfect z depth; moreover, every angle in the structure that is assumed to be 90° is a structural approximation in comparison to what is actually feasible in the etching process. Such slight differences between the actual fabricated pixel, and the simulated pixel, can have high importance, since we conclude previously that structural variations of a few nanometers can impact the optical response of nanostructured pixels. To respect confidentiality, we cannot show pictures of fabricated pixels in this work. Anyway, the underlying problem of the test case definition is not specific to pixel benchmarks but to all benchmarks. In Fig. II.47, Delaunay

meshes of decreasing precision of a rabbit are shown. The question is: which meshes are mesh of rabbit, and which meshes are not ? Meaning, what is the criteria to select the required accuracy ? Up to which quality the physical phenomena enhanced by a rabbit are captured ? For the pixels, one could say that the required accuracy is the one that allows to improve the pixel performance. But the underlying problem remains, since the 'performances' dependencies on structure variations, even very small variations, must be clarified.

The initial problem was to select the best numerical methods for the simulation of CMOS pixels. This problem, even with high stakes for both the INRIA Atlantis team, and STMicroelectronics, is an ill-posed problem. In other words, it does not have a unique solution. Instead, the solutions lay in a space of at least two dimensions: it depends on the amount of effort provided to optimize a specific implementation of a numerical method, and it depends on the structural simplification that can, or cannot, be accepted.

So a single answer cannot be provided to the problem introduced in this chapter.

About simulation times, one should have remarked the lack of simulation time comparisons between the FDTD and DGTD numerical methods on the square pixel. One example will show the trend: on the nanostructured pixel study in section II.6.3, the FDTD simulation last 26 hours on 15 CPUs, while the corresponding DGTD results was obtained after 3 hours, on 448 CPUs. Assuming equivalent CPU performances, and assuming that the simulation times scales linearly according to the number of CPUs, the DGTD solver is 3.4 times slower than the FDTD solver. This particular example shows that the DIOGENeS software is usable only by companies that can access hundred of CPUs. STMicroelectronics has not such ressources, and thus cannot use this software for the daily optical simulations of their pixels. However, this conclusion must be tempered, since the two compared softwars did not receive the same amount of optimizing efforts.

If the physical effect is depending highly on the structure precision, then the conformal Delaunay mesh used by the DGTD method is highly beneficial. This precision might come at a high numerical costs, and FDTD will be preferred, even with a highly refined cartesian mesh, when the DGTD simulations are too long. This is exactly the strategy of STMicroelectronics for optical simulations at the moment. RCWA allows to quickly compute the optical response of a large volume, as long as this volume is homogenous in the z dimension, while both FDTD and DGTD simulation time, using meshes, will increase with the volume size. However, whenever a complex geometry is involved, like a lens, or a not perfectly straight DTIs, the RCWA layers definition will increase linearly the simulation time, and the FDTD or DGTD methods could be preferred for achieving both faster and more precise results.

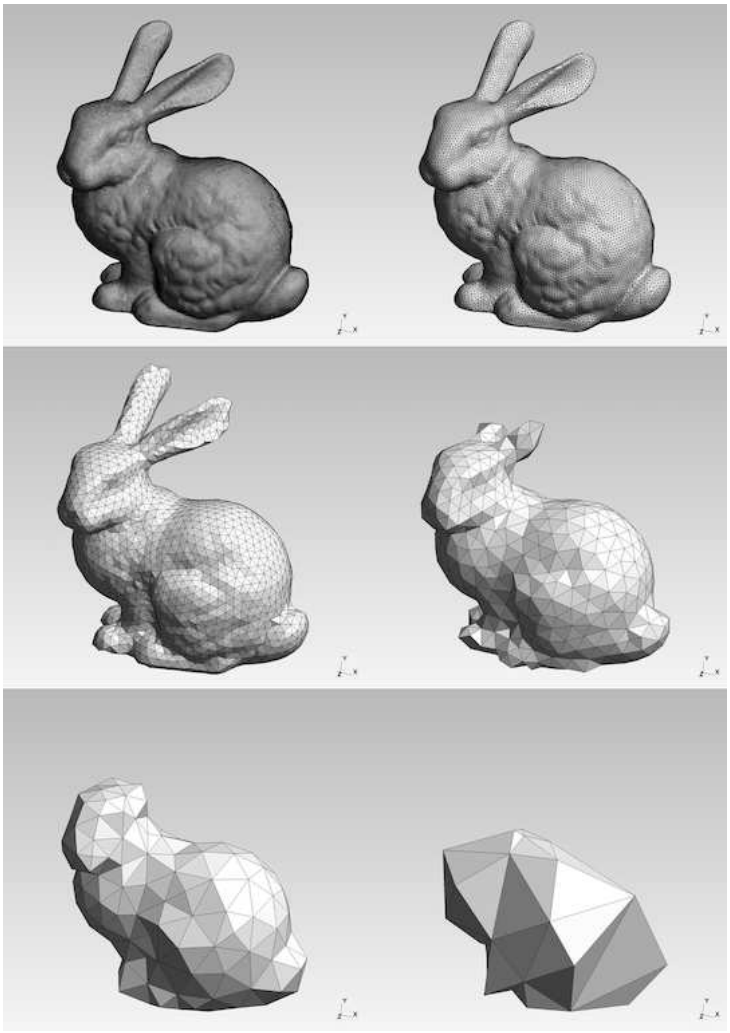


Figure II.47: Delaunay meshes of a rabbit, of various precision. Taken from `https://gmsh.info/`.

Chapter III

Nanostructuration optimization

Contents

III.1 Introduction

CMOS image sensors (CISs) for the near infrared range (NIR) are extensively used in smartphones, laptops, digital cameras, and in various areas such as biological inspection

[77], Time-of-Flight (ToF) [78] and fiber optic communication [79]. NIR light, in particular the wavelength of 940 nm, is mostly used because of its invisibility to the human eyes, allowing constant illuminations required, for example, by distance measurement or facial recognition. However, CIS tends to show low efficiency at such wavelengths mainly because silicon, due to its indirect bandgap, is almost transparent and its absorption is low.

One way to improve NIR light absorption efficiency in CISs is to use a thicker silicon layer for a longer optical path. However, this reduces imaging quality, as crosstalk between pixels is enhanced, and increases the fabrication cost of deep trench isolation (DTI) [80]. Another approach is to use an alternative material, for instance Germanium (Ge) [14], that exhibits a higher absorption in the NIR light due to its direct bandgap. But the associated fabrication process is still challenging mainly due to the appearance of defects [81].

Major improvements so far have been accomplished by the use of a nanostructured pattern on top of silicon substrate. Both plasmonic metal patterning [82]-[83] yielding strong electric field enhancement by resonant coupling between photons and electrons in metal, and diffractive patterning [84] allowing to increase light propagation length and effective silicon thickness, have shown drastic increase in NIR light sensitivity of CISs. Among the various patterns used, such as the rectangular array [82], [83] or nanopillar array [85]-[86], the Inverted Pyramid Array (IPA) [87]-[88] have been used in mass production [89].

The design of such nanoscale patterning schemes heavily relies on numerical modeling and, in most cases, multi-parametric simulations are performed to obtain an exploitable picture of the role of each geometrical parameter. Depending on the complexity of the considered structures, these numerical studies can require a considerable amount of computational resources, especially in the general three-dimensional setting. An alternative and attractive approach is to resort to a numerical optimization approach for discovering optimal sets of geometrical parameters. Although such numerical optimization strategies have been extensively considered in the recent years for metamaterial design and metasurfaces (see in particular [90], their development for nanostructured CMOS image sensors seems to be less remarkable.

Following these research, a recent work [88] has shown the importance of using a superlattice in order to introduce dissymmetry and improve light absorption. By using a numerical optimization methodology and considering a two-dimensional model structure, an improvement of 28.7% in the averaged (from 800 nm to 1000 nm) light absorption compared to identical pyramid array is demonstrated. In [88], the Covariance Matrix Adaptation Evolution Strategy (CMAES) optimization method is used in combination with the Rigorous Coupled Wave Analysis (RCWA) method for the numerical characterization of each candidate design.

In this work, we introduce an inverse design approach that combines optical solvers for the numerical characterization of light absorption in a nanostructured CMOS image sensor, with a statistical learning-based global optimization method, for goal-oriented discovery of the optimal patterning parameters.

Firstly, this work optimizes parameters grating of a realistic 2D SPAD with DTI, computing the light absorption with our 2D RCWA in-house solver, reaching an absorption of 83% at 940 nm. Secondly, following [88], various 3D pattern shapes (ellipsoid, cylinder, rectangle, pyramidal etc.), for symmetric and unsymmetric gratings, are optimized on a

simple silicon slab in order to determine which shapes enhance the most absorption in the 920-980 nm range. This second optimization takes advantage of the geometry versatility of a high order DGTD fullwave solver. Both optimizations are performed with the Efficient Global Optimization (EGO) method, achieving a convergence to the optimum within a reasonable number of solver evaluation [90].

III.2 State of the art

As previously mentioned, diffractive grating arrays are used in production for CMOS imagers in order to enhance light absorption. The problem of improving the performance of such gratings is thus widely spread within the literature [86, 85, 89, 82, 83]. In the current section, the concepts and procedures of such studies are presented. We rely mainly on the article [82] for introducing the physics concepts.

This article proposes a second order plasmonic grating and the concept of resonant-chamber-like pixels to enhance the near-infrared sensitivity of Si image sensors. It is demonstrated that second order plasmonic diffraction is efficient, and that an Si absorption of 48% at 940 nm is obtained. One must remark that the authors focused on improving plasmonic gratings, whose pillars are made of highly reflective metals, such as Ag, Cu, or Al, while we focus in this work on diffractive gratings, usually made of SiO_2. Despites this difference, this article is canonical to present the concept of effective light trace (or effective light path length).

On a flat pixel, the light coming from top is reflected on the metal layer at the bottom of the pixel. One could say that the effective path length is equal to two times the thickness of the Si layer. On Fig. III.1, extracted from [82], the effect of a grating is represented schematically. The light is coming from the top, with a null incident angle. Then the light is diffracted by the grating, impinging with an incidence on the DTIs. And then the light is reflected on the DTI, bouncing back and forth. From, firstly the diffraction by the grating, and secondly the multiple reflections on the DTI, a longer "effective light path" comes into play.

Two equations govern the effective light path: firstly, the light diffraction angles, noted θ_d, of an incident light of angle θ_i, on a symmetric Si grating of period p, is:

$$\theta_d = \arcsin\left(\sin\left(\theta_i\right) + \frac{l\lambda}{n_{Si}p}\right), \tag{III.1}$$

where l is the diffraction order ($l \in \mathbb{Z}$), λ is the wavelength in vacuum, and n_{Si} is the Si refractive index. Secondly, the reflectance on the DTIs, R, is computed with Snell law, that links the angle of the incident light (on the DTI), $\theta_{i,DTI}$, and the angle of the transmitted light, θ_t, and by the Fresnel coefficient

$$\begin{aligned} r_p &= \frac{N_t\cos(\theta_{i,DTI}) - n_i\cos(\theta_t)}{N_t\cos(\theta_{i,DTI}) + n_i\cos(\theta_t)}, \\ r_s &= \frac{n_i\cos(\theta_{i,DTI}) - N_t\cos(\theta_t)}{n_i\cos(\theta_{i,DTI}) + N_t\cos(\theta_t)}, \\ R_p &= (r_p)^2, \\ R_s &= (r_s)^2, \end{aligned} \tag{III.2}$$

$$n_i \sin(\theta_{i,DTI}) = n_t \sin(\theta_t) \qquad \text{(Snell law)}. \tag{III.3}$$

where n_i is the real index of refraction of the incident medium; N_t is given by $N_t = n_t - \mathrm{i}k_t$, where n_t is the real index of refraction of the substrate medium, and k_t is its extinction coefficient. We provide only the single interface Fresnel coefficient in Eq. III.2. These equations are equivalent to considering DTIs of infinite thickness. For DTIs of fixed thickness (as also considered in [82]), one must use the multiple interface Fresnel coefficients, available in [24]. Finally, the reflectance on the Si/DTI interface is a function of both the incident light angle on the DTI, $\theta_{i,DTI}$. and the DTI thickness. An example of such reflectance map is shown in Fig. III.2.

From Eq. III.1 and Eq. III.2, the procedure to improve light absorption is the following: the angle of diffraction, θ_d, given by Eq. III.1, is equal to the angle of the incident light on the DTI, $\theta_{i,DTI}$, of Eq. III.2 and thus it must be chosen to maximize the reflectance on the DTIs. The multiple reflections on the DTIs increase drastically the absorption, and result in a "resonant-chamber-like pixel" [82]. The author even pushed forward such theory by analyzing the serie resulting from the multiple reflections on the DTIs, thus creating a new figure of merit by computing the effective light trace as a function of R (see Fig. III.3).

This contribution is surely interesting, since it provides a physical intuition of the resonances phenomenon. But, as always, the claim of a breakthrough result is hiding the underlying hypothesis of such analysis: what is not explicitly written in [82] ? what is the axiom that they did not clarify ?

The hidden axiom is the following: the analysis of a "resonant-chamber-like pixel" can be performed independently on the different parts of the pixel. In other words, it is possible to isolate both the grating on one hand, and the DTI on the other hand, in order to model the resonances that occur in the whole pixel.

In our work, we claim that the analysis of the resonances cannot be fully performed by summing the modeling of, first, the grating, and then, the DTI. Furthermore, the resonances are always coupled resonances, thus depending on all parameters not independently. In contrast to [82], we take a step back, and the resonances resulting from a diffraction grating, of a SPAD pixel with DTIs are considered as a function of all geometrical parameters. A resonance is here defined simply as a peak in the absorption profile (see for instance Fig. III.15a). Finally, increasing the light absorption at 940 nm in a pixel is equivalent to positioning a resonance exactly at 940 nm, instead of increasing the effective path length.

Mathematically, this methodology consists of treating the optical absorption, of the inner Si, as a black-box function,

$$f(\lambda, p_1, p_2, \ \dots \ , p_n) := A_{\lambda, p_1, p_2, \ \dots \ , p_n}, \tag{III.4}$$

where λ is the wavelength of the incident light, and p_i are all the geometrical parameters that will be defined in the following. Positioning a resonance exactly at 940 nm is equivalent to the maximization problem

$$\max_{p_1, p_2, \ \dots \ , p_n} f(940, p_1, p_2, \ \dots \ , p_n), \tag{III.5}$$

and increasing absorption light on an NIR interval, for instance $[920, 980]$ nm, is equivalent to the maximization problem

$$\max_{p_1, p_2, \ \dots \ , p_n} \int_{920}^{980} f(x, p_1, p_2, \ \dots \ , p_n) \mathrm{d}x. \tag{III.6}$$

The black-box function f can be evaluated on a set of parameters p_i for $i = 1, ..., n$ with an optical solver. Then the maximization problem can be solved as an optimization problem.

In order to minize the total number of evaluation, the optimization is performed with the Bayesian Optimization method, also referred as Efficient Global Optimization (EGO). With such optimizer, the black-box function f is modeled by a surrogate gaussian process (see section III.3.2.3).

For simplicity, since we aim to demonstrate the validity of our methodology, we optimize in this work only TM polarized incident light.

In the following, firstly the optimization methodology is described in section III.3. Secondly, in section III.4 an optimization of a 2D realistic SPADs is performed in order to position a resonance exactly at 940 nm. Thirdly, in section III.5, a 3D optimization is performed on a simple Silicon slab, in order to investigate the performance of various grating shapes.

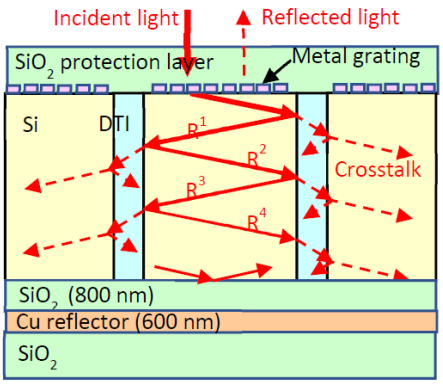


Figure III.1: This figure is extracted from [91]. Schematic diagram of resonant-chamber-like pixels with plasmonic diffraction. Effective light trace is increased by reflections.

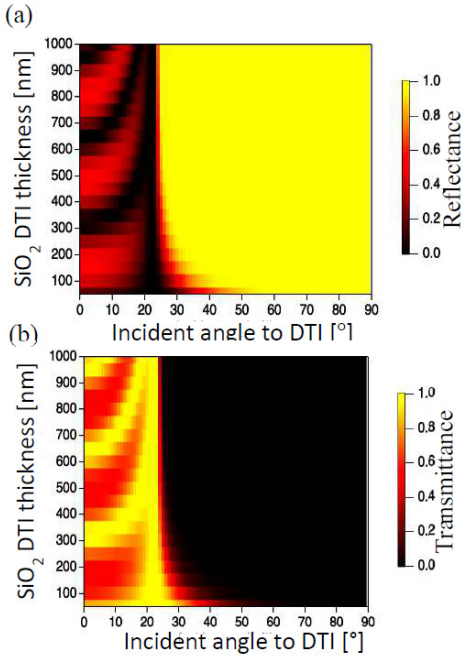


Figure III.2: This figure is extracted from [82]. Dependance of reflectance (a) and transmittance (b), of SiO2 DTI on incident angle and DTI thickness. These heatmaps are obtained with the multiple interface Fresnel coefficient (see [24]).

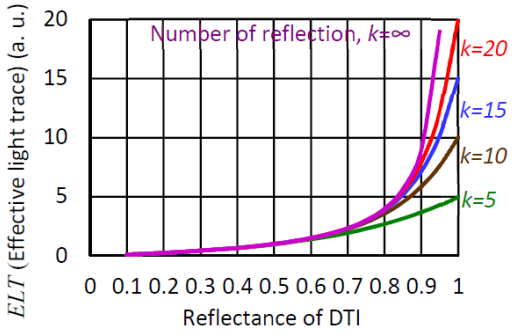


Figure III.3: This figure is extracted from [82]. ELT is computed using $ELT = R(1 - R^k)/(1 - R)$, where k is the number of reflections and R is the reflectance of the Si/DTI interface.

III.3 Bayesian optimization

In this section, the Bayesian optimization framework and the EGO method in particular, as well as all the underlying probability concepts, are recalled.

III.3.1 Introduction

Optimization is an essential tool for finding the best solution among the set of all feasible solutions. When coupled to a numerical solver, it allows efficient prototyping by reducing the total number of solver evaluations, especially when compared to standard parameter sweeps. When the number of parameters is high, or equivalently when the set of all feasible solutions is of high dimension, or when parameters effects are coupled, using an efficient optimizer leads to a drastic reduction of computational cost.

Bayesian optimization [1] is defined by Jonas Mockus in [92] as an optimization technique based upon the minimization of the expected deviation from the extremum of the studied function. The objective function is treated as a black-box function. A Bayesian strategy sees the objective as a random function and places a prior over it. The prior, a surrogate gaussian process model for EGO, captures our beliefs about the behavior of the function. After gathering the function evaluations, which are treated as data, the prior is updated to form the posterior distribution over the objective function. The posterior distribution, in turn, is used to construct an acquisition function (often also referred to as merit function) that determines what the next query point should be.

One of the earliest bodies of work on Bayesian optimization that we are aware of are [93] and [94]. Kushner used Wiener processes for one-dimensional problems. Kushner's decision model was based on maximizing the probability of improvement (see Eq. III.57), and included a parameter that controlled the trade-off between 'more global' and 'more local' optimization, in the same spirit as the Exploration/Exploitation trade-off.

Meanwhile, in the former Soviet Union, Mockus and colleagues developed a multidimensional Bayesian optimization method using linear combinations of Wiener fields, some of which was published in English in [92]. This paper also describes an acquisition function that is based on myopic expected improvement of the posterior, which has been widely adopted in Bayesian optimization as the Expected Improvement function (see Eq. III.58).

In 1998, Jones [95] used Gaussian processes surrogate together with the expected improvement function to successfully perform derivative-free optimization and experimental design through an algorithm called Efficient Global Optimization, or EGO.

In the rest of this document, Bayesian optimization will always refer to EGO.

III.3.2 Preliminaries on probability theory

First we recall the usual probability definitions. In order to describe briefly but accurately the theoretical background of Bayesian optimization, we first define random variables, secondly random vectors are defined, followed in the third step by random

[1] This brief historical background is freely inspired from the documentation of the SMT python package.

processes, focusing on gaussian processes. These definitions are of increasing complexity: the random vectors are defined upon the random variables, then the random processes are defined upon the random vectors.

III.3.2.1 Random variables

For a proper definition of measure, probability space, tribe, borelian, borelian measure, measurable function, independent random variable, we refer to [96]. We focus in this section only on a minimal introduction.

A **real random variable** (r.v.) is a measurable function, noted X, from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with value in $\mathbb{R}$ provided with the borelian tribe $\mathcal{B}(\mathbb{R})$. The **law of a r.v.**, X, is the measure of probability, noted $\mathbb{P}_X$ defined on $\mathbb{R}$ by,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(\omega \in \Omega | X(\omega) \in A), \quad \forall A \in \mathcal{B}(\mathbb{R}). \tag{III.7}$$

A r.v. is said to be **discrete** if its values are in a discrete subset of $\mathbb{R}$. A r.v. is said to have a **density**, with density f, if there exists a function f, defined on $\mathbb{R}$, positive or zero, integrable on $\mathbb{R}$, continuous almost everywhere, and of integral equals to 1, such that we have:

$$\forall A \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}_X(A) = \int_A f \mathrm{d}\lambda, \tag{III.8}$$

or equivalently, for all a, b in $\mathbb{R}$ such as $a < b$, we have:

$$\mathbb{P}_X([a,b]) = \int_a^b f(x)\mathrm{d}x. \tag{III.9}$$

In the following, we focus only on real random variables that admit a density.

The usual quantity associated to a r.v., with density f, are:

- the **expectation**, also named the mean, noted $\mathbb{E}[X]$ and defined as:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x)\mathrm{d}x. \tag{III.10}$$

  The expectation is linear. A r.v. is said to be **centered** when $\mathbb{E}[X] = 0$.

- the **variance**, noted $\mathrm{Var}(X)$ and defined as:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right], \tag{III.11}$$

  A r.v. is said to be **reduced** when $\mathrm{Var}[X] = 0$.

- the **standard deviation**, noted σ_X and defined as:

$$\sigma_X = \sqrt{\mathrm{Var}(X)}. \tag{III.12}$$

The **covariance** of two r.v. X_1 and X_2, having finite variance is noted $\mathrm{Cov}(X_1, X_2)$, and it is defined by:

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])\right]. \tag{III.13}$$

The covariance of two r.v. is a bilinear application. We clearly have $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. So the variance is a quadratic application. If X_1 and X_2 are independent, then one has $\mathrm{Cov}(X_1, X_2) = 0$. The reciprocal is false.

Given $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$, a **gaussian**, or **normal**, variable of expectation μ and variance σ^2 is a r.v. whose law is noted $\mathcal{N}(\mu, \sigma^2)$ and its density, f, is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{III.14}$$

Fig. III.4 shows two examples of gaussian distribution corresponding to the gaussian law $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(1, 0.75)$. A gaussian variable, Z, of law $\mathcal{N}(\mu, \sigma^2)$ is also defined as the translated and inflated of the centered reduced gaussian law, Z_0, of law $\mathcal{N}(0, 1)$. *I.e.* we have the following law equality:

$$Z = \mu + \sigma Z_0. \tag{III.15}$$

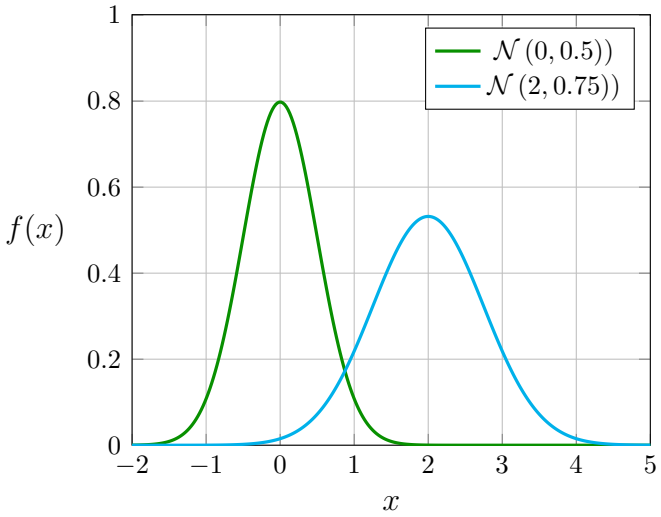


Figure III.4: Density of various gaussian laws.

To summarize, a r.v. is given by its value space and its law of probability. The expectation is equivalent to the mean and the variance is describing the dispersion around this mean value.

III.3.2.2 Random vectors

A **random vector** $\mathbf{X} = (X_1, ..., X_n)$ is an measurable application from $(\Omega, \mathcal{F}, \mathbb{P}(\Omega))$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. For all i, the random variable X_i is called the ith marginal. Each random vector has a law of probability on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, noted $\mathbb{P}_{\mathbf{X}}$. A random vector is said to be discrete if the set of all its value $\mathbf{X}(\Omega)$ is discrete in $\mathbb{R}^n$. A random vector admits a **density**, also named **joint density**, noted $f(x_1, ..., x_n)$, if:

$$\mathrm{d}\mathbb{P}(\mathbf{x}) = f(x_1, ..., x_n)\mathrm{d}x_1...\mathrm{d}x_n \Leftrightarrow \mathbb{P}_{\mathbf{X}}(A) = \int_A f(x_1, ..., x_n)\mathrm{d}x_1...\mathrm{d}x_n, A \in \mathcal{B}(\mathbb{R}^n). \tag{III.16}$$

In such a case, we have, as expected, $f \geq 0$ on $\mathbb{R}^n$ and $\int_{\mathbb{R}^n} f = 1$. The **expectation** of a random vector $\mathbf{X}$ is the vector of the expectation of its marginals:

$$\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], ..., \mathbb{E}[X_n]). \tag{III.17}$$

The **covariance matrix** of a random vector $\mathbf{X}$ is the square, symmetric, positive matrix:

$$K_{\mathbf{X}} \;=\; \mathrm{Cov}(\mathbf{X}) \;=\; \left(\mathrm{Cov}\left(X_i, X_j\right)\right)_{1\leq i,j\leq n}. \tag{III.18}$$

If $\mathbb{E}[\mathbf{X}] = 0_n$, the vector $\mathbf{X}$ is said to be centered. If a random vector, $\mathbf{X}$, admits a density, f, then the ith marginal X_i admits as density:

$$f_{X_i}(x_i) \;=\; \int_{\mathbb{R}^{n-1}} f(x_1, ..., x_n)\mathrm{d}x_1...\mathrm{d}x_{i-1}\mathrm{d}x_{i+1}...\mathrm{d}x_n. \tag{III.19}$$

From the density of a vector, one can get the density of its marginal. However the contrary is not true in all generality. Actually, the density of the vector is the product of the density of all marginals if and only if all its marginals are independent. If all marginals of a random vector are independent, then its covariance matrix is a diagonal matrix.

One example of random vectors is given by gaussian vectors. A random vector $\mathbf{X}$ is said to be gaussian if and only if all linear combinations of its marginals are real gaussian random variables. Namely:

$$\langle \mathbf{a}, \mathbf{X}\rangle \;=\; a_1X_1 + ... + a_nX_n, \tag{III.20}$$

is a gaussian random variable, for all $\mathbf{a}$ in $\mathbb{R}^n$.

If $\mathbf{X}$ is a gaussian vector, then the real random variable $\langle \mathbf{a}, \mathbf{X}\rangle$ has for law

$$\langle \mathbf{a}, \mathbf{X}\rangle \;\sim\; \mathcal{N}\left(\langle \mathbf{a}, \mathbb{E}[\mathbf{X}]\rangle,\; \mathbf{a}^t\mathrm{Cov}(\mathbf{X})\mathbf{a}\right), \tag{III.21}$$

where $\mathbf{a}$ is a column vector of $\mathbb{R}^n$, $\mathbb{E}[\mathbf{X}]$ is the expectation of $\mathbf{X}$ (see Eq. III.17), and $\mathrm{Cov}(\mathbf{X})$ is the covariance matrix of $\mathbf{X}$ (see Eq. III.18).

Indeed, by hypothesis, $\langle \mathbf{a}, \mathbf{X}\rangle$ is gaussian. And its expectation and its variance are:

$$\mathbb{E}\left[\langle \mathbf{a}, \mathbf{X}\rangle\right] = \mathbb{E}\left[\sum_{i=1}^{n} a_iX_i\right] = \sum_{i=1}^{n} a_i\mathbb{E}\left[X_i\right] = \langle \mathbf{a}, \mathbb{E}[\mathbf{X}]\rangle, \tag{III.22}$$

$$\mathrm{Var}\left(\langle \mathbf{a}, \mathbf{X}\rangle\right) = \mathrm{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i,j=1}^{n} a_ia_j\mathrm{Cov}\left(X_i, X_j\right) = \mathbf{a}^t\mathrm{Cov}\left(\mathbf{X}\right)\mathbf{a}, \tag{III.23}$$

where we used the linearity of the expectation and the bilinearity of the covariance (see section III.3.2.1). With $\mathbf{a} = (1, 0, ..., 0)$ in Eq. III.21, one finds, as expected, that X_1 has for law $\mathcal{N}\left(\mathbb{E}[X_1], \mathrm{Var}(X_1)\right)$. More generally, for all $1 \leq i \leq n$, X_i has for law $\mathcal{N}\left(\mathbb{E}[X_i], \mathrm{Var}(X_i)\right)$.

Given a gaussian vector with an invertible covariance matrix, its joint density is given by:

$$\mathbb{P}_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{1}{\left((2\pi)^n\det\left[\mathrm{Cov}(\mathbf{X})\right]\right)^{1/2}}\exp\left(-\frac{1}{2}\langle(\mathbf{x} - \mathbb{E}[\mathbf{X}]), \mathrm{Cov}(\mathbf{X})^{-1}(\mathbf{x} - \mathbb{E}[\mathbf{X}])\rangle\right)\mathrm{d}\mathbf{x}. \tag{III.24}$$

This law is noted $\mathcal{N}(\mathbb{E}[\mathbf{X}], \mathrm{Var}(\mathbf{X}))$ and it is fully known once given its expectation vector and its covariance matrix.

Let $\mathbf{X} = (X_1, ..., X_n)$ and $\mathbf{Y} = (Y_1, ..., Y_m)$, with n and m positive integer, be jointly Gaussian vectors, where:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbb{E}[\mathbf{X}] \\ \mathbb{E}[\mathbf{Y}] \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right) = \mathcal{N}\left(\begin{bmatrix} \mathbb{E}[\mathbf{X}] \\ \mathbb{E}[\mathbf{Y}] \end{bmatrix}, \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{C}^T & \tilde{B} \end{bmatrix}^{-1} \right), \tag{III.25}$$

then the *marginal* law of $\mathbf{X}$ is:

$$\mathbf{X} \sim \mathcal{N}(\mathbb{E}[\mathbf{X}], A), \tag{III.26}$$

and the *conditional* law of $\mathbf{Y}$ given $\mathbf{X}$ is:

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\mathbb{E}[\mathbf{Y}] + C^T A^{-1}(\mathbf{X} - \mathbb{E}[\mathbf{X}]), B - C^T A^{-1} C), \tag{III.27}$$

or

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\mathbb{E}[\mathbf{Y}] + \tilde{B}^{-1}\tilde{C}^T(\mathbf{X} - \mathbb{E}[\mathbf{X}]), \tilde{B}^{-1}). \tag{III.28}$$

The exact definition of a conditional law is fairly complex, and, in our view, not enough explained though the definition of conditional events by the usual Bayes formula. Thus, we refer to [97, 98] for a complete definition.

III.3.2.3 Random processes

A random process, $\mathcal{Z} := (\mathcal{Z}_x)_{x \in D}$ is a set of real random variables indexed by a set D. If D is a finite set, then $\mathcal{Z}$ is a random vector. If $D = \mathbb{N}$, then $\mathcal{Z}$ is a series of random variables. More generally, if $D \subset \mathbb{Z}$, then $\mathcal{Z}$ is said to be discrete. Usually, a random process is interesting by itself when $D \subset \mathbb{R}^d$ for a given d in $\mathbb{N}^*$, or $D = \mathbb{R}^+$. In such case, when the process is indexed by time ($D = \mathbb{R}^+$), the notation $\mathcal{Z}_t$ is used. And when the process is indexed by one or several spatial parameters ($D \subset \mathbb{R}^d$), the notation $\mathcal{Z}_x$ is preferred. In this work, we will use only "spatial" gaussian process, therefore the notation $\mathcal{Z}_x$ is used.

A random process depends on two parameters: $\mathcal{Z}_x(\omega)$ depends on $x \in D$ (a spatial parameter), and the random $\omega \in \Omega$. Both marginals have a different interpretation:

- For a given x in D, $\omega \mapsto \mathcal{Z}_x(\omega)$ is a real random variable;
- For a given ω, $x \mapsto \mathcal{Z}_x(\omega)$ is a real function on D, called a **realization** or trajectory of the random process $\mathcal{Z}$.

At first sight, a random process provides two types of information: all elements of D are associated to a real r.v., and all these r.v. can be simulated in order to produce one realization of the random process.

A random process, $\mathcal{Z}$, is characterized by all its finite dimensional laws, defined as, the law of all finite vectors $(X_{t_1}, X_{t_2}, ..., X_{t_n})$, for all $t_1, ...t_n$ in D, n in $\mathbb{N}^*$. So defining a random process is equivalent to defining all its finite dimensional laws. And a random process has no law by itself, only all its finite dimensional vectors have a law.

In the current work, we focus on the **gaussian process**, defined as a random process whose finite dimensional laws are gaussian vectors (see section III.3.2.2). For a complete introduction, please refer to [99]. A gaussian process, noted $\mathcal{Z} = (\mathcal{Z}_x)_{x\in D}$, is entirely characterized by its mean function:

$$\mu_{\mathcal{Z}} \ : \ D \to \mathbb{R}, \quad \mu_{\mathcal{Z}}(x) = \mathbb{E}\left[\mathcal{Z}_x\right], \tag{III.29}$$

and its covariance kernel:

$$K_{\mathcal{Z}} \ : \ D \times D \to \mathbb{R}^+, \quad K_{\mathcal{Z}}(x_1, x_2) = \mathrm{Cov}\left(\mathcal{Z}_{x_1}, \mathcal{Z}_{x_2}\right). \tag{III.30}$$

Indeed, given n in $\mathbb{N}^*$ and $\mathbf{X} := \{x_i \mid i \in [[1, n]]\}$ in D^n, one gets the mean vector and the covariance matrix of the gaussian vector $\mathcal{Z}_{(1,n)} := (\mathcal{Z}_{x_1}, \ ... \ , \mathcal{Z}_{x_n})$ from the mean function, $\mu_{\mathcal{Z}}$, and the covariance kernel, $K_{\mathcal{Z}}$, by:

$$\mathbb{E}\left[\mathcal{Z}_{(1,n)}\right] \ := \ \begin{bmatrix} \mathbb{E}\left[\mathcal{Z}_{x_1}\right] \\ ... \\ \mathbb{E}\left[\mathcal{Z}_{x_n}\right] \end{bmatrix} \ = \ \begin{bmatrix} \mu_{\mathcal{Z}}(x_1) \\ ... \\ \mu_{\mathcal{Z}}(x_n) \end{bmatrix} \tag{III.31}$$

and,

$$\mathrm{Var}\left(\mathcal{Z}_{(1,n)}\right) \ := \ \left[\mathrm{Cov}\left(\mathcal{Z}_{x_i}, \mathcal{Z}_{x_j}\right)\right]_{1\leq i,j\leq n} \ = \ \left[K_{\mathcal{Z}}\left(x_i, x_j\right)\right]_{1\leq i,j\leq n} = K_n^{\mathcal{Z}}(\mathbf{X}). \tag{III.32}$$

where $K_n^{\mathcal{Z}}\left(\mathbf{X}\right)$ is the covariance matrix associated to the covariance kernel $K_{\mathcal{Z}}$, *i.e.*,

$$\forall n \in \mathbb{N}^*, \quad \forall \mathbf{X} \in D^n, \qquad K_n^{\mathcal{Z}}\left(\mathbf{X}\right) := \ \left[K_{\mathcal{Z}}\left(x_i, x_j\right)\right]_{1\leq i,j\leq n}. \tag{III.33}$$

The covariance kernel of a gaussian process is, by definition, definite positive. Explicitly $K_{\mathcal{Z}}$ has the following property:

$$\forall n \in \mathbb{N}^*, \quad \forall \mathbf{X} \in D^n, \quad \forall \mathbf{A} \in \mathbb{R}^n, \qquad \left\langle \mathbf{A}, K_n^{\mathcal{Z}}\left(\mathbf{X}\right)\mathbf{A}\right\rangle \geq 0. \tag{III.34}$$

where $\langle \cdot, \cdot \rangle$ is the usual scalar product of $\mathbb{R}^n$, and $K_n^{\mathcal{Z}}$ is defined by Eq. III.33.

In the next section, we focus on how one can simulate a realization of a gaussian process given its mean function and its covariance kernel. Then the different types of covariance kernels are described.

III.3.2.4 Simulation of a Gaussian random process

Generating a realization of a gaussian process is equivalent to simulating a gaussian vector. For instance, suppose given a centered gaussian process, $\mathcal{Z}$, on the interval $[-6, \ 6]$ of known covariance kernel Σ. Generating a realization of $\mathcal{Z}$ is done by picking a sample, $\{x_i \mid i \in [[1, n]]\}$, in $[-6, \ 6]$, and simulating the corresponding gaussian vector $(\mathcal{Z}_{x_1}, \ ... \ , \mathcal{Z}_{x_n})$. The more points in the sample the more refined is the realization. Thus, in the following, we describe how to simulate a gaussian vector.

Simulating a gaussian vector is possible from multiple simulations of independent real gaussian variables. At least three methods (Cholesky, Singular Value Decomposition and Eigen decomposition) exist and all of them rely on the decomposition of the covariance matrix Σ. In the following we present the simulation method based on the Cholesky decomposition.

The covariance matrix Σ is, by definition, symmetric positive, and has a Cholesky decomposition, $\Sigma = LL^T$, where L is a triangular inferior matrix. From the Cholesky decomposition, one can compute the inverse matrix by $\Sigma^{-1} = (L^{-1})^T L^{-1}$ and the determinant of Σ. Given a gaussian vector, $\mathcal{Z}_n \sim \mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$, to simulate. Then, with the Cholesky decomposition $\Sigma_n = LL^T$ and the vector of n independent gaussian variables, $g \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_{nn})$, one can writes:

$$\mathcal{Z}_n = L\mathbf{g} + \boldsymbol{\mu}_n, \tag{III.35}$$

and we have as expected,

$$\begin{aligned}
&\mathbb{E}[\mathcal{Z}_n] = L\mathbb{E}[\mathbf{g}] + \boldsymbol{\mu}_n = \boldsymbol{\mu}_n, \\
&\mathrm{Cov}(\mathcal{Z}_n) = \mathbb{E}\left[\mathcal{Z}_n \mathcal{Z}_n^T\right] = L\mathbb{E}\left[\mathbf{g}\mathbf{g}^T\right] L^T = \Sigma_n.
\end{aligned}$$

Since all marginals of $\mathbf{g}$ are independent, the simulation of the gaussian vector $\mathcal{Z}_n$ is reduced to the simulation of n independent gaussian variables. Finally, these n simulations can be accomplished, for instance, with the standard Box-Muller method.

Other simulation methods rely on the same principle but on a different decomposition of the covariance matrix: either with the Singular Value Decomposition or the Eigen value decomposition [2]. The simulation of a correlated (conditional) gaussian vector is straightforward by combining Eq. III.27 and Eq. III.35.

On Fig. III.5, one can see five uncorrelated realizations of a centered gaussian vector with an exponential kernel (see Table III.1) of parameter $\sigma^2 = 1$. On Fig. III.6, one can see five correlated realizations of the same gaussian vector.

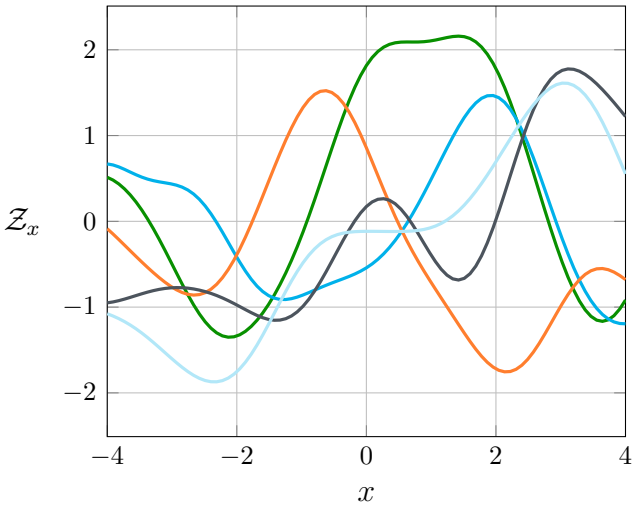


Figure III.5: Five uncorrelated realizations of a gaussian process.

[2]For instance, the numpy Python package implements all three aforementioned simulation methods (see the Multivariate normal numpy function).

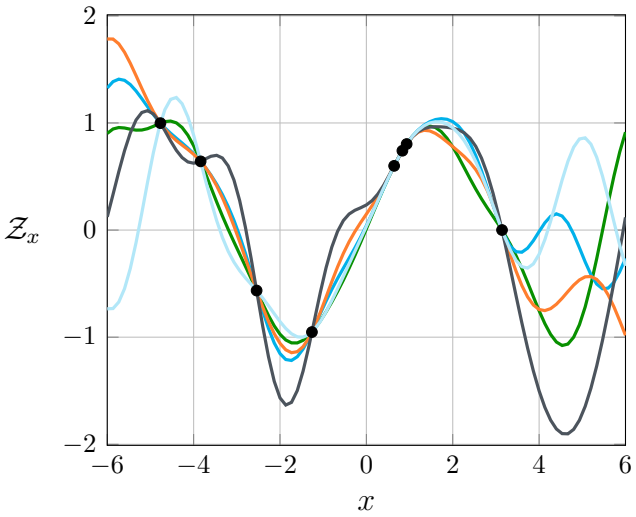


Figure III.6: Five correlated realizations of a gaussian process.

III.3.2.5 Covariance kernel

The covariance kernel of a gaussian process, also simply referred as its kernel, contains all the assumptions on the regularity of the objective function: the regularity of f is directly determined by the choice of the kernel. For instance if f is periodic, then periodic kernels are preferred. It is the essential component of the gaussian process surrogate.

For a one dimensional gaussian process (indexed by $D \subset \mathbb{R}$), the standard covariance kernels are given in the Tab. III.1 and illustrated in Fig III.7. Kernels can be combined by elementwise multiplication of their corresponding covariance matrix. For instance the local periodic kernel is obtained by the multiplication of the gaussian and the periodic kernel (see Tab. III.1), and is expressed as:

$$\forall (x_a, x_b) \in \mathbb{R}^2, \qquad K_{lp}\left(x_a, x_b\right) \;:=\; \sigma^2 \exp\left(-\frac{2}{\ell_p^2}\sin^2\left(\pi\frac{|x_a - x_b|}{p}\right)\right)\exp\left(\frac{-\,|x_a - x_b|^2}{2\ell_{eq}^2}\right), \tag{III.36}$$

where $(\sigma^2, \ell_p, \ell_{eq}, p) \in (\mathbb{R})^4$ are its parameters.

For multidimensional gaussian process (indexed by $D \subset \mathbb{R}^d$), covariance kernels are built as tensorial product of unidimensional kernels. The kernel parameters are then indexed by i in $[[1, d]]$, and allow varying the regularity of the objective function f per dimension.

For instance, a d dimensional gaussian kernel is expressed as,

$$\forall \left(\mathbf{X}, \mathbf{X}'\right) \in \left(\mathbb{R}^d\right)^2, \qquad K_{g,d} \;=\; \theta_1 \exp\left(-\frac{|x_i - x_i'|^2}{\ell_i^2}\right) \tag{III.37}$$

where $\mathbf{X} = (x_1,\ \ldots, x_n)$ and $\mathbf{X}' = (x_1',\ \ldots, x_n')$. The vector of all its parameters

$$\Theta := (\theta,\ \ell_1,\ \ldots,\ \ell_n) \tag{III.38}$$

is also called the **hyperparameters** of the kernel $K_{g,d}$.

Kernel name	Notation	$K(x_a, x_b)$, $\forall (x_a, x_b) \in \mathbb{R}^2$	Parameters
White noise	K_{wn}	$\sigma^2 \mathbb{1}_{x_a = x_b}$	$\sigma^2 \in \mathbb{R}^+$
Gaussian	K_g	$\sigma^2 \exp\left(\frac{-\lvert x_a - x_b\rvert^2}{2\ell^2}\right)$	$(\sigma^2, \ell) \in (\mathbb{R}^+)^2$
Rational quadratic	K_{rq}	$\sigma^2 \left(1 + \frac{\lvert x_a - x_b\rvert^2}{2\alpha\ell^2}\right)^{-\alpha}$	$(\sigma^2, \alpha, \ell) \in (\mathbb{R}^+)^3$
Periodic	K_p	$\sigma^2 \exp\left(-\frac{2}{\ell^2}\sin^2\left(\pi\frac{\lvert x_a - x_b\rvert}{p}\right)\right)$	$(\sigma^2, \ell, p) \in (\mathbb{R}^+)^3$

Table III.1: Various covariance kernels illustrated in Fig III.7. Refer to section III.3.2.3 for the definition of a gaussian process and its covariance kernel.

All the preliminaries on probability and the gaussian process are now recalled. In the next section, we focus on how gaussian processes are used as surrogate models of black-box function, and then how this gaussian process modeling is used to form an optimizer called EGO.

III.3.3 Gaussian process surrogates

The regression by gaussian process, or surrogate gaussian process modeling, comes initially from geology. The theoretical formalization was proposed by Georges Matheron [100] in the 1960s, naming these methods "kriging" in reference to Danie Krige, an engineer known for his work on ore deposits [101]. The main idea is to use the prediction of surrogate models, built on a finite set of observations, to perform interpolation in the phase set, *i.e.* in the set of the parameters considered.

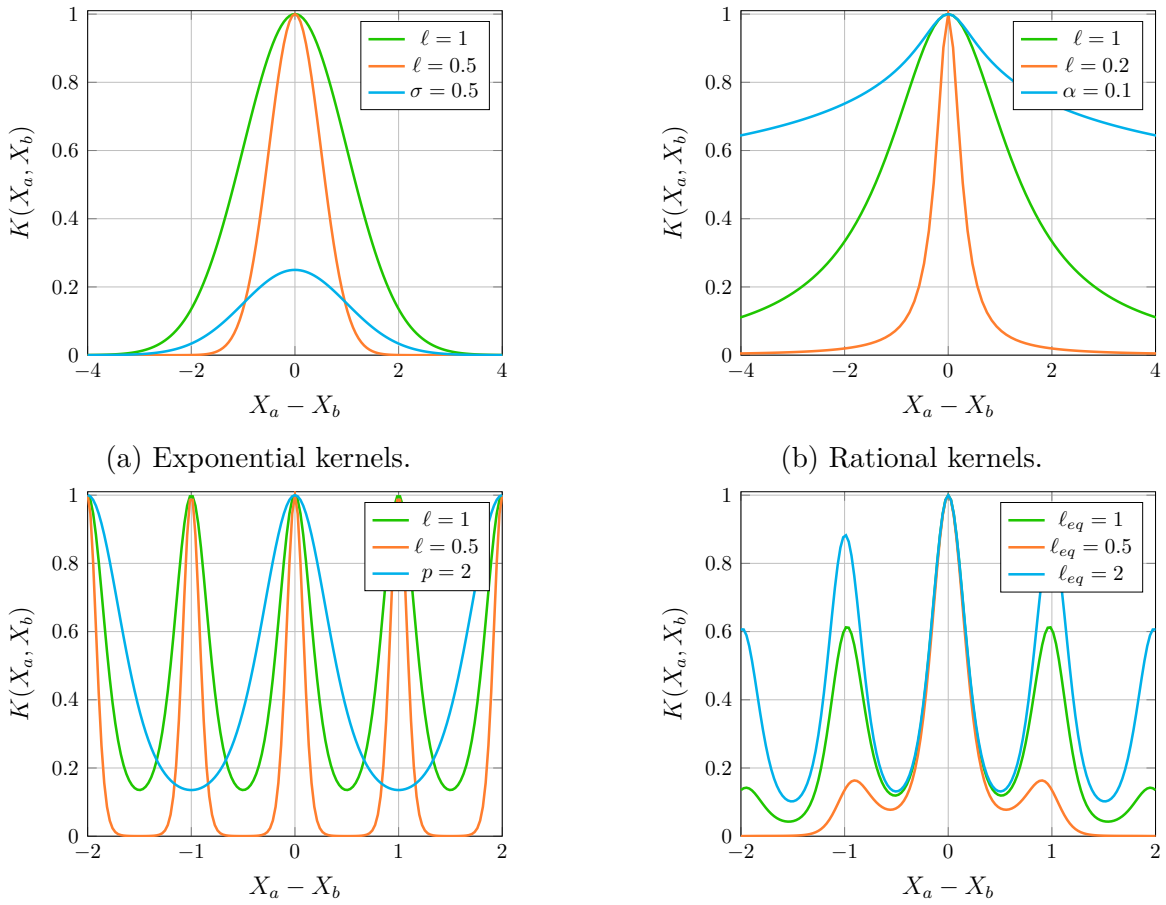
III.3.3.1 Noise free predictions

In this section the noise free prediction of a gaussian process surrogate, given a set of observations, is presented.

In this section, the predictions of a gaussian process surrogate on noise-free observations are presented. Basically, it explains how the function f can be interpolated by a surrogate gaussian process, based on the n already known samples.

With $D \subset \mathbb{R}^d$, let $f : D \mapsto \mathbb{R}$ be a black-box function to be predicted and suppose given n samples $\{x_i \mid i \in [[1, n]]\}$ in $(D)^n$ yielding the noise free responses $\{y_i \mid i \in [[1, n]]\}$ in $(\mathbb{R})^n$, *i.e.* we have, for a given $n \in \mathbb{N}^*$

$$\forall i \in [[1, n]], \quad y_i = f(x_i). \tag{III.39}$$



(a) Exponential kernels.

(b) Rational kernels.

(c) Periodic kernels.

(d) Local periodic kernels.

Figure III.7: Kernels of Table III.1 illustrated. If not specified, parameters value are set to 1.

First, we suppose that f is **a realization of a centered gaussian process** $\mathcal{Z} = (\mathcal{Z}_x)_{x \in D}$, with covariance kernel K *i.e.* we suppose that:

$$\exists \mathcal{Z},\ \exists \omega_0 \subset \Omega,\ \text{such as } \forall x \in D,\quad f(x) = \mathcal{Z}_x(\omega_0) \tag{III.40}$$

and that n events already occurred:

$$\forall i \in [[1, n]],\ \exists \omega_i \in \Omega,\quad \text{such as}\quad y_i = f(x_i) = \mathcal{Z}_{x_i}(\omega_i). \tag{III.41}$$

The prediction of $f(x_{n+1})$ for x_{n+1} in $D \backslash \{x_i \mid i \in [[1, n]]\}$ can then be rewritten as a real conditional random variable:

$$\mathcal{Z}_{x_{n+1}} \mid \mathcal{Z}_{(1,n)}, \tag{III.42}$$

where $\mathcal{Z}_{(1,n)}$ is the centered gaussian vector associated with the n known samples,

$$\mathcal{Z}_{(1,n)} \ := \ (\mathcal{Z}_{x_1},\ \dots, \mathcal{Z}_{x_n})^T, \tag{III.43}$$

with covariance matrix K_n, and where the T exponent marks the transpose operator.

In all generality, the law of the gaussian vector $\mathcal{Z}_{(1,\,n+1)}$ is given by definition as,

$$\mathcal{N}\left(\mathbf{0}_{n+1},\ K_{n+1}\right), \tag{III.44}$$

By rewriting K_{n+1} as a block matrix,

$$K_{n+1} \ = \ \begin{bmatrix} K_n & C \\ C^T & D \end{bmatrix} \tag{III.45}$$

where

$$C \ = \ \begin{bmatrix} K\left(x_1, x_{n+1}\right) \\ \vdots \\ K\left(x_n, x_{n+1}\right) \end{bmatrix}, \tag{III.46}$$

is a vector of size $(n,\ 1)$ and

$$D \ = \ K\left(x_{n+1}, x_{n+1}\right), \tag{III.47}$$

is real. The law of $\mathcal{Z}_{x_{n+1}} \mid \mathcal{Z}_{(1,n)}$ is directly known by applying Eq. III.27:

$$\mathcal{Z}_{x_{n+1}} \mid \mathcal{Z}_{(1,n)} \quad \sim \quad \mathcal{N}\left(C^T K_n^{-1} \mathcal{Z}_{(1,n)},\, D + C^T K_n^{-1} C\right), \tag{III.48}$$

Since we supposed that the n events defined by Eq. III.41 already occurred, the prediction of $f(x_{n+1})$ is a gaussian real variable, and it has for law:

$$\mathcal{Z}_{x_{n+1}} \quad \sim \quad \mathcal{N}\left(C^T K_n^{-1} Y_n,\, D + C^T K_n^{-1} C\right), \tag{III.49}$$

where

$$Y_n = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{III.50}$$

To summarize, the use of a gaussian process modeling allows to define, for each element of D, a corresponding gaussian variable whose mean and variance are depending on the kernel of the gaussian process and the observations. The two mean and variance function of the predictions are usually noted $\hat{\mu}$ and $\hat{\sigma}$. In the case of noise-free predictions, we have:

$$\begin{aligned} \forall x \in D, \quad & \hat{\mu}(x) = C^T K_n^{-1} Y_n, \\ & \hat{\sigma}(x) = D + C^T K_n^{-1} C. \end{aligned} \tag{III.51}$$

On Fig. III.8, the mean and variance prediction functions of a noise-free gaussian process, conditioned by eight observations, are shown for a prediction of the sin function.

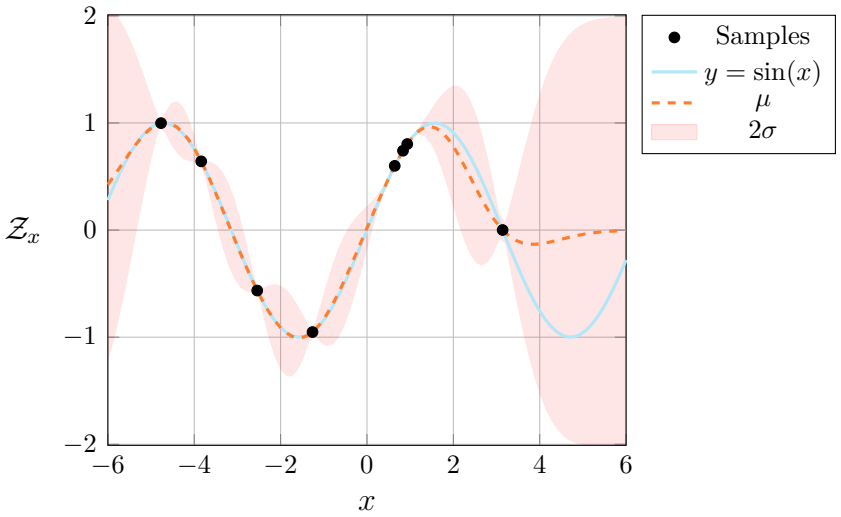


Figure III.8: Mean and variance function of a noise-free gaussian process, conditioned by eight observations. In this example, f is the usual sin function.

III.3.3.2 Noisy predictions

In this section, the predictions of a gaussian process surrogate on noisy observation are presented.

Firstly the noise in observation is taken into account by adding a small value to the function evaluation. We have,

$$\forall i \in [[1, n]], \quad y_i \; = \; f(x_i) \; + \; \varepsilon_i, \tag{III.52}$$

where $\varepsilon_i \in \mathbb{R}$ are supposed to be independent and identically distributed, following a centered gaussian variable of variance σ_ε, *i.e.* we supposed that:

$$\forall i \in [[1, n]], \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon). \tag{III.53}$$

Repeating the steps of section III.3.3.1, one gets, for each element of D, a prediction following a gaussian variable whose mean and variance are depending on the kernel of the gaussian process and the observations. In the case of noisy predictions, we have:

$$\begin{aligned} \forall x \in D, \quad \hat{\mu}(x) &= C^T \left(K_n + \sigma_\varepsilon I_n\right)^{-1} Y_n, \\ \hat{\sigma}(x) &= D + \sigma_\varepsilon + C^T \left(K_n + \sigma_\varepsilon I_n\right)^{-1} C. \end{aligned} \tag{III.54}$$

where Y_n is the column vector whose marginals are defined with Eq. III.52, C, D and K_n are identical to section III.3.3.1. On Fig. III.9, the mean and variance prediction functions of a noisy gaussian process, conditioned by five observations, are shown for a prediction of the sin function.

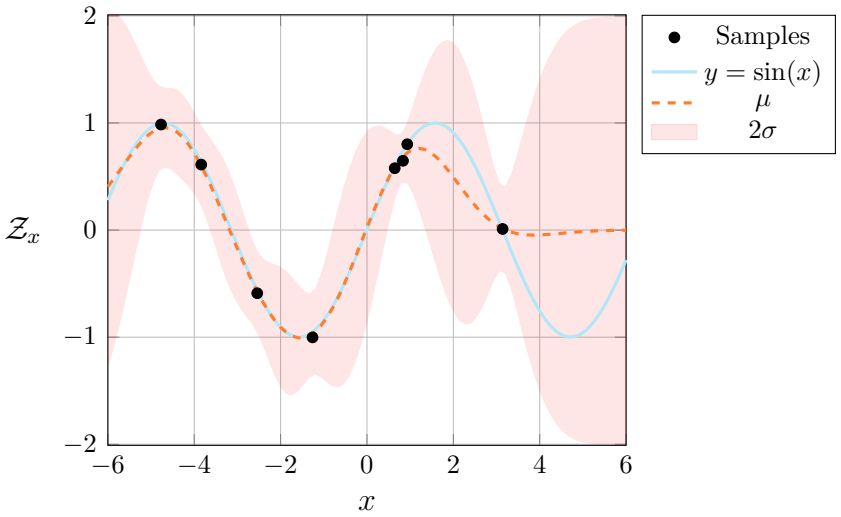


Figure III.9: Mean and variance function of a noisy gaussian process, conditioned by five observations. In this example, f is the usual sin function.

In the case of the EGO, the choice of considering noise-free or noisy predictions is included within the algorithm itself, and σ_ε is an hyperparameter of the gaussian process surrogate, identically to the covariance kernel hyperparameters defined in section III.3.2.5. In section III.3.4.2, the method to calibrate the hyperparameters is described.

III.3.3.3 Design of experiment

Previously in section III.3.3.1 and III.3.3.2, we supposed given the f function evaluation on a sample of the domain $D \subset \mathbb{R}^d$. We explain now how the sample, also referred as the **design of experiment** (DoE), is chosen. To simplify notations, we suppose for the current section that $D = [0, 1]^d$.

The simplest method to build a DoE is to select linearly spaced dots. This method is also referred as the uniform sampling methodology, and it is illustrated in Fig. III.10, where a 2D uniform sampling of 25 dots is performed on the unit square $[0, 1]^2$. Even if this method might seem interesting at first sight, it is not recommended. Indeed, the size of the DoE increases exponentially with the dimension of D. And f might be determined by a subset of D of lower dimension, leading to a large number of unnecessary function evaluations.

The standard method for building a DoE for a Bayesian optimiser is the Latin Hypercube Sampling (LHS) one [102]. It consists of dividing each dimension of D in n uniform intervals in order to mesh D, then each subinterval will contain a sample, for each dimension. An example of a 25 dots LHS of the unit square is shown in Fig. III.11. The uniform distribution within each dimension is clearly visible: each the five intervals of length 0.2 ($[0, 0.2]$, $[0.2, 0.4]$, $[0.4, 0.6]$, $[0.6, 0.8]$ and $[0.8, 1.0]$), on both dimensions, or axis, contain five sample dots. And more precisely, the 25 intervals of length $\frac{1}{25}$, on each dimension, or axis, all contain one sample dot.

In order to select among the $n!^{d-1}$ possible LHS, various criteria exist, such as a statistical criterion from [103] or a uniform criterion where the minimal distance between two observations is maximized (see [104]). In practice, only a few LHS are generated and then the best, according to one of the aforementioned criteria, is selected.

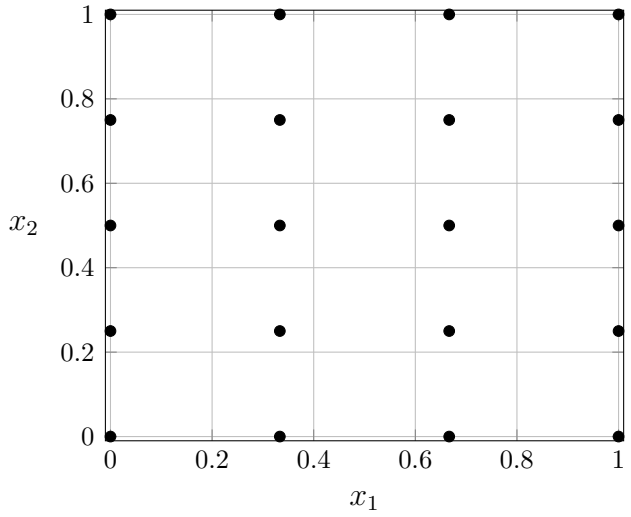


Figure III.10: Uniform 2D sampling of 25 dots of the unit square.

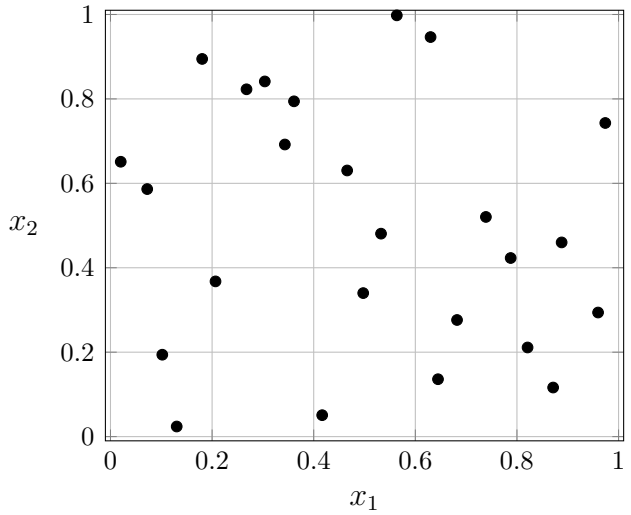


Figure III.11: 25 dots Latin Hypercube sampling.

The predictions of a black-box function with a gaussian process surrogate are now clarified for both noise-free and noisy observations. In the next section, the EGO algorithm is presented.

III.3.4 EGO

In this section, the EGO algorithm is presented, followed by its two specificities, the determination of the hyperparameters, and the merit function (or acquisition function). Finally a 1D example is provided.

III.3.4.1 Algorithm of EGO

This section summarizes the main steps of a global efficient algorithm. The inputs are:

- The parameter space, D, a subset of $\mathbb{R}^d$,
- The size of the DoE, n in $\mathbb{N}^*$,
- The covariance kernel, K,
- The merit function f_{merit}.

The output is the solution of the minimisation problem, x_{sol} in D. The steps are the followings:

- Construction of the LHS (see section III.3.3.3) of size n on the domain D.
- Evaluation of the objective function f on each dot of the LHS.
- For each iteration of the optimization loop,
  - Determine the hyperparameters of the covariance kernel K (see section III.3.4.2).
  - Determine the new observation x_{n+1} (see section III.3.4.3).
  - Compute the objective function f on the new observation x_{n+1}.
  - The value of $f(x_{n+1})$ is added to the DoE.

The ending criterion of the iterative optimization loop is either a number of iterations, or a lower bound the merit function must reach, or a standard deviation on the convergence of the gaussian surrogate model.

In the following sections, we describe, first, how the hyperparameters of the covariance kernel are determined, and then how the new observation is selected.

III.3.4.2 Determination of hyperparameters

At the starting point of the EGO algorithm, and for each iteration, the underlying gaussian process covariance kernel hyperparameters (see section III.3.2.5) must be calibrated. These hyperparameters can be determined by the, often costly, cross-validation method [105], or the maximum log-likelihood method [106], which is described in the following paragraphs.

The log-likelihood is usually used in order to avoid working with the law value of the density of probability. It consists in finding $(\sigma_\varepsilon, \Theta)$ that maximizes:

$$\mathcal{L}(\sigma_\varepsilon, \Theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left(\det\left(K_n + \sigma_\varepsilon I_n\right)\right) - \frac{1}{2}Y_n^T\left(K_n + \sigma_\varepsilon I_n\right)^{-1}Y_n, \quad \text{(III.55)}$$

where Y_n is defined by Eq. III.52, K_n by Eq. III.45 and σ_ε by Eq. III.53. Finding this maximum is by itself an optimization problem. For instance, in the DiceOptim R package [107] it is solved with the CMA-ES [108] global optimization algorithm.

The determination of the hyperparameters of the gaussian process surrogate is described. In the next section, the selection of the next evaluation point is clarified.

III.3.4.3 Merit function

The merit function, also called the acquisition function, aims to select a new observation $x_{n+1} \in D$, given the prediction on D of the gaussian model $\mathcal{Z}$, conditioned to the n first observations. More precisely, the merit function provides a criterion to define the best next observation, once given $\hat{\mu}(x)$ and $\hat{\sigma}(x)$, for all x in D. We present here three possible merit functions, more definitions are available and compared in [109].

Lower Confidence Bound (LCB). This first merit function was introduced by [110] and consists in minimizing the LCB defined as,

$$\forall x \in D, \qquad LCB(x) \;=\; \hat{\mu}(x) - \rho\hat{\sigma}(x), \quad \text{(III.56)}$$

where $\rho \in \mathbb{R}$ allows customizing the weights between the exploring and the simple prediction of the gaussian surrogate model. If $\rho = 0$, then the $LCB(x)$ is $\hat{\mu}$ and the minimization of the merit function is equivalent to the minimization of the prediction. Inversely if $\rho > 1$, then the exploration of the gaussian model is enhanced. The choice of ρ relies *in fine* on the user.

Probability of Improvement (PI). Introduced by [111], the PI is computed as,

$$\forall x \in D, \qquad PI(x) \;=\; \Phi\left(\frac{y_{min} - \hat{\mu}(x)}{\hat{\sigma}(x)}\right), \quad \text{(III.57)}$$

where y_{min} is the minimum on all previous observation, and Φ is the distribution function of the centered reduced normal law ($\mathcal{N}(0,\ 1)$). The main disadvantage of this merit function is that it tends to select dots close to the current minimum.

Expected Improvement (EI). This merit function was first introduced by [112] and more developed by [113]. It is expressed as the expectation of a r.v. on D, and defined as,

$$\forall x \in D, \qquad EI(x) \;=\; \mathbb{E}\left[\max\left(y_{min} - Q_x,\ 0\right)\right], \quad \text{(III.58)}$$

where y_{min} is the minimum on all previous observation, and Q_x is the r.v. of the prediction (see section III.3.3.1), *i.e.*,

$$\forall x \in D, \qquad Q_x \;:=\; \mathcal{Z}_x \mid \mathcal{Z}_{(1,n)}, \quad \text{(III.59)}$$

and, given the n previous observations, has for law $\mathcal{N}(\hat{\mu}(x), \hat{\sigma}(x))$ at $x \in D$ (see Eq. III.49). After a computation, involving an integration by part, one gets,

$$\forall x \in D, \qquad EI(x) \;=\; \hat{\sigma}(x)\left(u(x)\Phi\left(u(x)\right) + \phi\left(u(x)\right)\right), \tag{III.60}$$

where Φ (resp. ϕ) is the distribution function (resp. the density of probability) of the centered reduced gaussian law.

Since all merit functions must be minimized, the computation of the next observation implies using another optimizer in order to compute this minimum. For instance, for the DiceOptim R package, [107], this minimization problem is solved with the CMA-ES algorithm [108].

III.3.4.4 Simple 1D example of EGO

On Fig. III.12, a simple 1D example is provided. It consists in minimizing the 1D black function $f(x) = x\sin(x)$ on the interval $[-6, 6]$. Starting from a DoE of three dots, the six EGO iterations are illustrated.

On each subfigure of III.12, the blue curve is the function f. The dashed blue curve is the mean function of the underlying gaussian process, $\hat{\mu}$. The green area is the 2σ representation of the covariance function of the gaussian process, $\hat{\sigma}$ and the red curve is the EI (see Eq. III.58). Each iterations consists in firstly determining the hyperparameters of the gaussian model, resulting in the computation of $\hat{\mu}$ and $\hat{\sigma}$, and secondly minimizing the EI in order to get the next point to evaluate, which is marked in red. During the EGO iterations, one clearly sees the green area getting smaller, illustrating the decreasing of the uncertainties. Looking to the sixth iterations, one sees the majority of the function evaluation where performed toward the actual minimum of $4,8$ of the function of interest f.

The EGO is now presented, we focus in the rest of this chapter on the actual nanostructuration optimization results we achieved.

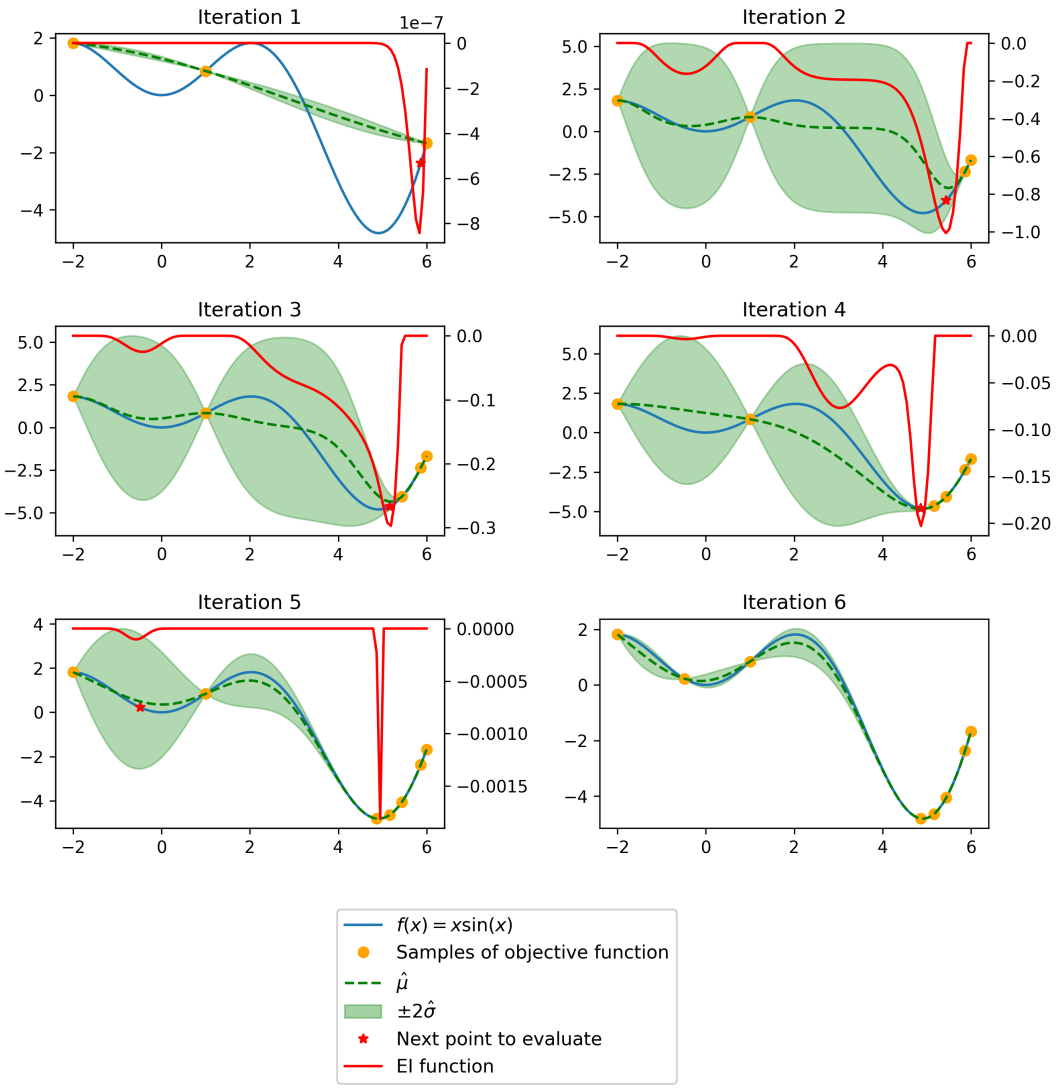


Figure III.12: Illustration of the EGO methodology on a simple 1D analytical function.

III.4 Grating optimization in 2D

In this section, we focus on the 2D optimization of grating parameters on a realistic SPADs pixel. For this purpose, we use the Matlab Bayesian algorithm coupled with our 2D in-house RCWA solver (see section II.5). As previously mentioned, we aim to identify the maximal absorption possible at 940 nm and the corresponding grating parameters using Bayesian optimization, rather than by using the usual parameters sweep, in order to minimize the number of total simulation runs.

Following conclusion of section III.2, this maximization problem is equivalent to finding the geometrical parameters that position a resonance exactly at 940 nm.

Firstly, the realistic SPAD geometry is defined, as well as its parametrization. Secondly, the choice of parameters of interest is clarified with a sensibility analysis. Thirdly, the optimization setup is described, including the number of DoE elements and EGO iterations, the range of the variables to optimize and the RCWA convergence analysis. Fourthly, the results of optimizations are presented.

III.4.1 Structure definition

In this section, a realistic 2D nanostructured SPAD structure is defined. All the following 2D grating optimizations are performed on this structure. In Fig. III.14, a structure example is provided.

In the z dimension, the structure is defined as a stack of layers, according to the RCWA requirements (see section II.4.3.1). From top to bottom, the layers are:

- An air layer of 3900 nm thick;
- An antireflective layer of TA2O5, surrounded by a Tungstène shield, 100 nm thick;
- A grating layer of thickness L_{depth}, including SiO2 DTI (Deep Trench Isolation). The grating trenches are made of SiO2;
- A Si layer of thickness $L_{epi} - L_{depth}$, including SiO2 DTI;
- A reflective Cu layer 200 nm thick.

Three components are defined on the x-axis: the tungsten shield in the antireflective layer, the DTI in both the grating layer and the Si layer, and the grating in the corresponding layer. Only the grating parameters will be optimized in this study, keeping the DTI and tungsten shield definitions constant.

On the x-axis, the DTI are $L_{dti} = 200$ nm thick and the DTI center is at a distance of $L_{rl,dti} = 500$ nm to the boundaries of the simulation. The tungsten shield is covering both the DTI and the outer Si, so it is at a distance of $L_{rl,dti} + L_{dti}/2 = 750$ nm to the x-boundaries.

As illustrated in Fig. III.13, the grating is defined by four parameters:

- Its depth, corresponding to the thickness of the grating layer, noted L_{depth};
- The size of one of its period, noted L_{pitch};

- Its number of period, noted nb_{pitch};
- Its fill factor, noted f, determining the size of the trench in one period of the grating. The fill factor is relative, so $f \in [0, 1]$.

In one period, we choose that a grating period ends with the SiO2 trench (see Fig. III.13). So, in order to center the trenches and to respect an equal distance between the SiO2 trenches and the left and the right DTI, an extra Si pillar of length $(1 - f)L_{Pitch}$ is added on the right of the grating. The total size of the grating, inside the DTI, in the x-axis, is:

$$L_{grating} = (nb_{Pitch} + 1 - f)L_{Pitch}, \tag{III.61}$$

and the total size of the whole structure, in the x-axis, is:

$$L_{total} = L_{grating} + L_{dti} + 2L_{rl,dti}. \tag{III.62}$$

Material permittivity used in this structure are given in Table III.2. The exact value at 940 nm is obtained by linear interpolation.

For this current work, in order to reduce the total number of parameters considered, we set two of the grating parameters, f and nb_{pitch}:

- $f = 0.5$, in order to work with a symmetric pattern, whose Si pillar and Si02 trenches are of equal size in the x axis. This symmetry hypothesis will be specially studied in the 3D grating optimization section below (see Sec. III.5).
- $nb_{pitch} = 10$, in order to work on a SPAD with a realistic L_{total}. Indeed, with $L_{pitch} \in [300, 650]$ nm, one gets $L_{total} \in [4350, 8025]$.

The realistic 2D SPAD structure being described, we can focus on the next section on the parameters selected for the optimization.

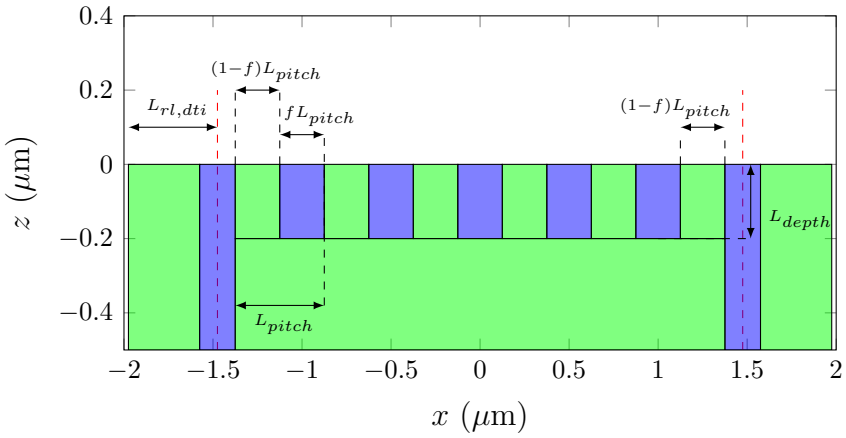


Figure III.13: Definition of L_{pitch}, f, $L_{rl,dti}$ and nb_{pitch} of the grating. In this example, nb_{pitch} is set to 5. Green (respectively blue) rectangles are made of Si (resp. SiO2). An extra Si pillar of length $(1-f)L_{pitch}$ is added on the right to ensure that the grating is centered, and that the distance between the trenches and the two DTI (left and right) are equal. In this example, we have: $L_{pitch} = 500$ nm, $f = 0.5$, $L_{rl,dti} = 500$ nm and $L_{depth} = 200$ nm. And thus, we have $L_{grating} = 2750$ nm and $L_{total} = 3950$ nm.

Name	Ref	ε_r	ε_i
Air	-	1	0
Si	Palik[114]	12.9507	0.0097
SiO2	Palik [114]	2.1060	0
W	Palik [114]	-0.1655	20.0019
TA2O5	Bright [115]	4.1612	0
Cu	Palik [114]	-43.4555	4.3978

Table III.2: Permittivity value at 940 nm and material references for the realistic SPAD 2D optimization. W refers to the tungsten.

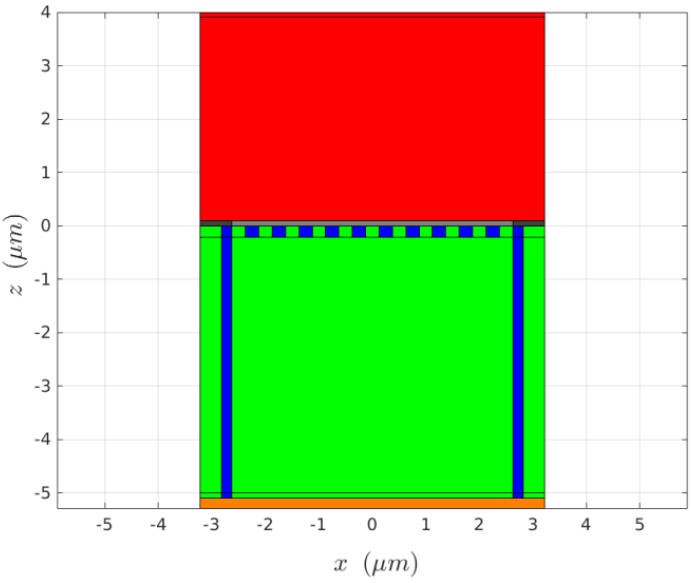


Figure III.14: Example of 2D nanostructured SPAD, with parameters, in μm: $L_{epi} = 5$, $L_{pitch} = 0.5$, $L_{depth} = 0.218$, $L_{dti} = 0.2$ and $L_{rl,dti} = 0.5$. nb_{pitch} is equal to 10. Red is air, black is tungsten, grey is TA2O5, blue is SiO2, green is Si and orange is Cu materials.

III.4.2 Parameters sensibility analysis

In this section, we present a preliminary study of the grating parameters impact on the inner Si absorption. This study will allow us to define the design space for the optimization.

The efficiency of a SPAD is correlated to the light absorption in its inner Si volume. In Fig. III.14, it corresponds to the green rectangle inside the DTI, including the grating. To get this figure of merit from a 2D RCWA simulation, one must compute the volumic absorption on the inner Si volume. Details on the implementation of the volumic absorption calculus from our in-house 2D RCWA software (see section II.5) are not provided to keep the current work concise.

So the figure of merit of interest is the absorption in the inner Si volume, at $\lambda = 940$ nm. In the rest of our work, it is noted A_{940}. The following parameters study consists in observing the influence of one grating parameter variation on A_{940}.

In the following, we set the structure parameters, if not varying, to $L_{epi} = 5$, $L_{pitch} = 0.5$, $L_{depth} = 0.218$. Firstly, one can observe on Fig. III.15a, as expected, the resonant and oscillating nature of the inner Si absorption according to the wavelength. The whole goal of our optimization is to position an absorption peak exactly at 940 nm.

Secondly, the L_{depth} influence on A_{940} is shown in Fig. III.15b. The maximum reached is 57% and the minimum reached is 14%. The variation of amplitude occurs on a wider range than the other parameters (see below). From this graph, we choose the range of interest as $[300, 600]$ nm.

Thirdly, the influence of L_{pitch} is shown in Fig. III.15c. This parameter exhibits multiple resonances of increasing amplitude, especially when $L_{pitch} < 940$. One must remark

that increasing L_{pitch} actually increases the ratio $\frac{L_{grating}}{L_{total}}$, leading to higher absorption. This is for instance visible in the increasing of the lower peak of Fig. III.15c. But despite the increasing of the ratio, the maximum absorption of 80% is obtained for $L_{pitch} = 685$ nm. This shows the well-known fact that the grating is actually diffracting the light, and thus enhancing absorption, only when its period is at the same magnitude as the wavelength of the incident light. Since 80% is the higher absorption reached within this parameter study, we choose for L_{pitch} the range $[300, 650]$ nm to exclude this maximum and to challenge the Bayesian optimization to perform equally or better than this maximum.

Fourthly, the influence of L_{epi} is shown in Fig. III.15d. The multiple resonances allow a maximum absorption of 66%. We choose for the optimization the range $[4.5, 7]$ μm.

Finally, this preliminary study allows us to define the range of interest for the grating parameters L_{depth}, L_{pitch} and L_{epi}, and to observe the influence of single parameters variations on A_{940}. In the next section the optimization setup is described.

III.4.3 Optimization setup

In this section, the optimization setup for the 2D grating optimization is described, including the Bayesian algorithm parameters and the RCWA convergence analysis.

As mentioned in section III.3.4.1, the EGO method requires five initial parameters: the parameter space, noted D, the size of the DoE, a covariance kernel K, a merit function, and a stopping criterion.

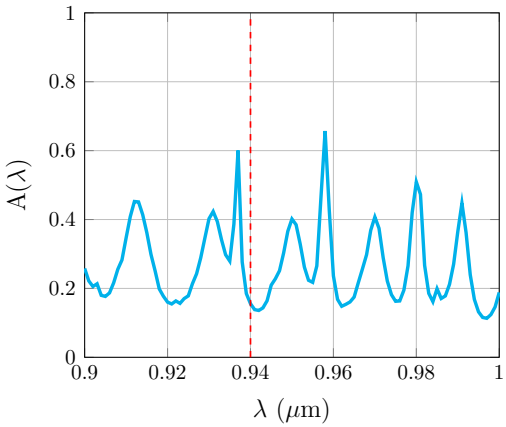
For the optimization of the inner volumic absorption at $\lambda = 940$ nm, on the realistic 2D nanostructured SPAD described above (see section III.4.1), we choose the following input parameters:

- K is a square exponential kernel (see Tab. III.1);
- The DoE contains 40 elements;
- The merit function is the Expected Improvement (see Eq. III.58);
- The stopping criterion is the number of EGO iterations, set to 460.

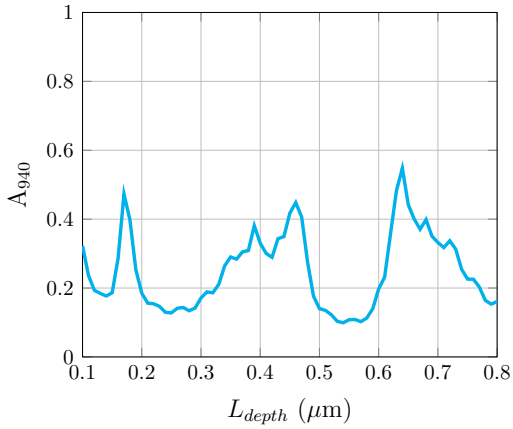
We define two optimizations setups, depending on the number of parameters to optimize. The parameters range selected correspond to the range of interest defined in section III.4.2.

The first setup, denoted as **setup 1**, is an optimization on both (L_{pitch}, L_{depth}), on the space $D_1 = [300, 650] \times [300, 600]$ nm^2, keeping $L_{epi} = 6.8$ μm constant.

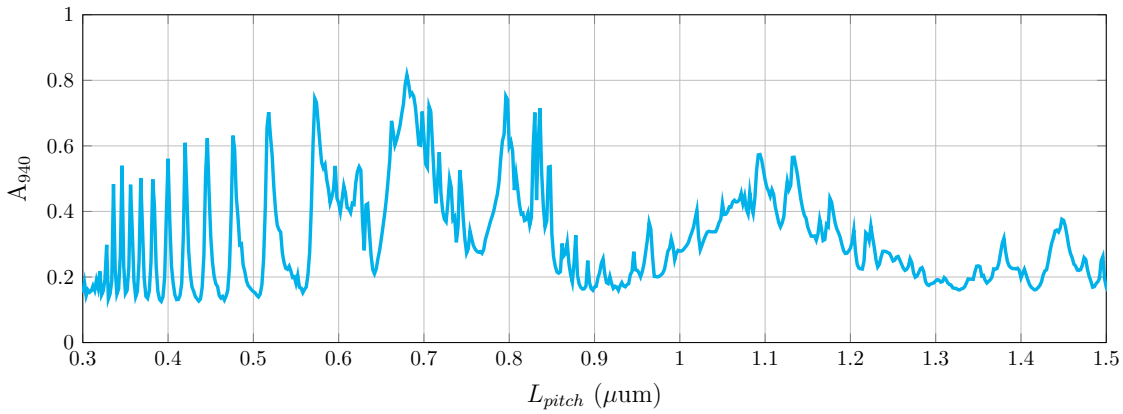
The second setup is an optimization on three parameters, (L_{pitch}, L_{depth}, L_{epi}) on the space $D_2 = [300, 650] \times [300, 600] \times [4500, 7000]$ nm^3. This setup is denoted as **setup 2**.



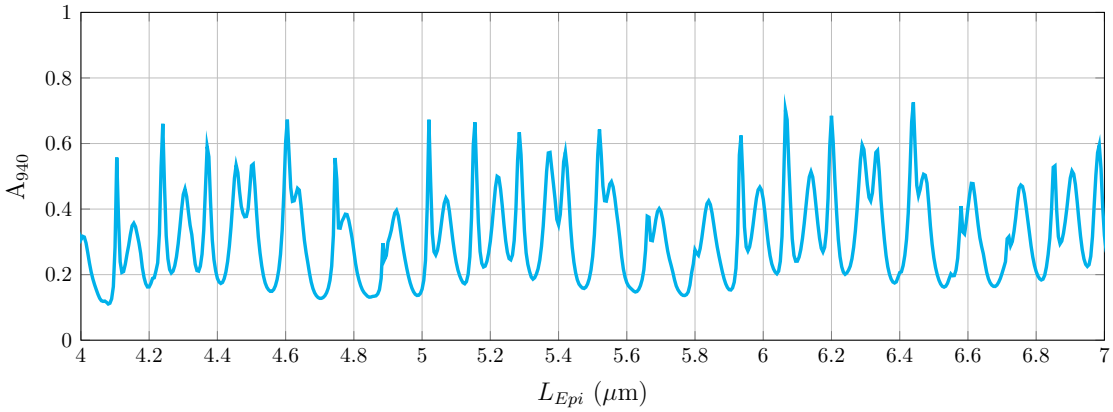
(a) Wavelength sweep, 101 dots.



(b) L_{depth} sweep, 71 dots.



(c) L_{pitch} sweep, 851 dots.



(d) L_{Epi} sweep, 1001 dots.

Figure III.15: RCWA 2D computation of the inner Si absorption, as a function of various parameters, of the structure described in section III.4.1. If not varying, we have $L_{epi} = 5$ μm, $L_{pitch} = 500$ nm and $L_{depth} = 218$ nm. A_{940} denotes the inner Si absorption at 940 nm. In each subcation, "X dots" indicates the number of equally spaced x-axis dots that were simulated.

In order to perform this optimization, we use a slightly modified version of the *bayesopt*[3] function from the Matlab Global Optimization toolbox. The modification was performed on the inner Gaussian process modeling in order to custom the underlying kernel function and to accept a squared exponential kernel.

The convergence study of the 2D RCWA solver, for the structure described in section III.4.1, with parameters identical to Fig. III.14, is shown on Fig. III.16. One can see that 201 plane waves is enough. For the optimization, since L_{total} might be higher, we select a conservative truncation of 401 plane waves. The simulation time for 401 plane waves truncation is approximately 1 min on a single CPU.

Finally, the input parameters for the Bayesian optimization are defined, the corresponding results of both **setup 1** and **setup 2** are provided in the next section.

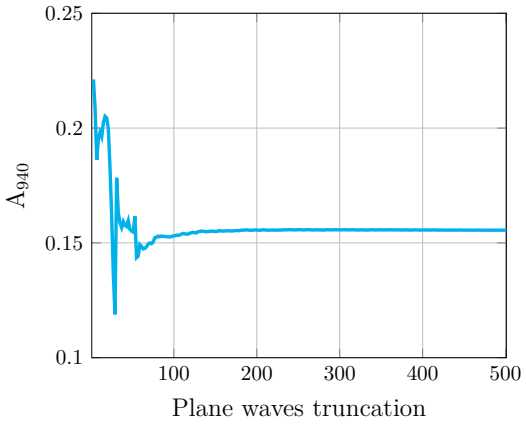


Figure III.16: RCWA 2D convergence study. A_{940} is the inner Si absorption at $\lambda = 940$. On the x-axis is the number of plane waves truncation. 250 simulations were run: all odd numbers from 3 to 501.

III.4.4 Optimization results

In this section, the results of the optimization setups defined in the above section are presented.

III.4.4.1 Setup 1

In this section, the results of the optimization **setup 1** are presented.

In Fig. III.17, the objective function for each iterations of both the DoE and the EGO phases are shown. The blue dots represent the 40 DoE evaluation while the red dots represent the 460 EGO iterations. The maximum reached, visible on Fig. III.17 as a black triangle, and the corresponding optimized parameters are:

$$A_{940,max,1} = 0.77527, \qquad \text{with} \quad (L_{pitch,max,1}, L_{depth,max,1}) = (643.53, 356.43) \text{ nm}^2. \tag{III.63}$$

[3] `https://fr.mathworks.com/help/stats/bayesopt.html`

The absorption profile of the optimized structure is visible on Fig. III.18. The peak at the 940 nm wavelength is clearly visible and the optimization thus accomplished the desired goal in positioning a resonance exactly at 940 nm.

In order to extract not only the maximum reached but also the distribution of the best performing parameters, we represent, in Fig. III.19, the classification of the evaluated parameters according to the value of the absorption reached. Basically, one first defines the following three parameters classes:

$$\begin{aligned} C_1^1 &= \left\{ (L_p, L_d) \mid g\left(L_p, L_d\right) \in [c_1^1, c_2^1] \right\}, \\ C_2^1 &= \left\{ (L_p, L_d) \mid g\left(L_p, L_d\right) \in [c_2^1, c_3^1] \right\}, \\ C_3^1 &= \left\{ (L_p, L_d) \mid g\left(L_p, L_d\right) \in [c_3^1, c_4^1] \right\}, \end{aligned} \tag{III.64}$$

where $g\left(L_p, Ł_d\right)$ denotes the black-box function, *i.e* the inner volumic absorption computed by the 2D RCWA solver on the structure with parameters $(L_{pitch}, L_{depth}) = (L_p, L_d)$, and where,

$$\begin{aligned} c_1^1 &= 0.8, \\ c_2^1 &= 0.75, \\ c_3^1 &= 0.725, \\ c_4^1 &= 0.7. \end{aligned}$$

The distribution of classes C_1^1, C_2^1 and C_3^1 can be seen by plotting these classes on a $L_{pitch} \times L_{depth}$ graph, or more precisely on the search space D_1. as done in Fig. III.19. In Fig. III.19b, which is already a zoom (recalling that $D_1 = [300, 650] \times [300, 600]$ nm^2), the distribution of C_1^1, C_2^1 and C_3^1 is clearly concentrated in a small region of the whole desing space. It appears that the coupling between the resonances in L_{depth} and L_{pitch}, that were observed in Fig. III.15b and Fig. III.15c, is maximal only in the subspace $D_{1,sub} := [640, 650] \times [340, 440]$ nm^2 of D_1. This means, supposing that L_{pitch} and L_{depth} are searched within D_1 and that $L_{epi} = 6.8$ μm, that achieving at least 70% of absorption can only be done within $D_{1,sub}$. Furthermore, the maximal absorptions are obtained not only on a square within D_1, but rather on a line, as clearly visible in Fig. III.19b. For informations, recalling that 500 iterations were performed in total, the cardinal of C_1^1, C_2^1 and C_3^1 are:

$$\#C_1^1 = 48, \qquad \#C_2^1 = 41, \quad \text{and} \quad \#C_3^1 = 53. \tag{III.65}$$

Finally, the Bayesian optimization not only achieved to determine a maximum of 77% absorption at 940 nm, but it also managed to generate 142 designs that achieve an absorption of more than 70%. So the topology of the resonance, and the neighbor of the reached maximum, are known. Thus we obtained both one maximum and a qualitative information on the resonance, in spacial distribution over the design space.

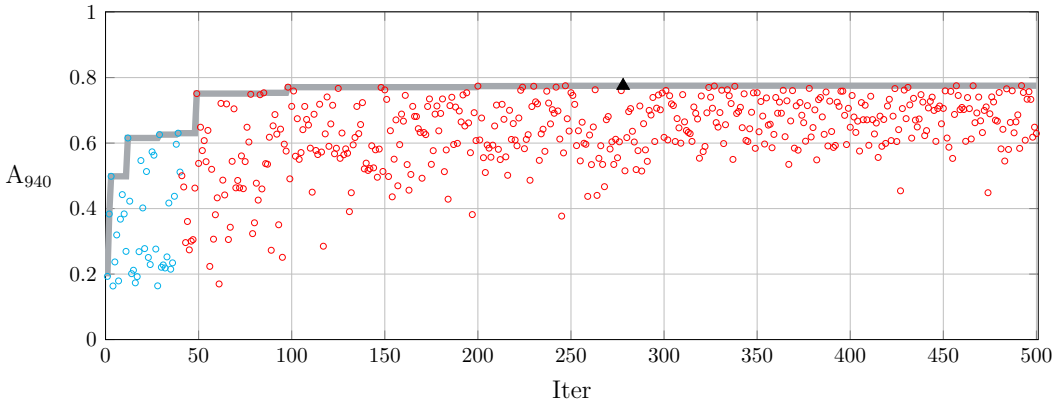


Figure III.17: Objective function according to the iterations, for the Bayesian optimization of the **setup 1** (defined in section III.4.3). DoE (resp. EGO) iterations are displayed in blue (resp. red). The maximum reached is 77% and is marked as a black triangle, with parameters given in Eq. III.63. The gray line is the maximum reached during the optimization.

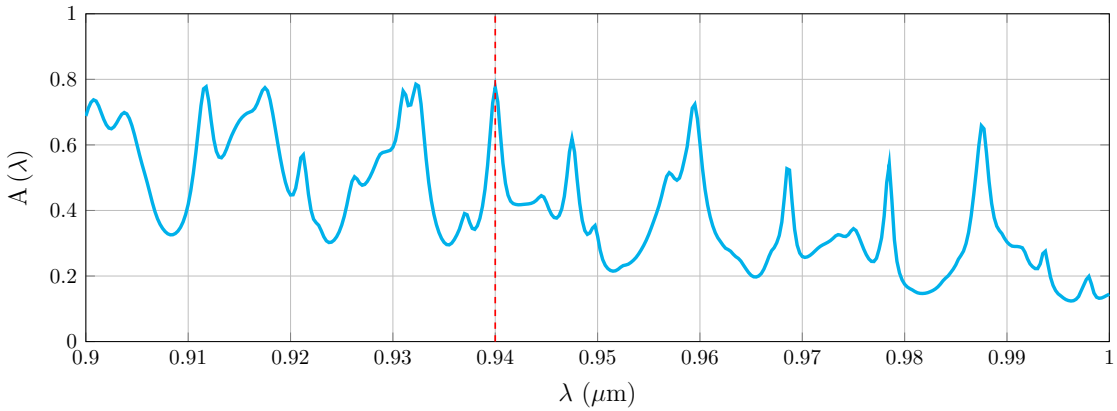


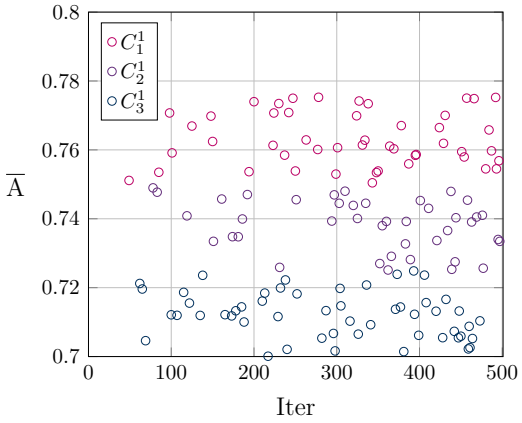
Figure III.18: Absorption profile of the best design for the **setup 1**. The optimal parameters are given in Eq. III.63. In order to catch the exact profile, 401 equally spaced wavelengths between 900 and 1000 nm are simulated.

III.4.4.2 Setup 2

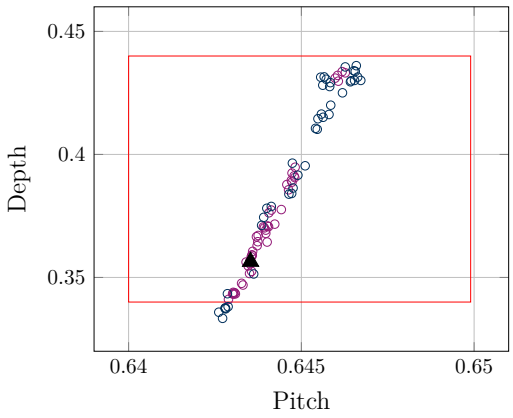
In this section, the results of the optimization **setup 2**, defined in section III.4.3, are presented.

In Fig. III.20, the objective function values for each iteration of both the DoE and the EGO phase are shown. The maximum reached, visible on Fig. III.20 as a black triangle, and the corresponding parameters are:

$$\begin{aligned} A_{940,max,2} &= 83.953\%, \\ L_{pitch,max,2} &= 621.195017622117 \quad \text{nm}, \\ L_{depth,max,2} &= 283.694395416066 \quad \text{nm}, \\ L_{epi,max,2} &= 4643.31087203558 \quad \text{nm}. \end{aligned}$$



(a) Definition of C_i^1.



(b) C_i^1 distribution.

Figure III.19: Definition and distribution of classes C_i^1 defined by Eq. III.64. On the left, an illustration on C_i^1 definition is shown. On the right, the distribution of C_i^1 is shown. The figure on right is already a zoom on D_1 defined in section III.4.3. The red square is the area of interest chosen for the further analysis (see section III.4.5).

Since parameters defined at a precision of 10^{-21} m are actually physically meaningless, the absorption profile for the best design (respectively the best design with truncated parameters) is shown in Fig. III.21 in blue (resp. orange). As visible on this figure, the truncated optimal parameters only provide an absorption of 75.558%, illustrating the high sensitivity of the resonance peak in all three parameters considered, and confirming the observation of section III.4.2.

Similarly to the presentation of the optimization results of **setup 1**, we investigate the distribution of the best performing parameters by defining the following classes:

$$\begin{aligned}
C_1^2 &= \left\{(L_e, L_p, L_d) \mid g\left(L_e, L_p, L_d\right) \in [c_1^2, c_2^2]\right\}, \\
C_2^2 &= \left\{(L_e, L_p, L_d) \mid g\left(L_e, L_p, L_d\right) \in [c_2^2, c_3^2]\right\}, \\
C_3^2 &= \left\{(L_e, L_p, L_d) \mid g\left(L_e, L_p, L_d\right) \in [c_3^2, c_4^2]\right\}, \\
C_4^2 &= \left\{(L_e, L_p, L_d) \mid g\left(L_e, L_p, L_d\right) \in [c_4^2, c_5^2]\right\},
\end{aligned} \tag{III.66}$$

where $g\left(L_e, L_p, Ł_d\right)$ denotes the black-box function, *i.e* the inner volumic absorption computed by the 2D RCWA solver on the structure, with parameters $(L_{epi}, L_{pitch}, L_{depth}) = (L_e, L_p, L_d)$, and where,

$$\begin{aligned}
c_1^2 &= 1.0, \\
c_2^2 &= 0.8, \\
c_3^2 &= 0.75, \\
c_4^2 &= 0.725, \\
c_5^2 &= 0.7.
\end{aligned}$$

In Fig. III.22, both the definition of C_i^2 and their distribution are shown. Similarly to the results of **setup 1**, the higher absorption designs are spread within determined subspaces

of the search space D_2. More precisely, the optimal design (with parameters given by Eq. III.66) lies in the resonance regions $D_{2,sub\,1}$, while a wider regions, noted $D_{2,sub\,2}$, contains the majority of the C_i^2 classes members, where we defined:

$$D_{2,sub\,1} = [4500, 4700] \times [610, 625] \times [270, 300] \quad \text{nm}^3, \tag{III.67}$$

$$D_{2,sub\,2} = [6000, 6500] \times [620, 645] \times [250, 380] \quad \text{nm}^3, \tag{III.68}$$

and where $D_{2,sub\,i}$ are described on the triplets $(L_{epi}, L_{pitch}, L_{depth})$. For informations, the cardinal of the classes C_i^2 are:

$$\#C_1^2 = 6, \qquad \#C_2^2 = 43, \qquad \#C_3^2 = 45, \quad \text{and} \quad \#C_4^2 = 66. \tag{III.69}$$

Finally, the optimization on three parameters not only provided a higher maximum than 80% initially found in section III.4.2 but also a dispersion of the higher performing designs, allowing a deduction of the area of high coupled resonances. And despite the maximum of 83% corresponds to an unphysical precision, the Bayesian algorithm successfully found more than 49 designs achieving an absorption higher than 75%.

In the next section, we perform a further analysis of the optimization results of both **setup 1** and **2**, trying to take advantage of the distribution observed on the best performing parameters.

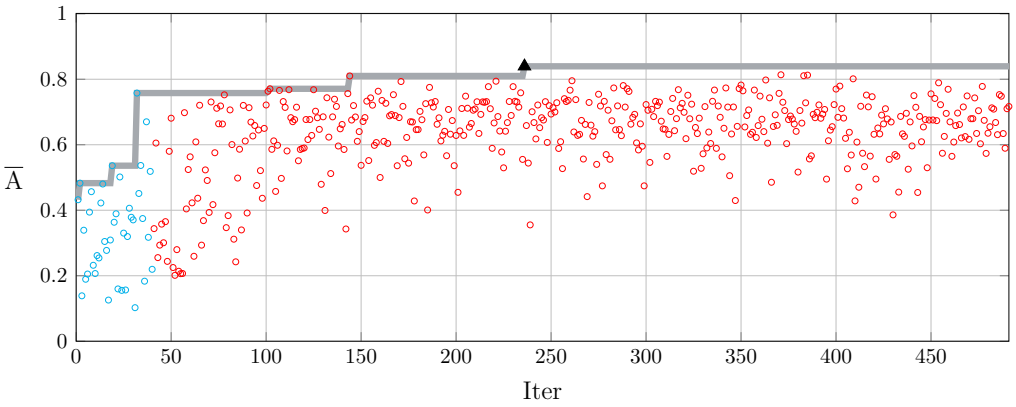


Figure III.20: Objective function according to the iterations, for the Bayesian optimization of the **setup 2** (defined in section III.4.3). DoE (resp. EGO) iterations are displayed in blue (resp. red). The maximum reached is 83% and is marked as a black triangle, with parameters given in Eq. III.66. The gray line is the maximum reached during the optimization.

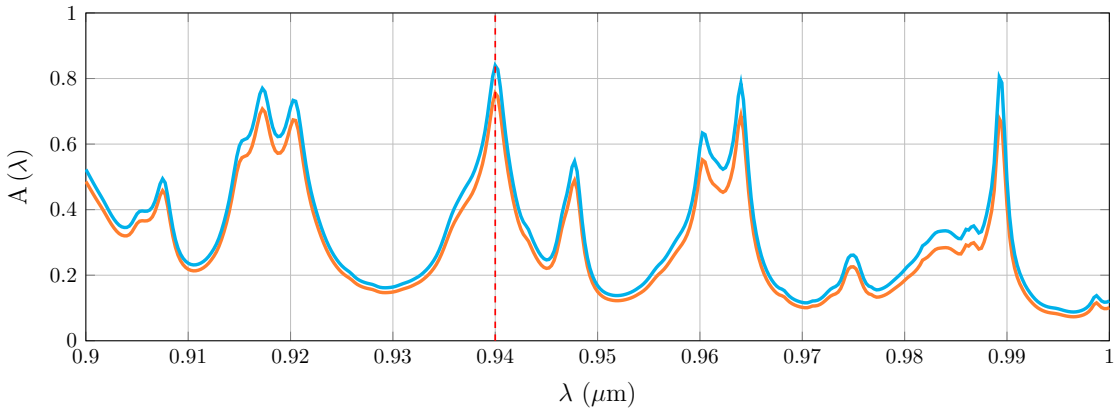
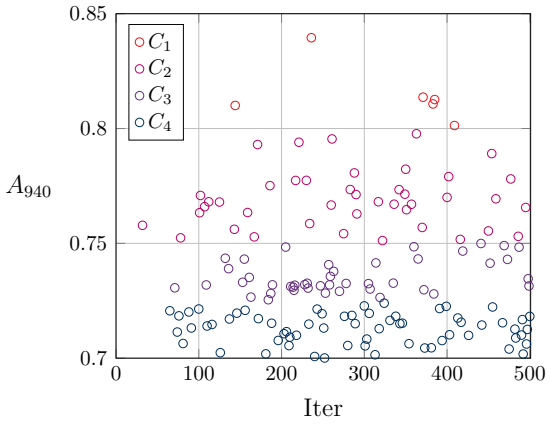
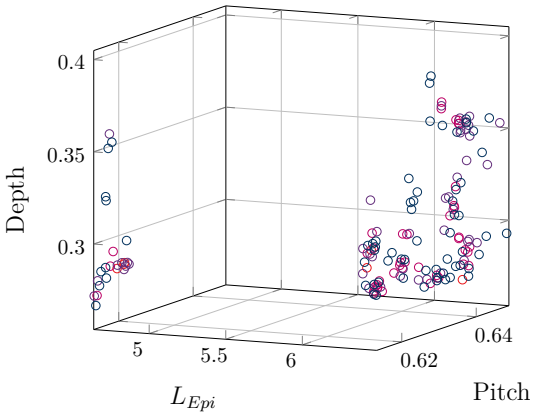


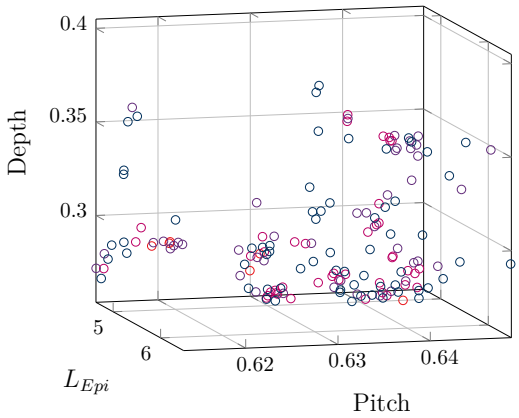
Figure III.21: Absorption profile of the best design (blue curve) for the **setup 2**. The optimal parameters are given by Eq. III.66. The orange curve is the absorption profile of the best design whose parameters are truncated, *i.e.* with parameter $(L_{epi},\ L_{pitch},\ L_{depth}) = (4643.3,\ 621.2,\ 283.6)$ nm^3. In order to catch the exact profile, 401 equally spaced wavelengths between 900 and 1000 nm are simulated.



(a) Definition of C_i.



(b) C_i distribution.



(c) C_i distribution.

Figure III.22: Definition and distribution of classes C_i^2 defined in section III.4.3. On top, an illustration on C_i^2 definition is shown. On bottom, the distribution of C_i^2 is shown. The figures on bottom are already a zoom on D_2 defined in section III.4.3.

III.4.5 Further analysis

In this section, we investigate the optimization results presented in the previous section. More precisely, we first test if the optima reached with the **setup 1** (resp. **2**) can be improved, by refining the searching space to $D_{1,sub}$ (resp. $D_{2,sub\,2}$). Secondly, we perform a linear sweep on $D_{1,sub}$ in order to visualize the exact neighborhood of the best performing designs found for **setup 1**.

III.4.5.1 Optimization on a finer parameters space

Setup 1

Since the optimum reached for the **setup 1** and all the best performing designs are within the subspace $D_{1,sub}$ of D_1, this second optimization aims to check if the optimum reached is actually an optimum, *i.e.* if it can be improved.

In Fig. III.23, the objective function for each iteration of both the DoE and the EGO phases are shown. The blue dots represent the 40 DoE evaluation while the red dots represent the 460 EGO iterations. The maximum reached, visible on Fig. III.23 as a black triangle, and the corresponding optimized parameters are:

$$A_{940,max,1} = 0.77508, \qquad \text{with} \quad (L_{pitch,max,1}, L_{depth,max,1}) = (643.63, 359.6)\ \text{nm}^2. \tag{III.70}$$

This optimum is equal to 77%, and so it is identical to the one reached for the first corresponding optimization, whose results are presented in section III.4.4.1. Actually, the only difference lies in the dispersion of the objective function to the maximum: comparing Fig. III.17 and Fig. III.23, one clearly sees that the objective function is closer to the maximum of 77% on Fig III.23.

Finally, this second optimization on the finer search space $D_{1,sub}$ confirms that the first optimization on D_1 found the optimum of 77% absorption at 940 nm.

Setup 2

Since the optimum reached for the **setup 2** is within $D_{2,sub\,1}$ while the best performing design is inside the subspace $D_{2,sub\,2}$, this second optimization aims to check if the optimum reached is a **global** optimum.

In Fig. III.24, the objective function values for each iteration of both the DoE and the EGO phase are shown. The maximum reached, visible on Fig. III.24 as a black triangle, and the corresponding parameters are:

$$\begin{aligned} A_{940,max,2} &= 80.714\%, \\ L_{pitch,max,2} &= 649.01 \quad \text{nm}, \\ L_{depth,max,2} &= 371.55 \quad \text{nm}, \\ L_{epi,max,2} &= 6352.7 \quad \text{nm}. \end{aligned} \tag{III.71}$$

This optimum is lower than the 83% reached in the first corresponding optimization on D_2, confirming that the optimum was reached in this first optimization. Similarly to the optimization on $D_{1,sub}$, the dispersion of the objective function, for the optimization on the finer space $D_{2,sub\,1}$, is lower on Fig. III.24 than the dispersion of the objective function of the optimization on the larger space D_2, visible in Fig. III.21.

Finally, this second study shows that the optimum previously reached is the global optimum.

III.4.5.2 Setup 1 complete response

In this section, we aim to visualize on $D_{1,sub}$ the neighborhood of the optimum reached for the first optimization done on D_1. Basically, we select 21 linearly interpolated L_{pitch} in the $[640, 650]$ nm interval, and 101 linearly interpolated L_{depth} on the $[340, 440]$ nm interval, and we compute the objective function, namely A_{940}, at each parameters couple of this grid. The absorption on such a grid is shown on Fig. III.25 as well as the best performing designs found by the optimization on D_1 (on Fig. III.27), or on $D_{1,sub}$ (on Fig III.26). The main observation are:

- The linear sweep shows the ridge (or line) of coupling resonances that was inducted by the first optimization results on section III.4.4.1;
- This ridge of high absorption is itself composed of gaps and peaks, as visible on Fig. III.25;
- This ridge of high absorption exhibits the higher absorption on the $[340, 380]$ nm L_{depth} interval, as visible on Fig. III.25;
- The first optimization on D_1 identified both this ridge of high absorption and its maximum in the $[340, 380]$ nm L_{depth} interval, as visible on Fig. III.27;
- The 60 best performing designs of the optimization on $D_{1,sub}$ are positioned on this ridge, as visible in Fig. III.26.

The total number of simulations of this linear sweep, on $D_{1,sub}$, is $21 \times 11 = 2121$. This is four times the number of simulations performed in the first optimization on the larger space D_1. This simple comparison clearly shows that linear sweeps are not efficient and relevant for finding coupling resonances. This also justifies why such study cannot be realistically performed on the **setup 2**, due to the number of needed simulations being too high.

Finally, this closer study, on the optimum reached for the **setup 1**, shows that the Bayesian optimization allows to induce the high coupling resonances area. Considering the high number of simulations run on this linear sweep, the Bayesian optimization accomplished such results by minimizing the number of cost function evaluations.

III.4.6 Conclusion on 2D structure optimization

The initial problem that motivates these optimizations of a 2D realistic SPAD was to identify the geometrical parameters of interest and to optimize the absorption at 940 nm, and so position a resonance exactly at 940 nm.

The two optimization setups (defined in III.4.4.1 and III.4.4.2) found optimal designs that exhibit a resonance exactly at 940 nm (see Fig. III.18 and Fig III.21), achieving at most an absorption of 77% for the **setup 1** and 83% for the **setup 2**. In view of the reached absorption, and the corresponding absorption profile, we consider our initial problem solved.

Treating the inner Si absorption as a black-box function, and optimizing it with EGO, allowed us to perform better than the 47% absorption found in [82]. Also, our methodology finds an optimum with lower numerical cost compared to a usual linear parameters sweep (see section III.4.5.2).

Furthermore, using the Gaussian Process surrogate allows to identify the area of coupling resonances on the searched parameters space. Combining the best performing design of **setup 1** and **setup 2**, the prediction of the underlying Gaussian Process allows to generate more than 97 ($\#C_1^1 + \#C_1^2 + \#C_2^2 = 97$) designs presenting an absorption higher than 75%. This efficiency of the use of gaussian process surrogates must be remembered for further studies. In particular, the absorption exhibiting multiple coupled resonances (see Fig. III.15), using a local optimizer, such as conjugate gradient for instance, is not adapted, and thus it must be avoided.

About the numerical optical solver, the absorption high sensitivity on the geometrical parameters, as shown in Fig. III.21, points out the importance of using a conformal numerical method, *i.e.* a numerical method that simulate exactly the structure considered. The FDTD, for instance, using a cartesian mesh, intrinsically introduces approximation on the structure interfaces, and thus it adds a crucial error. On the contrary, the DGTD method, using a conformal mesh, avoids *a priori* such errors. RCWA is well suited as long as all the z structure interfaces are normal to the x-axis (respectively the x-y plane) for 2D simulations (resp. 3D simulations). In this grating optimization of 2D SPADs, since all the z-interfaces are normal to the x-axis, we initially avoided such errors.

Finally, we can *a posteriori* claim that the parameter sweep in L_{pitch} is the most interesting for investigating the maximum, since it exhibits, in Fig. III.15c, an absorption peak at 940 nm of the same magnitude as the optimum found for both **setup 1** and **setup 2**.

The main results of the 2D grating optimization of a realistic SPAD are now recalled. The next paragraphs aims to identify all the limitations of this 2D optimization.

About the SPAD geometry considered, we claimed that the structure of Fig III.14 is a realistic SPADs since it includes DTIs, and presents a realistic L_{total} and L_{epi}. However, we excluded two structural elements that must be taken into account to optimize a structure closer to the real device, *i.e.*, the lens and the non-perfect metal reflector. Firstly, a lens, by focusing the light inside the SPADs, has a great influence on the grating response, and it should be included in a complete optimization, for instance by varying its radius of curvature. Secondly, the bottom Cu layer of our SPADs (see Fig. III.14) is a perfect metal reflector, while in reality the carrier collection imposes the presence of electrodes

that prohibits the use of a perfect reflector. Thus the use of a partial Cu layer at the bottom of our structure should be considered.

On polarization, our study is performed entirely in the TM mode. A complete optimization should average the TM and TE polarization.

On the optimum reached 83%, we showed in Fig. III.21 that the corresponding optimum design with truncated parameters reaches only an absorption of 75%, proving the high geometrical sensitivity of such optimum. This inherently limits the race to the higher absorption design, since the optimum actually lies in physically meaningless geometrical variation. However, we briefly presented how the process of surrogate modeling can provide more information than a single optimum: the EGO by itself successfully identified more than 97 designs exhibiting an absorption higher than 75%. Actually, the use of gaussian process surrogate is well suited in order to perform both sensibility and classification analysis. In our work, we presented only a simple classification from the EGO evaluations, but exact classification methodology using gaussian processes are available, for instance in [116, 117]. About the sensitivity analysis, gaussian process surrogates are also well-suited and already used in literature, for instance in [118, 119, 120]. We limited our work to EGO, but both the classification and the sensitivity analysis from gaussian process surrogate seems promising.

On the parameters of the gaussian process surrogate model, we used only gaussian kernels and the choice of this kernel was not investigated. Since the gaussian process kernel contains all the *a priori* information we have on the black-box function, *i.e.* the absorption at 940 nm, and since we showed that the absorption exhibits multiples resonances in L_{pitch} and L_{epi}, the use of a periodic kernel, or local periodic kernel (see table III.1) seems promising.

On the DTI material and parameters, we fixed $L_{rl,dti} = 500$ nm, $L_{dti} = 200$ nm and the use of SiO2 materials. Even if an appendix (see appendix A) investigated the use of an absorbing material for DTIs, we did not investigate the influence of the DTI thickness.

More globally, the grating parameters f (see Fig. III.13) is kept equal to 0.5. Thus, we studied only symmetrical gratings, while we know from [88] that the use of unsymmetrical grating enhances better absorption. This limitation of the 2D grating optimization is treated in the 3D grating optimization section below III.5.

Finally, optimizing only a 2D structure intrinsically limits the scope of the optimum found, since 3D geometry brings specific difficulties. The first of these difficulties lies in the choice of the grating shapes considered, and it is investigated in the next section on 3D grating optimization III.5.

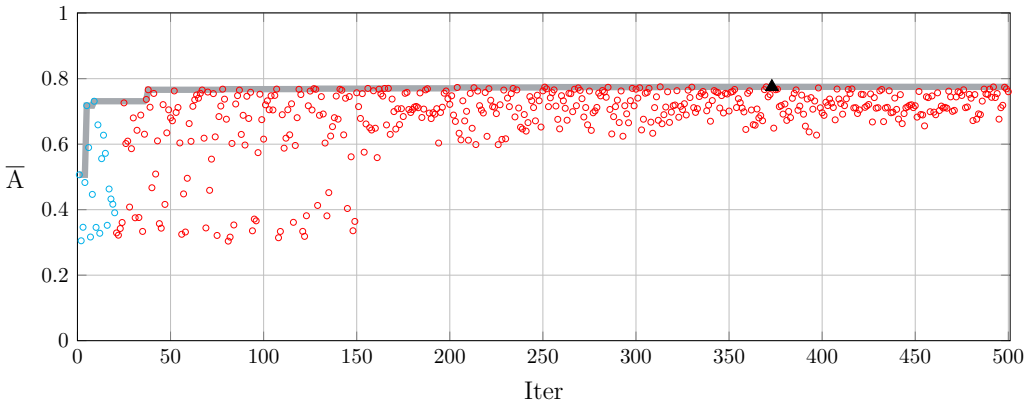


Figure III.23: Objective function according to the iterations, for the Bayesian optimization of the **setup 1** (defined in section III.4.3), on the parameters space $D_{1,sub}$ (defined in section III.4.4.1, and illustrated in Fig III.19b). DoE (resp. EGO) iterations are displayed in blue (resp. red). The maximum reached is 77% and is marked as a black triangle, with parameters given in Eq. III.70. The gray line is the maximum reached during the optimization.

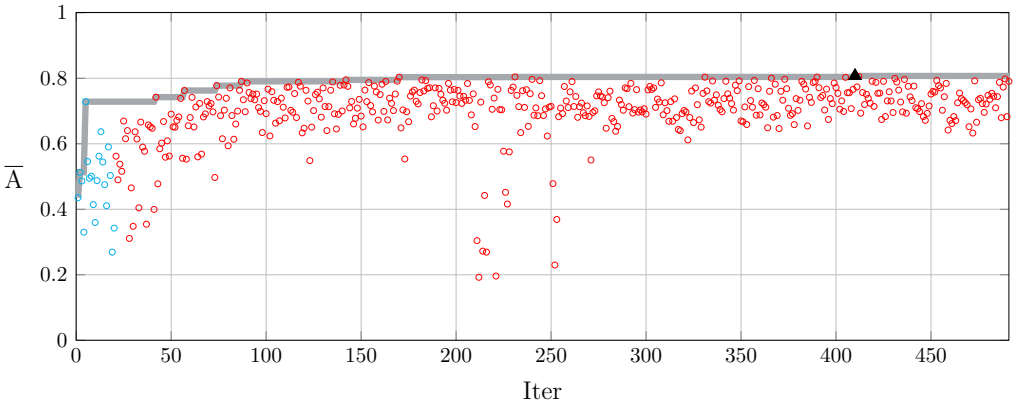


Figure III.24: Objective function according to the iterations, for the Bayesian optimization of the **setup 2** (defined in section III.4.3), on the parameters space $D_{2,sub\,2}$ (defined in Eq. III.68). DoE (resp. EGO) iterations are displayed in blue (resp. red). The maximum reached is 80% and is marked as a black triangle, with parameters given in Eq. III.71. The gray line is the maximum reached during the optimization.

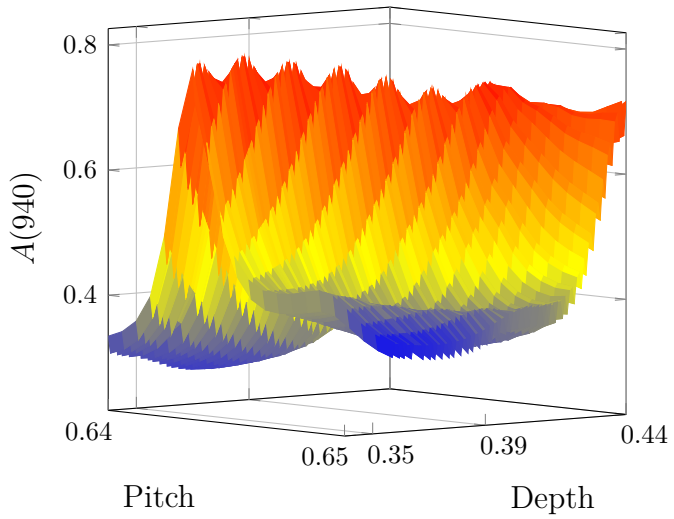


Figure III.25: Side view of the linear sweep responses on $D_{1,sub}$, defined in section III.4.5.2 and also shown in Fig. III.27 and Fig. III.26. This surface is generated with 21 linearly spaced dots on the Pitch axis and with 101 linearly spaced dots on the Depth axis.

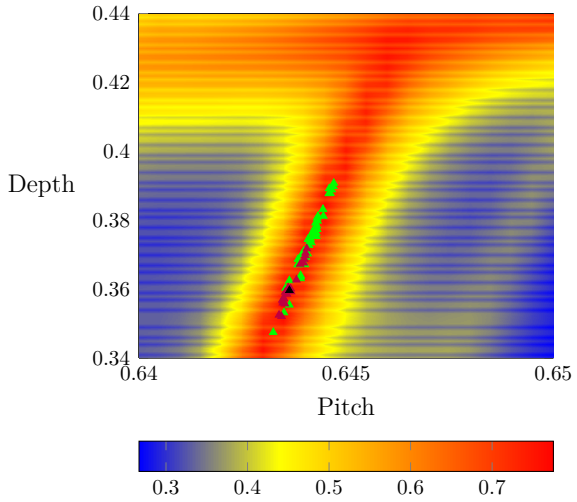


Figure III.26: Linear sweep response on $D_{1,sub}$ (defined in section III.4.5.2 and also shown in Fig. III.25 and III.27) and best performing parameters of the optimization of **setup 1** on $D_{1,sub}$. The black triangle marks the optimum reached, purple triangles mark the 20 first best performing parameters. Green triangles mark the 60 best performing parameters (except the first 20, already marked in purple or black). For the heat map, 21 dots are linearly spaced on the x-axis and 101 dots are linearly spaced on the y-axis.

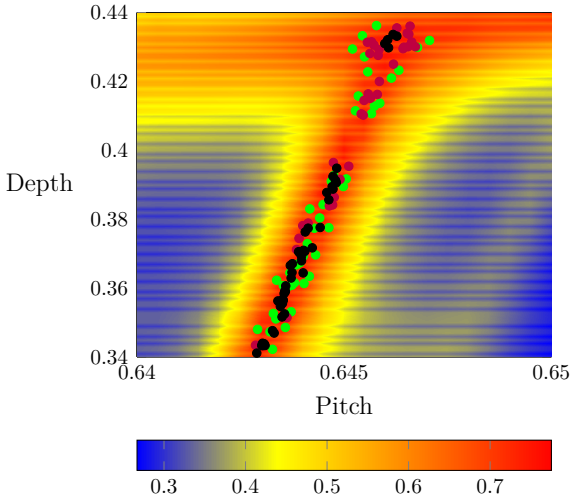


Figure III.27: Linear sweep response on $D_{1,sub}$ (defined in section III.4.5.2 and also shown in Fig. III.25 and III.26) and best performing parameters of the optimization of **setup 1** on D_1. Black circle marks the element of the class C_1^1, purple circles mark the element of C_2^1 and the green circles mark the element of C_3^1 (see Eq. III.64 for C_i^1 definitions). For the heat map, 21 dots are linearly spaced on the x-axis and 101 dots are linearly spaced on the y-axis.

III.5 Grating optimization in 3D

In this section, the investigation on the best performing 3D grating shape for a simple structured silicon slab is performed. Taking [88] as a guideline, the importance of breaking the symmetry in the grating geometry or equivalently the importance of using a superlattice, is also studied. So the goal of this section is to determine which shape is improving the best the volumic absorption, for both symmetric and unsymmetrical gratings, in a simple silicon slab.

For this study, we choose to investigate rectangular, cone, ellipsoid, cylinder, pyramidal and pyramidal with fixed angle gratings (more details are available on section III.5.1.). However, since one does not simply parameterize continuously 3D shapes from a cuboid to a cone, we consider various shapes and the optimizations are done independently on each shape. The numerical optical solver used is the DGTD fullvawe solver from [121] in order to, firstly, take advantage of the geometry versatility of this solver, and secondly, optimize on a wavelength range, as described in Eq. III.6.

Firstly, the simple silicon slab structure is described in section III.5.1. Secondly, the six optimization setups, as well as the geometrical parameters to optimize are clarified in section III.5.2. Finally the optimization results are presented in III.5.3, followed by the validation of the best design in III.5.4.

III.5.1 Structure definition

In this section the structure, and all the grating shapes studied, are described. Taking [88] as a guideline, we choose to optimize a more simplified structure than the 2D realistic SPAD of section III.4: an infinite x-y structured slab of Silicon surrounded by air and Cu in the z dimension, with absorbing condition in both z_{max} and z_{min}. The different layers and their thicknesses are, from top to bottom:

- A 200 nm PML air layer;
- A 200 nm SF air layer;
- A 200 nm TF air layer;
- A 400 nm structured Si layer, whose structuration are filled with air;
- A 200 nm TF Cu layer;
- A 200 nm SF Cu layer;
- A 200 nm PML Cu layer;

where SF and TF denote the enclosing TF/SF volume (the TF/SF decomposition is a tool to inject the source plane wave. Instead of being imposed on the boundary of the computational domain, the fields are imposed on an enclosing volume of the computational domain. See [58] for details). The source is injected on top to the Si layer at a distance of 200 nm. The superlattice is chosen of size $P = 600$ nm in both x and y dimensions, and the periodic boundary conditions are applied in x and y. In Fig. III.29, an example structure is shown, with the dimension provided above, where the structuration pattern is visible in the Si green layer.

The invariant geometrical parameters of the silicon slab being clarified, we now focus on the parametrization of the grating parameters, which constitute the variables to optimize. As previously mentioned, we aim to study the influence of different grating 3D shapes, as well as the symmetric hypothesis, on the absorption inside the Si layer. Six shapes are studied in this optimization: cone, ellipsoid, cylinder, pyramidal, pyramidal with fixed angle and rectangular gratings.

All these six grating shapes share a superlattice basis definition: on the x-y basis of dimension $P \times P$ nm^2, four smaller grating basis are defined, of size $a_1 \times a_1$, $a_2 \times a_2$, $a_2 \times a_1$ and $a_1 \times a_2$, where a_1, a_2 and P are linked by the following equation:

$$P = a_1 + a_2 + 2g_{be}, \tag{III.72}$$

and where g_{be} is the distance between each smaller basis. In Fig. III.29, an example of a grating basis is provided with views of the resulting mesh, illustrating the basis positions as well as the definition of g_{be}. Since a_1 and a_2 are linked by Eq. III.72, only one parameter, a_1, is a free variable. We choose for the following the range $[50, 250]$ nm for a_1. For five grating shapes, *i.e.* cone, ellipsoid, cylinder, pyramidal and rectangular, four additional parameters are necessary to set the height of each of the four structures. More precisely, we have the following parameters:

- h_1, the height of the grating structure with the $a_1 \times a_1$ basis,
- h_2, the height of the grating structure with the $a_2 \times a_1$ basis,
- h_3, the height of the grating structure with the $a_1 \times a_2$ basis,
- h_4, the height of the grating structure with the $a_2 \times a_2$ basis.

Since the Si layer is 400 nm thick, we choose for the parameters h_i the range $[50, 300]$ nm. For the pyramidal with fixed angle, the height is set by the following equation,

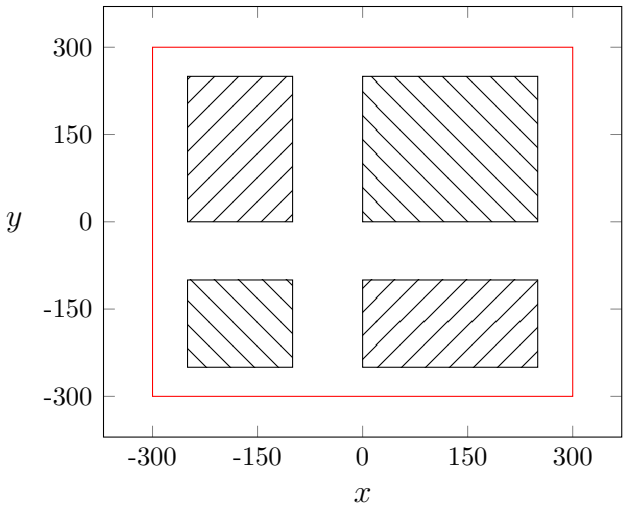
$$h = \frac{a_{small}}{2} \frac{\sin(54.7)}{\cos(54.7)}, \tag{III.73}$$

where a_{small} is the minimum side of the grating basis: *i.e.* a_{small} is equal to $\min(a_1, a_2)$ for the $a_1 \times a_2$ and the $a_2 \times a_1$ basis, a_1 for the $a_1 \times a_1$ basis and a_2 for the $a_2 \times a_2$ basis. The angle of 54.7 is chosen in agreement with the fabrication constraint on the Si pyramid [122]. For structuration presenting a peak, formerly the pyramidal, pyramidal with fixed angle and the cone, a truncation apex is applied for controlling the influence of this peak on the time step of the DGTD solver. The importance of introducing truncating structures is explained and described in section II.6.1.

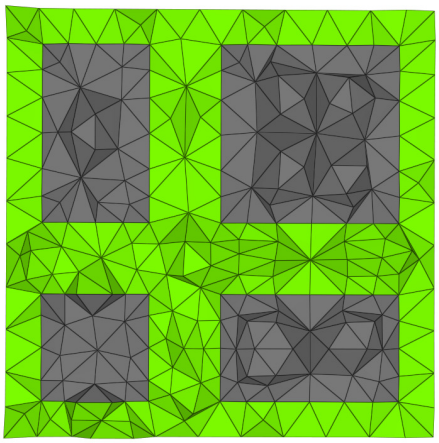
Palik references were used for the permittivity of both Si and Cu material [114].

In Fig. III.30, all the six grating shapes, for a fixed superlattice (with $a_1 = 150$ and $g_{be} = 100$) and, when relevant, fixed heights (with $h_1 = 100$, $h_2 = 150$, $h_3 = 200$ and $h_4 = 250$), are shown. The z-axis is inverted for visual clarity: the gratings are pointing upward.

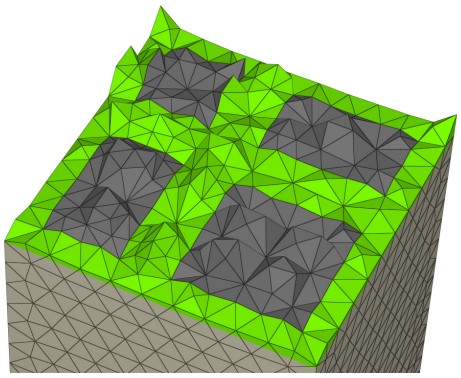
The structure is described, as well as the grating parameters and shapes. In the next section, we focus on the definition of the optimization setup.



(a) Schematic definition of the grating basis.



(b) Top view.



(c) Perspective view.

Figure III.28: Example of the basis grating definition. (a) The simulation domain in the x-y axis, of period P, is marked as a red square. The two squares are respectively of size $a_1 \times a_1$ (bottom left) and $a_2 \times a_2$ (top right). The rectangulars are of size $a_1 \times a_2$ (top left) and $a_2 \times a_1$ (bottom right). g_{be} is the distance between each basis, and $g_{be}/2$ is the distance of each basis to the simulation domain boundary (in red). For this example, $P = 600$ nm, $g_{be} = 100$ nm, $a_1 = 150$ nm and $a_2 = 250$ nm (given by Eq. III.72). (b-c) Corresponding mesh visualized with Vizir [123].

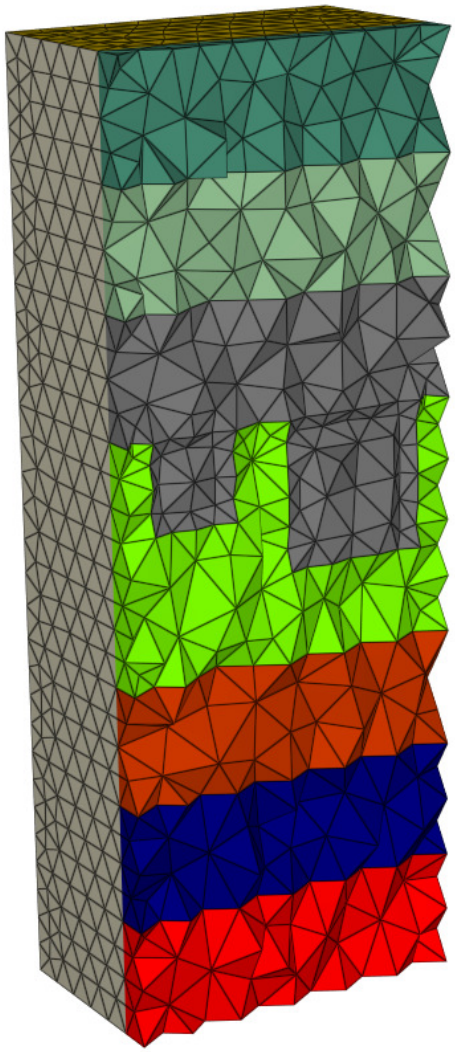


Figure III.29: Y-Z cut of the simple Si slab structure, with rectangular grating. From top to bottom, PML air is in dark green, SF air in pale green, air in gray, including the grating, Si in green, TF Cu in orange, SF Cu in blue and PML Cu in red. Visualization done with Vizir [123].

III.5.2 Optimization setup

In this section, the optimization setup for the 3D grating optimization is described, including the EGO parameters.

As mentioned in section III.3.4.1, an EGO required five initial parameters: the parameter space, noted D, the size of the DoE, a covariance kernel K, a merit function, and a stopping criterion. For the optimization of the inner volumic absorption on the simplified 3D Silicon slab described above (see section III.5.1), we choose the following input parameters:

- K is a square exponential kernel (see Tab. III.1);
- The DoE contains 20 iterations;
- The merit function is the expected improvement (see Eq. III.58);
- The stopping criterion is the number of EGO iterations, set to 20.

Since DGTD is a time domain solver, the objective function can be defined on a wavelength range rather than a single wavelength as done in section III.4. We choose to maximize the mean absorption on the wavelength range $[920, 960]$ nm, and denote this objective function as $\bar{\text{A}}$ in the following. We first define six optimizations setups, varying on the grating shape considered. The parameters range selected correspond to the range of interest defined in section III.5.1. For the pyramidal with fixed angle grating, we optimize on a_1, g_{be} and the truncation apex, while the other shapes are optimized on a_1, h_1, h_2, h_3, h_4 and $g_{be} = 40$ nm. These six first setup aims to investigate which shape performs better.

In order to test the symmetrical hypothesis, we also perform six optimizations with symmetric grating ($a_1 = a_2$, and $h_1 = h_2 = h_3 = h_4$), on the following parameters: only g_{be} for the pyramidal with 54.7 angle, g_{be} and h_1 on the five other grating shapes.

To perform these optimizations, we use the R DiceOptim package [107], coupled with the DGTD solver from the DIOGENeS software suite. The convergence study for the DGTD simulations of the optimized design is shown below in section III.5.4. The simulation time for one design is approximately 1 hour on 72 hyper threaded dual-Xeon Cascade Lake SP Gold 6240 @ 2.60GHz cores (144 threads).

Finally, the input parameters for the Bayesian optimization are defined, the corresponding results of the twelve optimization setups are presented in the next section.

III.5.3 Optimization results

In this section, the Bayesian optimization of the 3D grating parameters, on the twelve setup described in the previous section, are shown.

In Fig III.31, the mean absorption on the wavelength range $[920, 980]$ nm, *i.e.* the objective function to maximize, is shown for each DoE and EGO iterations, for all the six non-symmetric setup. The maximum reached for each shape, as well as the corresponding parameters, are available in Table III.3. The ellipsoid shape performs the best with a mean absorption of 12.5%. The cylinder (resp. rectangular) shape is the second (resp. third) best with an absorption of 11% (resp. 10%),

One must remark that, excepted for the pyramidal shapes, the optimization does not clearly reach a maximum, the dispersion of the objective function remaining high. For the 2D grating optimization, for instance in Fig. III.24, the optimizer was always reaching a maximum, lowering the dispersion of the objective function as the number of iterations increases. For all results of the 3D grating optimization, this is not always the case. It could be due to the number of parameters being high compared to the number of evaluations. Thus, hoping to clearly reach a maximum, we try to optimize the better performing shape, the ellipsoid grating, on the same parameters range, but with 30 DoE and 70 EGO iterations.

In Fig. III.32, the results of this optimization are shown. Even if the maximum reached is still 12.5%, the dispersion of the objective function remains high and 100 total iterations is not enough. We face here the main limitation of the 3D grating optimization: even on a simplified slab structure, the whole simulation time is the current limitation. In our case, the simulation time being one hour approximately for each iteration [4], the total time for such optimization is more than four days.

Even if we cannot claim to have reached the optimum, we present a clear trend: the ellipsoid grating is performing better than all other shapes considered. In Table III.4, the improvement of the best ellipsoid design on all other best performing designs are shown, exhibiting an improvement of minimum 12% of the best cylinder design, and an improvement of maximum 160% on the best cone design.

As mentioned in section III.5.2, we also perform the optimization on the symmetrical structure in order to test the unsymmetric hypothesis on the six considered shapes. The objective function, for each iteration, of the six corresponding optimizations, are shown in Fig. III.36. Similarly to the asymmetrical grating optimizations, we cannot clearly claim to have reached the optimum. However for all grating shapes, except the conical one, the best asymmetric grating is performing better than the best symmetric one. On Fig III.33, the best design mean absorption of all twelve optimizations are shown, clearly illustrating the drastic impact of the asymmetrical pattern to achieve efficient light trapping.

Following [88], we can explain this higher absorption because the asymmetrical grating pattern excites more guided resonances in the Si layer, allowing a better absorption. Indeed, in view of the light trapping theory, the enhancement of light absorption depends on the number of optical resonances supported by the structure that could be excited.

To complete these results, the absorption spectrum of the maximum reached mean absorption for ellipsoid grating, and the corresponding maximum reached for the symmetrical grating, are shown on Fig. III.34. On the one hand, one clearly sees the multiple resonances enhanced by the asymmetrical grating that allow a higher mean absorption. On the other hand, the choice of $\bar{A}$ as the objective function implies that the resonances

[4] On 72 hyperthreaded Xeon 3.2 GHz cores.

are not positioned exactly at 940 nm, unlike the 2D optimization, as visible in Fig. III.18 and III.21, where A_{940} was optimized.

The optimization results of the twelve setup being presented, we focus in the next section on the validation of the best design in order to prove the convergence of all iterations.

Structure type	a_1	h_1	h_2	h_3	h_4	g_{be}	$\bar{A}$
Pyr	148.5	253.7	143.1	163.0	300.0	40.0	9.8%
Rect	102.3	266.3	180.1	159.3	105.7	49.1	10.5%
Ellipse	199.4	134.9	226.5	92.2	151.7	69.0	12.5%
Cone	111.0	223.8	88.7	183.4	289.7	40.0	4.8%
Cylinder	163.5	103.4	176.5	170.6	171.1	51.5	11.1%
Structure type	a_1	Apex	-	-	-	g_{be}	$\bar{A}$
Pyr 54.7	88.3	0.27	-	-	-	30.0	9.3%

Table III.3: Optimal parameters for the six optimization of the six corresponding unsymmetrical gratings.

III.5.4 Validation of the best design

In this section, the convergence study of the DGTD solver is shown. In Fig III.35a, the best asymmetrical cylindric design absorption profile is shown for increasing interpolation degree. The energy threshold is set to 0.01%. The mesh, kept constant, contains 20174 cells, with minimum cell size of 11.7 nm and maximum cell size 127 nm. The absorption profile is already well captured by the P_2 curve. A similar study for the flat model (without structuration) is performed on Fig. III.35b. Similarly to the cylindrical design, P_2 interpolation is enough.

III.5.5 Conclusion on 3D grating optimization

This study aimed to find the best performing 3D grating shape for a simple structured silicon slab is performed. Taking [88] as a guideline, where only 2D optimization are performed, we extended this research to unsymetric grating pattern of various shapes. From considering 3D structures, specific challenges arise.

Firstly the computation time for one iteration drastically increases, requiring the use of a global optimizer which has proven in literature its efficiency in reducing the total number

Structure type	max $\bar{A}$	Ellipsis Improvement
Ellipsis	12.5%	*
Cone	4.8%	160%
Cylinder	11.1%	12%
Pyramidal 54.7	9.3%	34%
Rectangular	10.5%	19%
Pyramidal	9.8%	27%

Table III.4: Maximum mean absorption reached on the six non symmetrical setup, and the corresponding improvement of the ellipsoid best design. The "Ellipsis improvement" is the percentage of improvement of the best ellipsoid design on the maximum reached design for the shape considered. For instance, for the cone grating, we have $\frac{12.5\times100}{4.8} - 100 = 160\%$.

of iterations needed. Bayesian optimization was chosen for this property in particular. The computational cost is currently the main limitation for 3D optimization, justifying the remarquable lack of such results in the CMOS imager scientific community.

Secondly, adding an extra dimension increases the number of shapes that can be considered. Usually, square or pyramidal gratings are taken into account for 2D gratings. But a 2D square can be a slice of either a cylinder, an ellipsoid or a cuboid. Similarly a 2D pyramid, namely a triangle, can be a slice of a pyramid or a cone. Extrapolating 2D optimization results is thus not possible without first answering the question of the structural extension. Such choice is certainly impactful on the resulting light absorption, and must be taken into account. Our study, by optimization each shapes independently, is seeking to answer explicitly this question.

From a physical standpoint, unsymetric gratings enhances more resonances than symetric gratings, allowing a better averaged absorption. In Fig III.34, the symetric optimal design has higher absorption peak but a lower averaged absorption on the wavelengths range considered, since the unsymetric optimal design has three absorption peaks. The balance between a single but higher, or multiple but lower absorption peaks, is clarified in our comparison between symetric and unsymetric grating pattern.

The main limitation of our study relies in the fact that we cannot claim to have reached the optimum. This is due to the specificity of the 3D geometrical optimization, namely the computational cost. Even if a simple Silicon slab of 400 nm thickness, the simulation time for one iterations is one hour on 72 hyperthreaded Xeon 3.2 Gz cores.

This challenge justifies the choice of a fine Silicon slab compared to the ~5 μm of the 2D grating optimization of section III.4. Since the Silicon thickness is lower, the absorption enhanced is lower. Indeed in the absence of grating, a Silicon slab is a Fraby-Perrot cavity

whose resonnances are determined only by the silicon thickness. The 12.5% obtained by the best design of the best performing shape, the ellipsoid, cannot be fairly compared to the ~80% obtained in section III.4.

However the drastic impact of the grating is shown by the comparison to the flat case whose absorption is 1%, illustrating the importance of the nanostructuration.

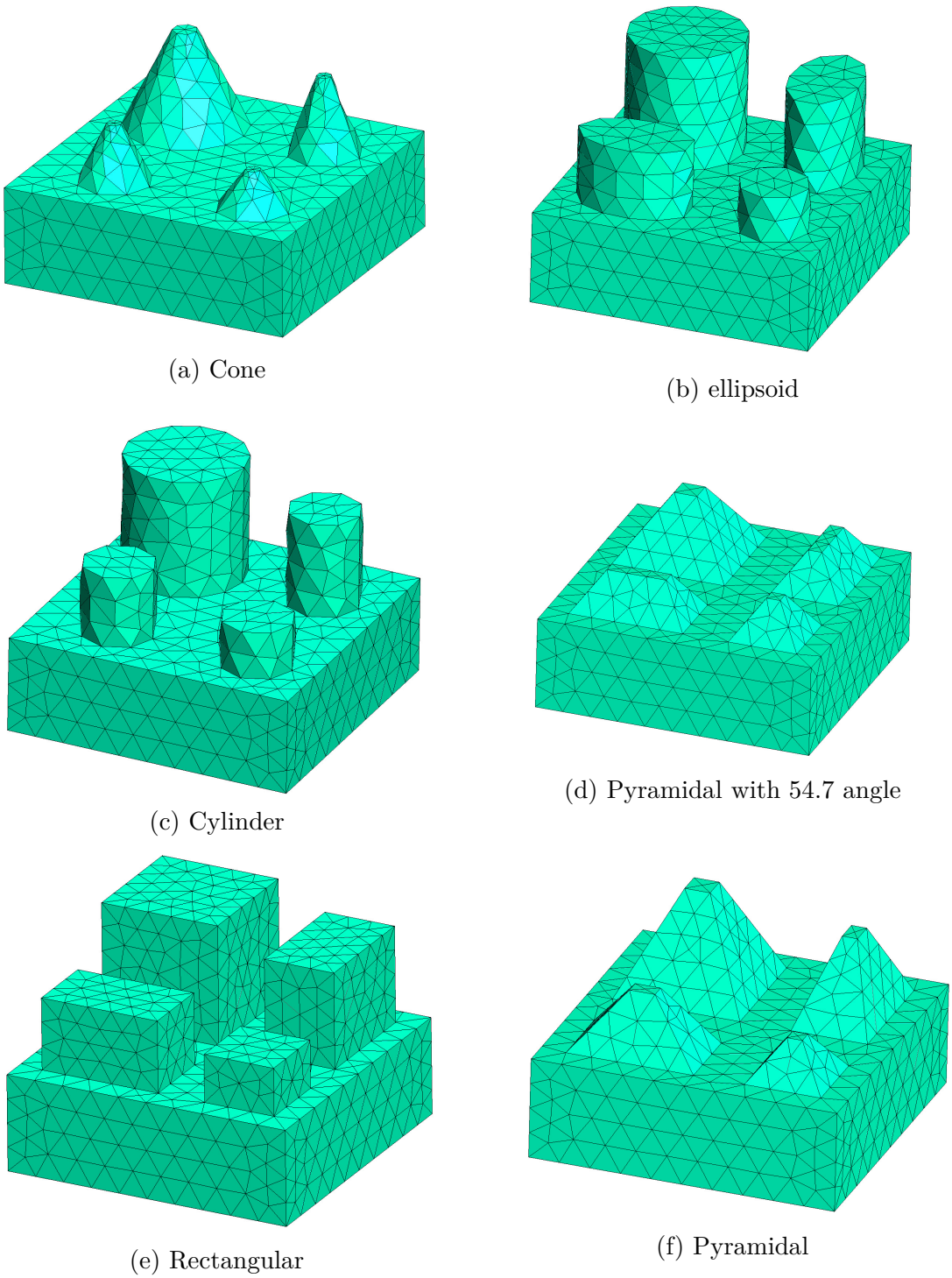
Finally, this work shows that ellipsoid grating is a good candidate to fabricate high performing 3D grating, compared to the cylindrical, rectangular, conical or usual pyramidal grating. In particular, the use of a superlattice has been demonstrated to enhance higher absorption for all shapes considered, excepted the conical grating.

III.6 Conclusion

The optimizations on 2D and 3D gratings performed in this chapter demonstrated that the EGO is a great candidate for investigating the dependence of optical resonances in CMOS imagers on the geometrical parameters. In particular the 2D grating optimization, exhibiting an absorption of 83% for SiO2 DTI, and 88% for tungsten DTIs (see appendix A), found an absorption at 940 nm higher that everything found in literature. Furthermore, the race to the maximum absorption reached a point of no return, where the absorption cannot be improved without geometrical parameter variations lower than the nanometer. Further studies on resonances must focus on both sensitivity analysis or classification in order to find geometrical parameters that provide high absorption in the range of feasible lithography precision.

To perform such optimization in 3D, the numerical cost is still the main bottleneck. In our work, we considered a simple silicon slab, far from a realistic SPADs, but still the convergence to the optimum remains unclear, even after 100 solver evaluations.

The cost function of the two optimizations are different: in 2D, the absorption at 940 nm was optimized and successfully positioned a resonance at 940 nm. While, in 3D, the mean absorption on the range $[920, 980]$ nm was optimized. The choice of this cost function *in fine* rely on the exact signal the SPADs received: optimizing the absorption at 940 nm, or a mean interval around this wavelength, must be selected exactly on the performance desired by industrial. Such a choice cannot be communicated in this work. However, from our two optimization results, the influence of this cost function is clear: optimizing exactly at 940 nm successfully positions a resonance at 940 nm, while optimizing on an interval only positions several resonances within this interval.



(a) Cone

(b) ellipsoid

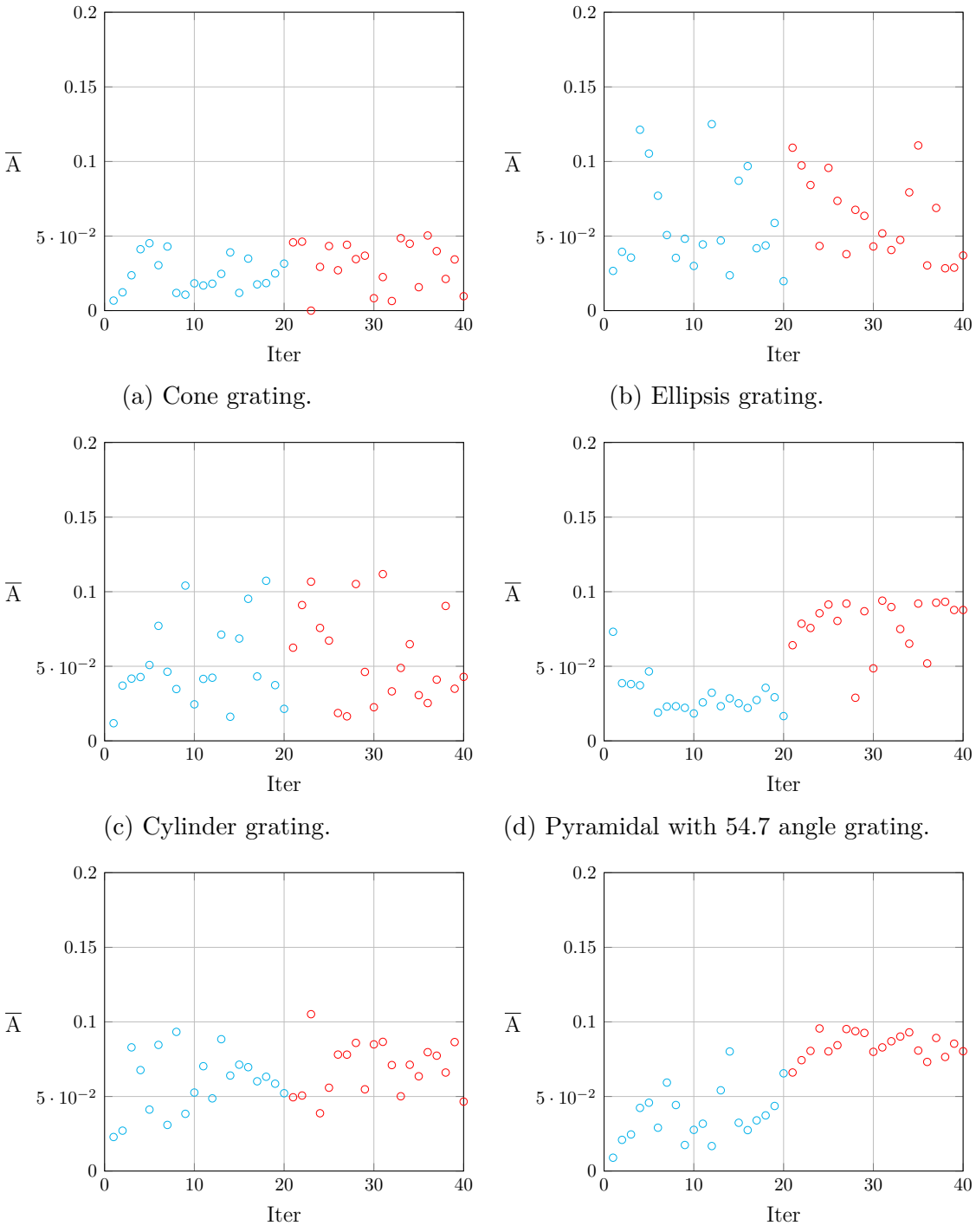
(c) Cylinder

(d) Pyramidal with 54.7 angle

(e) Rectangular

(f) Pyramidal

Figure III.30: Six grating shapes, with constant superlattice parameters ($a_1 = 150$ and $g_{be} = 100$) and, when relevant heights ($h_1 = 100$, $h_2 = 150$, $h_3 = 200$ and $h_4 = 250$). The z-axis is inverted to ease visualization. Meshes pictures are done with GMSH [124].



(a) Cone grating.

(b) Ellipsis grating.

(c) Cylinder grating.

(d) Pyramidal with 54.7 angle grating.

(e) Rectangular grating.

(f) Pyramidal grating.

Figure III.31: Objective function for all iterations of the six **unsymmetric** optimization setup described in III.5.2. Blue dots (resp. red) are DoE (resp. EGO) iterations. The best performing design is the ellipsoid one, reaching a maximum of 12.5% mean absorption, whose corresponding parameters are available in table III.3.

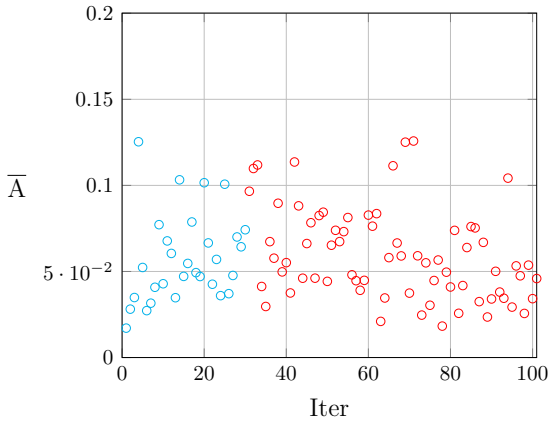


Figure III.32: Objective function according to iterations, for the ellipsoid grating, on 30 DoE and 70 EGO iterations, on the optimization setup described on section III.5.2 and III.5.3. The maximum reached is 12.5% and is equivalent to the previous maximum reached of Fig. III.31b.

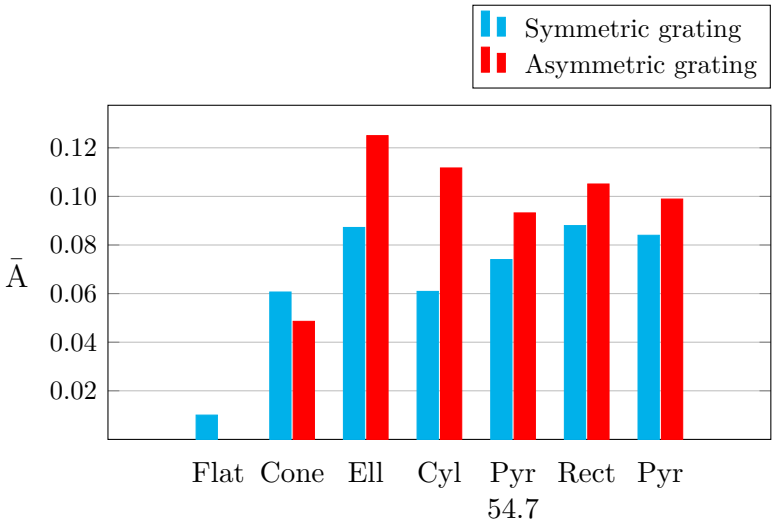


Figure III.33: Best performing designs of all twelve optimization setups defined in section III.5.3. Red (resp. blue) represents the symmetrical (resp. unsymmetrical) gratings. Each best performing design is the corresponding maximum reached within Fig. III.31 and III.36.

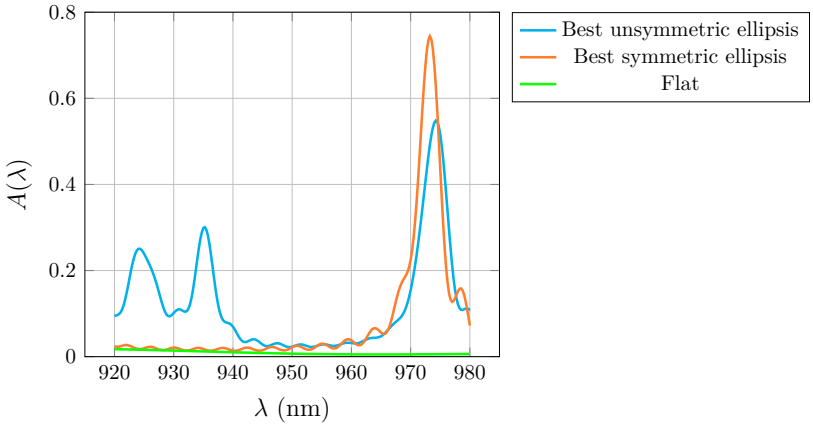
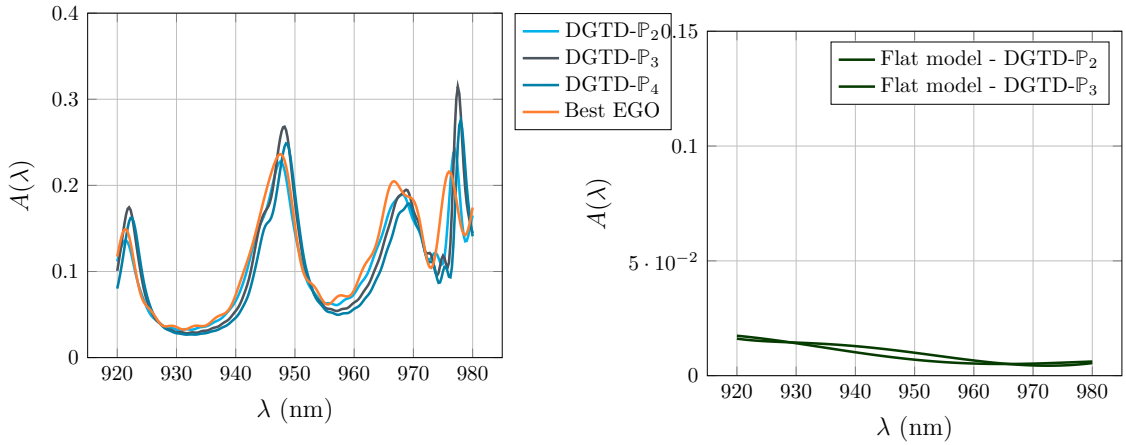


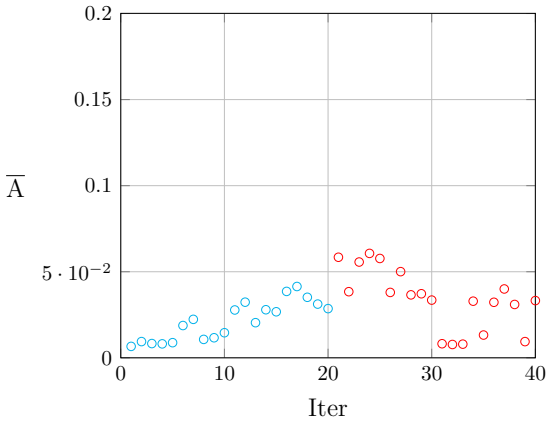
Figure III.34: Absorption spectrum of the best performing ellipsoid design, for the symmetric and unsymmetric optimization setup, compared to the flat design (without gratings).



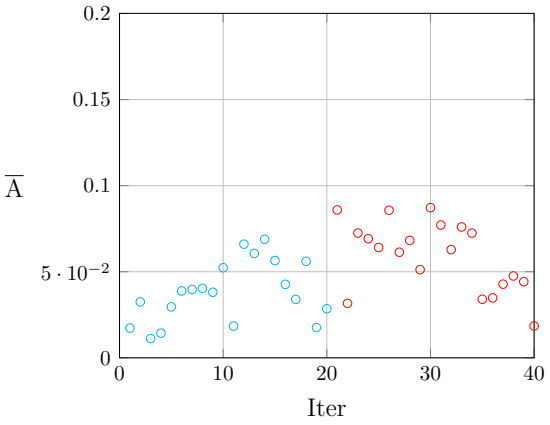
(a) Best performing cylinder superlattice.

(b) Flat model.

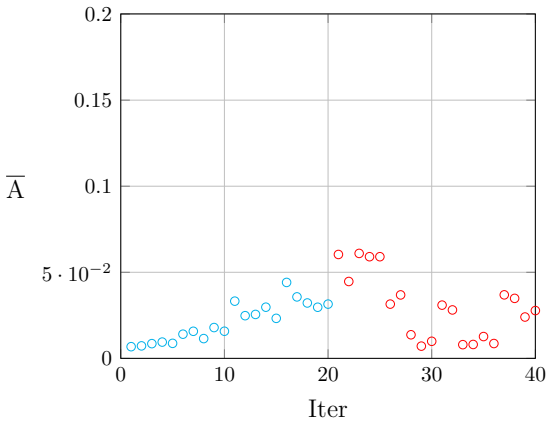
Figure III.35: On the left, validation of one best performing design for the DGTD solver. The degree of interpolation is increased, while the mesh is unchanged, with characteristics given in section **??**. On the right, convergence study of the flat model, without structuration.



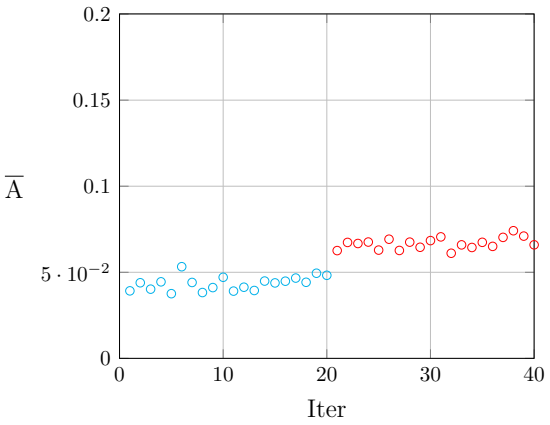
(a) Cone grating.



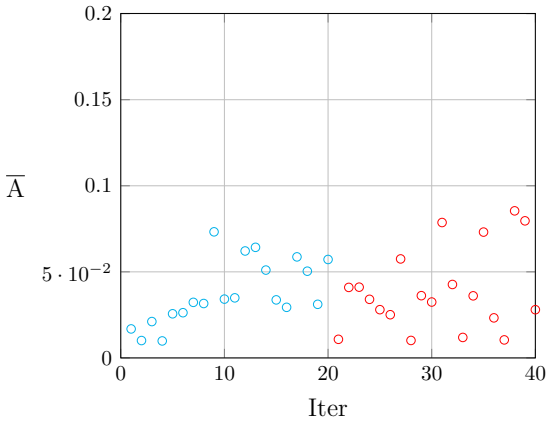
(b) Ellipsis grating.



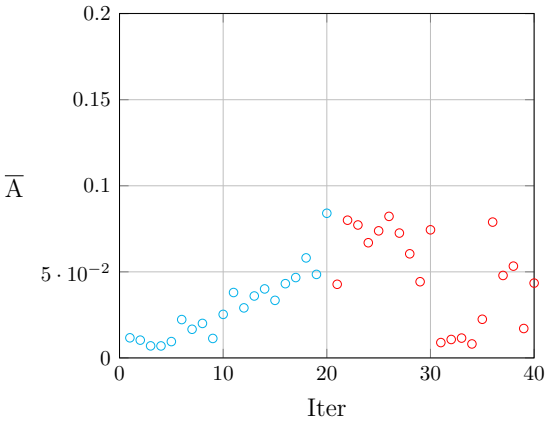
(c) Cylinder grating.



(d) Pyramidal with 54.7 angle grating.



(e) Rectangular grating.



(f) Pyramidal grating.

Figure III.36: Objective function for all iterations of the six **symmetric** optimization setup described in III.5.2. Blue dots (resp. red) are DoE (resp. EGO) iterations. The best performing design is the ellipsoid one, reaching a maximum of 8.7% mean absorption.

Conclusion

The purpose of this last chapter is to summarize the content of this manuscript and to identify possible extensions.

In the first chapter, a semi-empirical model for the optical properties of SiGe alloys is proposed, based on physical considerations and the summation of Tauc-Lorentz and parametric oscillators. The key parameters of the oscillators are intuited from the optical transition and the symmetry points in the band structure. The model is fitted on extensive experimental data, for different temperatures and germanium concentrations. This model can be used in optical simulation tools, in order to help the design of optoelectronic devices based on SiGe materials.

In the second chapter, a benchmark of the reference numerical method, the FDTD method, and two alternatives, the RCWA and the DGTD methods, has been performed on structures of increasing complexity, culminating with a nanostructured SPAD device. To this purpose, 2D and 3D RCWA solvers have been implemented in Matlab. The reference FDTD method remains faster than the other methods, even on the most complicated structure considered. Detailled explanations, available in section II.7, limit the impact of such empirical benchmark.

In the third chapter, we introduced an inverse design approach that combines optical solvers for the numerical characterization of light absorption in a nanostructured CMOS image sensor, with a statistical learning-based global optimization method for goal-oriented discovery of the optimal patterning parameters for maximising light absorption. Firstly, we optimize the grating parameters of a realistic 2D SPAD with DTIs, computing the light absorption with our 2D RCWA in-house solver, reaching an absorption of 83% at 940 nm. Secondly, various 3D pattern shapes (ellipsoid, cylinder, rectangle, pyramidal etc.), for symmetric and unsymmetric gratings on a simple silicon slab, are optimized in order to determine the shapes that maximize light absorption in the 920-980 nm range. This second optimization takes advantage of the geometry versatility of the DGTD method. Both optimizations are performed with the Efficient Global Optimization (EGO) method, achieving a convergence to the optimum within a reasonable number of solver evaluations.

Some perspectives of this work are the following:

- The model of the first chapter is said to be accounting for strain or unstrained SiGe. The strain is taken into account only in the band structure computation of strained/unstrained SiGe alloy. However, no comparison have been made between the prediction of our model on strain SiGe permittivity data. Right now, this is

only supposed to match. A future work would ensure that this model is correct for strained SiGe alloys.

- Benchmarks are always selective. In our work, we selected only the DGTD implementation from the DIOGENeS software suite, while a DGTD solver is actually available from Ansys. An identical remark can be said about the Reticolo RCWA solver: we tested only our in-house RCWA solver and not the Reticolo one. Benchmarks across more solvers could be done.

- The grating optimization performed in the chapter 3 must include, in future works, a sensitivity analysis. Indeed, the optimum is found at a precision with no physical meaning. This would allow to optimize within the range of what is feasible with the current fabrication process.

Epilogue

Un syllogisme est défini par Aristote comme une succession de proposition s'enchaînant logiquement. La forme du syllogisme, de l'implication ou de la contraposée par exemple, assure sa validité, et ainsi la vérité de la conclusion suivant les prémisses. Tout cela est bien connu des scientifiques, ils l'apprennent souvent tôt, nomment rigueur cette habilité à expliciter toutes les étapes d'un raisonnement, et s'efforcent sans cesse de l'enseigner, par une répétition ardue, un effort constant et des corrections implacables mais justes.

Il y a cependant une dimension des syllogismes qui est rarement explicitée, voire comprise, dans l'enseignement scientifique, et chez les stagiaires, doctorants, ingénieurs de recherches ou chercheurs qui quotidiennement travaillent, réfléchissent et publient des articles. Cette dimension, inhérente à tout syllogisme, est présentée par Aristote dans le même livre les décrivant, juste quelques pages plus loin, il s'agit du problème du troisième homme.

Aristote reprend une critique de la théorie platonicienne des idées, critique que Platon lui-même avait formulée dans le Parménide. Entre deux termes, l'idée de l'homme et un homme existant, un troisième terme, un troisième homme, décrivant la relation entre ces deux termes, se glisse irrémédiablement. Un syllogisme peut ainsi constamment être coupé en deux, un A_{bis} s'imbrique entre les deux termes initiaux, A et B, afin d'expliciter l'implication $A \mapsto B$. Ce troisième terme appelle lui-même deux implications $A \mapsto A_{bis}$, $A_{bis} \mapsto B$, qui à leur tour nécessitent un intermédiaire supplémentaire, et ceci à l'infini. Cela est problématique, car tel la flèche de Zénon, jamais le raisonnement n'atteindra son terme. Il sera comme perdu dans les limbes.

Comment ce problème antique resurgit dans nos travaux ? Où donc un troisième terme eut été nécessaire ? A ces questions nous souhaitons, comme d'ordinaire en science, répondre par une nième image. Non un graphique mais une simple métaphore mathématique. Si cette thèse est un parcours présentant un départ et une arrivée, d'un 0 à un 1, l'ensemble de ses chapitres, sections, paragraphes et phrases peut être compris comme un ensemble rationnel fini, distribué dans l'intervalle $[0, 1]$, le complémentaire de cet intervalle est lui-même un intervalle, non connexe cette fois, partagé entre d'une part le contexte, les motivations et les enjeux, $]\infty, 0[$, et les perspectives, $]1, \infty]$. En principe, un ensemble fini de nombre rationnel ne déroge pas au problème du troisième homme: entre deux nombres rationnels, un troisième existe toujours entre les deux premiers ; $\mathbb{Q}$ est dense dans $\mathbb{R}$. Mais la distribution des nombres ne s'arrêtent évidemment pas aux simples nombres rationnels: $\mathbb{I}$ l'ensemble des nombres irrationnels est également dense dans $\mathbb{R}$, et symétriquement un nombre irrationnel existe nécessairement entre tous couples de rationnels donnés, peu importe leur promiscuité. Le réel, ou les nombres réels, est plus implacable encore. $\mathbb{I}$, dans $\mathbb{R}$ muni de la mesure de Lebesgue, est de mesure 1, quand $\mathbb{Q}$ est de mesure nulle. $\mathbb{Q}$ n'est que dénombrable et les deux ensembles infinis $\mathbb{I}$ et

$\mathbb{Q}$ sont de nature différents, le premier étant plus vaste que le second. $\mathbb{Q}$ n'est ainsi qu'un peigne de dirac dans un intervalle continu. On le voit, ce problème du troisième homme est virtuellement partout, et actuellement nulle part.

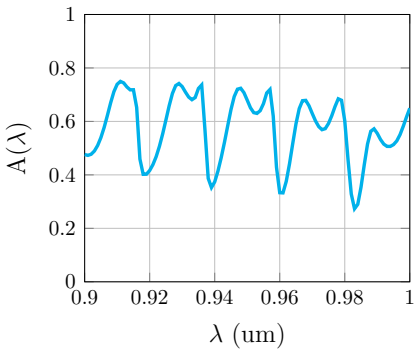
Une erreur qui revient souvent chez un doctorant est de dévaluer systématiquement son propre travail. L'apprenti chercheur apprend ainsi une règle implicite qui gouverne la pratique scientifique: rendre explicite ces termes intermédiaires est un travail réservé, non à l'auteur, mais à ses critiques. Ou suivant les mots de Wittgenstein, ce dont on ne peut parler, on doit le taire.

Appendix A

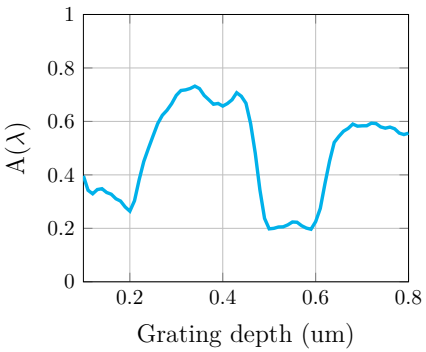
Grating optimization in 2D, with Tungsten DTI

The conclusion of the article that initially motivated our 2D grating parameters optimization [82], claims that using high absorbing materials in DTI enhance higher absorption. So this appendix is a replicate of the 2D grating parameters optimization, but with Tungsten, a high absorbing material, Deep Trench Isolation.

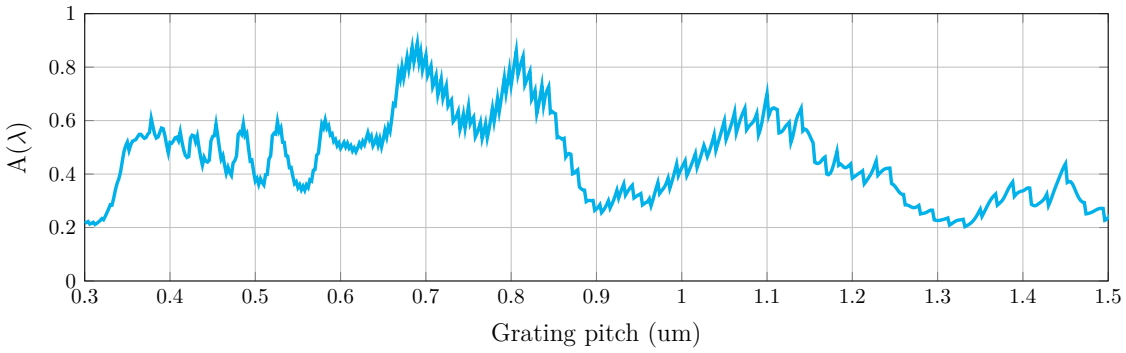
The grating parameters and the underlying structure are identical to the one described in section III.5.1. The parameters sensibility analysis, analogous to section III.4.2, is available in Fig. A.1. Two, respectively three, parameters optimization results (analogous to section III.5.2) are shown in Fig A.2, resp. A.3. The maximum reached in both optimizations is an absorption of 83%. The use of Tungsten did not increased the maximum reached. All conclusions provided in the chapter III apply to this study.



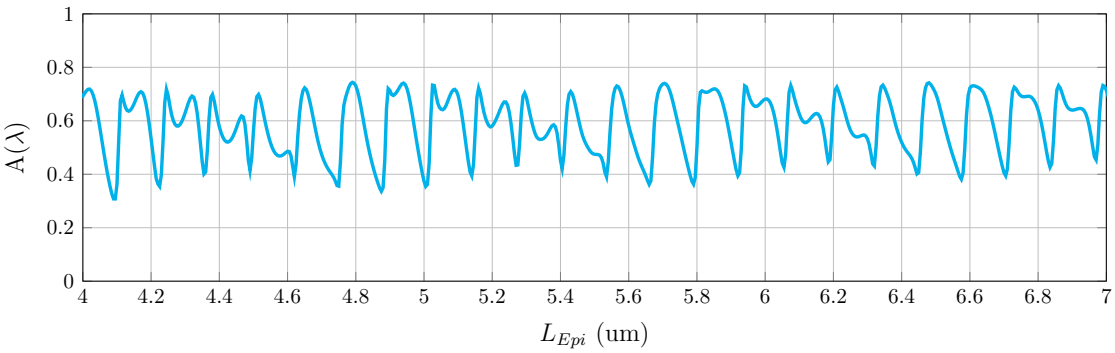
(a) Wavelength sweep, 101 dots.



(b) L_{depth} sweep, 71 dots.



(c) L_{pitch} sweep, 851 dots.



(d) L_{Epi} sweep, 1001 dots.

Figure A.1: RCWA 2D computation of the inner Si absorption, as a function of various parameters, of the structure described in section III.5.1, with Tungsten DTI. If not varying, we have $L_{epi} = 5$ μm, $L_{pitch} = 500$ nm, and $L_{depth} = 218$ nm. A_{940} denotes the inner absorption at 940 nm. In each subcation, "X dots" indicates the number of equally spaced x-axis dots that were simulated.

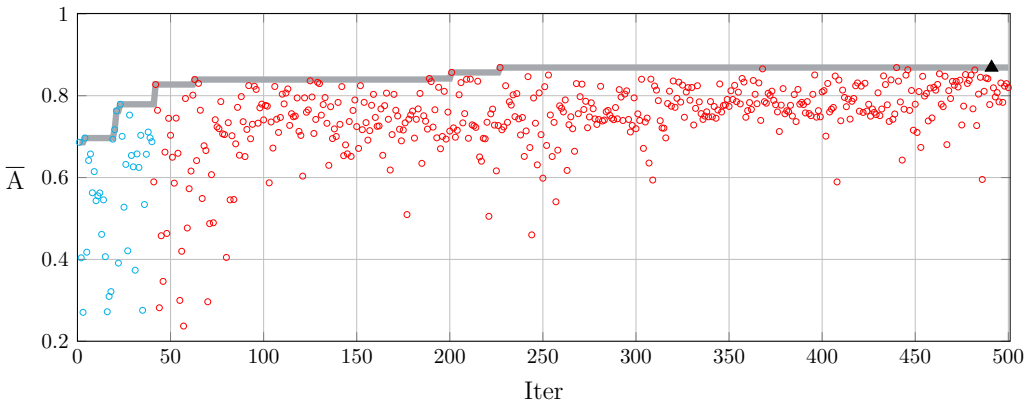


Figure A.2: Objective function according to the iterations, for the Bayesian optimization of the **setup 1** (defined in section III.4.3), on the parameters space $D_{1,sub}$ (defined in section III.4.4.1. DoE (resp. EGO) iterations are displayed in blue (resp. red). The maximum reached is 83% and is marked as a black triangle. The gray line is the maximum reached during the optimization.

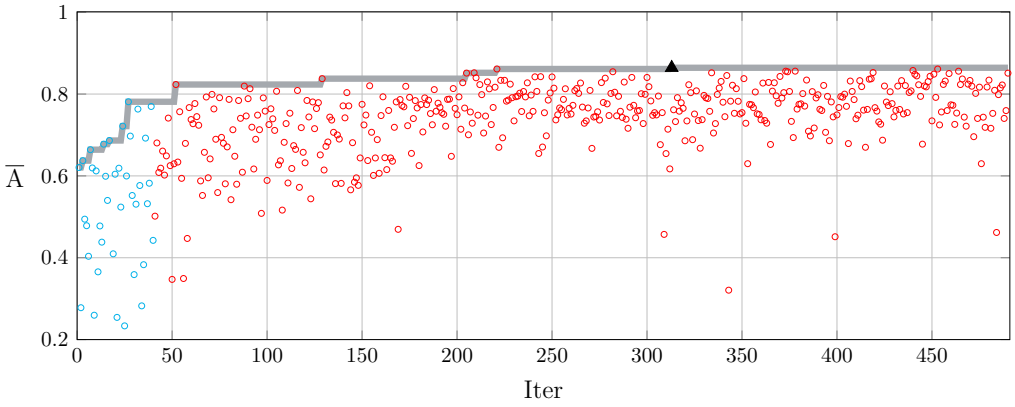
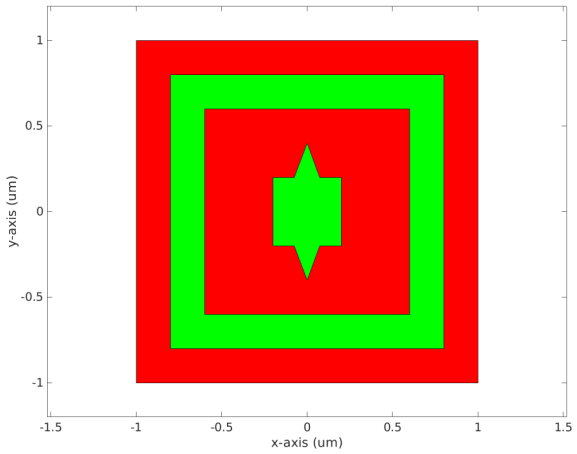


Figure A.3: Objective function according to the iterations, for the Bayesian optimization of the **setup 2** (defined in section III.4.3), on the parameters space $D_{2,sub\,2}$ (defined in Eq. III.68). DoE (resp. EGO) iterations are displayed in blue (resp. red). The maximum reached is 82% and is marked as a black triangle, with parameters given in Eq. III.71. The gray line is the maximum reached during the optimization.

Appendix B

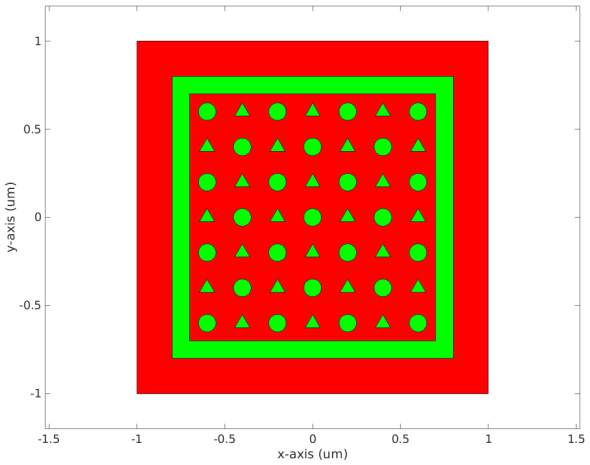
3D Structure generation for RCWA

In this appendix, several examples of 2D layers building, for 3D RCWA simulations, are shown. These examples illustrate the section II.5.2 where the geometry module for our RCWA implementation is described.



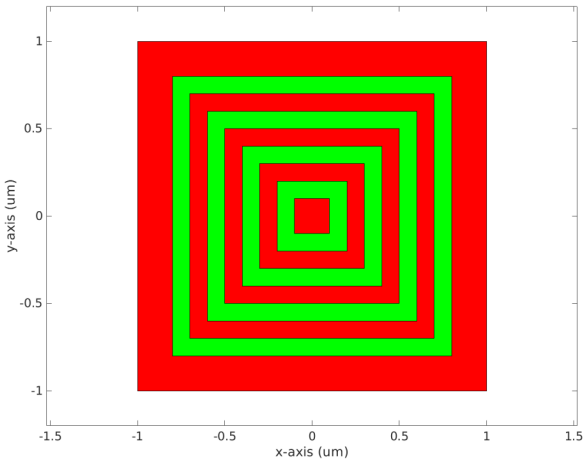

```
Array_of_circle = [];
Array_of_Polygon = [];

s = 0.8;
p1 = [  -s, -s;
         s, -s;
         s,  s;
        -s,  s];

s = 0.6;
p2 = [  -s, -s;
         s, -s;
         s,  s;
        -s,  s];

s = 0.2;
p3 = [  -s, -s;
         s, -s;
         s,  s;
        -s,  s];

p4 = [ 0.15, 0.0;
        0.0, -0.4;
      -0.15, 0.0;
        0.0,  0.4];

% Adding 4 polygones
Array_of_Polygon = [ Polygon2D(p1, epsSiO2),
                     Polygon2D(p2, epsSi),
                     Polygon2D(p3, epsSiO2),
                     Polygon2D(p4, epsSiO2)];

% Creating layer
Layers = [layer_2D(d,                          ...
                   eps_square,                 ...
                   Array_of_Polygon,           ...
                   Array_of_circle,            ...
                   current_z-d,                ...
                   current_z)];
current_z = current_z - d;
```

Figure B.1: Illustration of a RCWA layer with multiple including polygons.



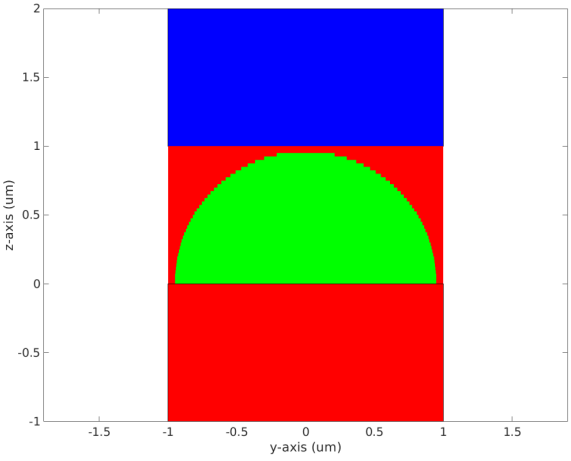

```
Array_of_circle = [];
Array_of_Polygon = [];

s = 0.8;
p1 = [  -s, -s;
         s, -s;
         s,  s;
        -s,  s];

s = 0.7;
p2 = [  -s, -s;
         s, -s;
         s,  s;
        -s,  s];

Array_of_Polygon = [ Polygon2D(p1, epsSiO2),
                     Polygon2D(p2, epsSi)];

grating_center = [-0.6;
                  -0.4;
                  -0.2;
                   0.0;
                   0.2;
                   0.4;
                   0.6];

s = 0.05; % grating size
counter = 1;
for x_c = grating_center
    for y_c = grating_center
        if mod(counter, 2) == 0
            pC = [ x_c - s*sqrt(3)/2,  y_c  - s/2 ;
                   x_c + s*sqrt(3)/2,  y_c - s/2  ;
                   x_c              ,  y_c + s    ];

            Array_of_Polygon = [  Array_of_Polygon,
                                  Polygon2D(pC,        ...
                                            epsSiO2)  ...
                                ];
        else
            Array_of_circle = [Array_of_circle,       ...
                               Disk_2D(s,             ...
                                       [x_c, y_c],    ...
                                       epsSiO2)       ...
                              ];

        end
        counter = counter + 1;
    end
end

% Creating layer
Layers = [layer_2D(d,                    ...
                   eps_square,           ...
                   Array_of_Polygon,     ...
                   Array_of_circle,      ...
                   current_z-d,          ...
                   current_z)];
current_z = current_z - d;
```

Figure B.2: Illustration of a RCWA layer using loops to build complex gratings.



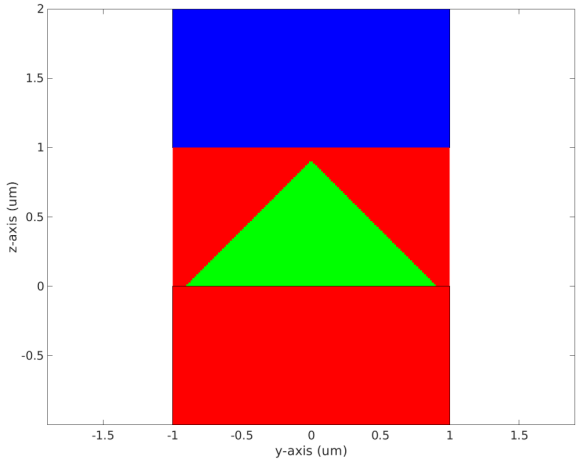

```
Array_of_circle = [];
Array_of_Polygon = [];

all_square_size = [0.2, 0.4, 0.6, 0.8];
for s = all_square_size
    p1 = [  -s, -s;
             s, -s;
             s,  s;
            -s,  s];

    p2 = [  -s + 0.1,  -s + 0.1;
             s - 0.1,  -s + 0.1;
             s - 0.1,   s - 0.1;
            -s + 0.1,   s - 0.1];

    Array_of_Polygon = [  Array_of_Polygon,
                          Polygon2D(p1, epsSiO2),
                          Polygon2D(p2, epsSi)];
end

% Creating layer
Layers = [layer_2D(d,                          ...
                   eps_square,                 ...
                   Array_of_Polygon,           ...
                   Array_of_circle,            ...
                   current_z-d,                ...
                   current_z)];
current_z = current_z - d;
```

Figure B.3: Illustration of a RCWA layer using multiple including square.




```
% Number of layer in z.
NL = 40;
air_guard_top = 0.1;
top_lense = current_z;
bot_lense = current_z - d;
radius_lense = top_lense - bot_lense - air_guard_top;
dz = (top_lense - bot_lense)/NL;

for iL = 1:NL
    % zmax, zmin and d of layer iL.
    zmax_layer = top_lense - dz *(iL-1);
    zmin_layer = top_lense - dz*iL;
    d_layer = dz;
    radius = sqrt(radius_lense -  (zmin_layer- bot_lense)^2 );

    % adding a circle only if radius is real
    if isreal(radius)
          Array_of_circle = [ Disk_2D(radius,
                                      [0.0, 0.0],
                                      epsSiO2) ...
                            ];
    else
          Array_of_circle = [ ];
    end

    Array_of_Polygon = [ ];

    % adding the layer
    Layers(end + 1) = [layer_2D(d_layer,
                                eps_square,
                                Array_of_Polygon,
                                Array_of_circle,
                                zmin_layer,
                                zmax_layer)];
    current_z = current_z - d_layer;
end
```

Figure B.4: Illustration of a 3D RCWA structure with a lens.




```
% Number of layer in z.
NL = 70;
air_guard_top = 0.1;
top_pyr = current_z;
bot_pyr = current_z - d;
heigh_pyr = top_pyr - bot_pyr - air_guard_top;
dz = (top_pyr- bot_pyr)/NL;

for iL = 1:NL
    % zmax, zmin and d of layer iL.
    zmax_layer = top_pyr - dz *(iL-1);
    zmin_layer = top_pyr - dz*iL;
    d_layer = dz;
    radius = heigh_pyr - (zmin_layer - bot_pyr);

    % adding a circle only if radius is real
    if radius > 0
          s = radius;
          p1 = [  -s, -s;
                   s, -s;
                   s,  s;
                  -s,  s];
          Array_of_Polygon = [ Polygon2D(p1, epsSiO2)];
    else
          Array_of_Polygon = [ ];
    end

    Array_of_circle = [ ];

    % adding the layer
    Layers(end + 1) = [layer_2D(d_layer,
                                eps_square,
                                Array_of_Polygon,
                                Array_of_circle,
                                zmin_layer,
                                zmax_layer)];
    current_z = current_z - d_layer;
end
```

Figure B.5: Illustration of a 3D RCWA structure with a pyramid.

Appendix C

Fourier transform formulas

In this annex, the Fourier transform formulas and proof of a 1D constant by part function, and of a 2D polygone, are given. The Fourier transform of the indicator function of polygon, or simply of a polygon, is also given in [76].

C.1 Constant by part 1D function

Definition C.1.1. Given $T \in \mathbb{R}^{+*}$ and given u a periodic real function of period T. One defines the Fourier decomposition of u with:

$$\begin{aligned} u \quad : \quad & \mathbb{R} \to \mathbb{R} \\ & x \quad \mapsto \quad \sum_{n=-\infty}^{+\infty} c_n \exp\left(i \, \frac{2\pi}{T} \, nx\right) \end{aligned} \tag{C.1}$$

$$\tag{C.2}$$

where

$$c_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) \exp\left(-i \frac{2\pi}{T} nt\right) \mathrm{d}t$$

Proposition (Simple case)**.** *Given $T \in \mathbb{R}^{+*}$ and $f \in]0,1[$. Given u a periodic real function of period T defined by:*

$$\begin{aligned} u \quad : \quad & \left]-\frac{T}{2}, \frac{T}{2}\right[\to \mathbb{R} \\ & x \quad \mapsto \quad \begin{cases} \epsilon_A & \text{if } x \in \left]-\frac{T}{2}, -\frac{fT}{2}\right[\; \bigsqcup \; \left]\frac{fT}{2}, \frac{T}{2}\right[\\ \epsilon_B & \text{if } x \in \left]-\frac{fT}{2}, \frac{fT}{2}\right[\end{cases} \end{aligned} \tag{C.3}$$

Then one has:

$$\forall n \in \mathbb{Z}, \quad c_n = \begin{cases} (\epsilon_B - \epsilon_A) \, f + \epsilon_A & \text{if } n = 0 \\ (\epsilon_B - \epsilon_A) \, \dfrac{\sin(\pi n f)}{\pi n} & \text{if } n \neq 0 \end{cases}$$

Proof. Given n in $\mathbb{Z}$:

$$\begin{aligned}
c_n &= \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} u(t)\exp\left(-i\frac{2\pi}{T}nt\right)\mathrm{d}t \\
&= \frac{1}{T}\epsilon_A\left(\int_{\frac{fT}{2}}^{\frac{T}{2}}\exp\left(-i\frac{2\pi}{T}nt\right)+\exp\left(i\frac{2\pi}{T}nt\right)\mathrm{d}t\right)+\frac{1}{T}\epsilon_B\left(\int_{-\frac{fT}{2}}^{\frac{fT}{2}}\exp\left(-i\frac{2\pi}{T}nt\right)\mathrm{d}t\right) \\
&= \frac{1}{T}\epsilon_A\left[\frac{T}{\pi n}\sin\left(\frac{2\pi}{T}nt\right)\right]_{\frac{fT}{2}}^{\frac{T}{2}}+\frac{1}{T}\epsilon_B\left[\frac{T}{2\pi in}\exp\left(i\frac{2\pi}{T}nt\right)\right]_{-\frac{fT}{2}}^{\frac{fT}{2}} \\
&= \frac{\epsilon_A}{\pi n}\left(\sin\left(\pi n\right)-\sin\left(\pi nf\right)\right)+\frac{\epsilon_A}{2\pi in}\left(\exp\left(if\pi n\right)-\exp\left(-if\pi n\right)\right) \\
&= \left(\epsilon_B-\epsilon_A\right)\frac{\sin\left(\pi nf\right)}{\pi n} \quad + \quad \epsilon_A\frac{\sin(\pi n)}{\pi n} \\
&= \begin{cases} \left(\epsilon_B-\epsilon_A\right)f+\epsilon_A & \text{if } n=0 \\ \left(\epsilon_B-\epsilon_A\right)\dfrac{\sin\left(\pi nf\right)}{\pi n} & \text{if } n\neq 0 \end{cases}
\end{aligned}$$

where we used several time the usual formula: $\exp(ix)=\cos(x)+i\sin(x)$; used a variable change $t'=-t$ in the second integral of the second line et used the limited developement of the sinus function in the last line. □

Proposition (Generic case)**.** *Given $T\in\mathbb{R}^+$, $m\in\mathbb{N}^+$ and $\left((x_i)_{i\in[[1,m+1]]}\right)$ in $[0,T]^{m+1}$ such as $0=x_1<x2<...<x_{m-1}<x_{m+1}=T$.*
Given u a periodic real function of period T, constant by part, defined by:

$$\begin{array}{rccl} u & : & [0,T]\setminus\{\,x_i\,|i\in[[1,m]]\,\} & \to & \mathbb{R} \\ & & x & \mapsto & e_i \quad \text{si} \quad x\in\,]x_i,x_{i+1}[\quad \text{et} \quad i\in[[1,m]] \end{array} \tag{C.4}$$

Then one has:

$$\forall n\in\mathbb{Z},\quad c_n=\begin{cases} \sum_{k=1}^{m}\dfrac{e_k}{T}(x_{k+1}-x_k) & \text{if } n=0 \\ \sum_{k=1}^{m}\frac{e_k}{2\pi in}\left(\exp\left(-i\frac{2\pi}{T}nx_k\right)-\exp\left(-i\frac{2\pi}{T}nx_{k+1}\right)\right) & \text{if } n\neq 0 \end{cases}$$

Proof. Given n in $\mathbb{Z}$:

$$\begin{aligned}
c_n &= \frac{1}{T}\int_0^T x(t)\exp\left(-i\frac{2\pi}{T}nt\right)\mathrm{d}t \\
&= \frac{1}{T}\sum_{k=1}^{m}e_k\int_{x_k}^{x_{k+1}}\exp\left(-i\frac{2\pi}{T}nt\right)\mathrm{d}t \\
&= \frac{1}{T}\sum_{k=1}^{m}e_k\left[\frac{-T}{2\pi in}\exp\left(-i\frac{2\pi}{T}nt\right)\right]_{x_k}^{x_{k+1}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{m} \frac{-e_k}{2\pi i n} \left(\exp\left(-i \frac{2\pi}{T} n x_{k+1} \right) - \exp\left(-i \frac{2\pi}{T} n x_k \right) \right) \\
&= \sum_{k=1}^{m} \frac{e_k}{2\pi i n} \left(\exp\left(-i \frac{2\pi}{T} n x_k \right) - \exp\left(-i \frac{2\pi}{T} n x_{k+1} \right) \right) \\
&= \begin{cases} \sum_{k=1}^{m} \frac{e_k}{T} (x_{k+1} - x_k) & \text{if } n = 0 \\ \sum_{k=1}^{m} \frac{e_k}{2\pi i n} \left(\exp\left(-i \frac{2\pi}{T} n x_k \right) - \exp\left(-i \frac{2\pi}{T} n x_{k+1} \right) \right) & \text{if } n \neq 0 \end{cases}
\end{aligned}
$$

where we used the limited developement of the exponential function in the last line. □

C.2 Fourier transform in the plane

Definition C.2.1. Given Ω an open space of $\mathbb{R}^2$, one define the indicator function of Ω, noted χ^Ω, by :

$$
\begin{aligned}
\chi^\Omega \quad : \quad & \mathbb{R}^2 \to \{0, 1\} \\
& (x, y) \mapsto \begin{cases} 0 & \text{if } (x, y) \notin \Omega \\ 1 & \text{if } (x, y) \in \Omega \end{cases}
\end{aligned} \tag{C.5}
$$

Proposition. *Given Ω an open space of $\mathbb{R}^2$, the distribution, noted T_Ω, associated to $\chi \in C_c^\infty(\mathbb{R}^2)$, has a weak derivative in x, noted $T_\Omega^{(x)}$:*

$$
\forall \phi \in C_c^\infty\left(\mathbb{R}^2\right), \qquad T_\Omega^{(x)} \quad : \quad \phi \quad \mapsto \quad - \int_{\partial\Omega} \phi \cdot \boldsymbol{v}_x \, \mathrm{d}\sigma
$$

Proof.

$$
\begin{aligned}
< \partial_x T_\Omega, \phi > \; &:= - \int_\Omega \partial_x \phi(x, y) \mathrm{d}x \mathrm{d}y \\
&= - \int_{\partial\Omega} \phi \cdot \boldsymbol{v}_x \, \mathrm{d}\sigma
\end{aligned}
$$

thanks to Stokes formula. ($\boldsymbol{v}_x$ designates the x component of the normal exterior to Ω). □

Proposition. *If Ω is a polygon with n vertices; whose boundary (in positive orientation) is $p_1 \to p_2 \to ... \to p_n \to p_1$, $p_j = (x_j, y_j)$; then:*

$$
T_\Omega^{(x)} = - \sum_{j=1}^{n} \delta_{[p_j \to p_{j+1}]}
$$

where $\delta_{[p_j \to p_{j+1}]}$ is the distribution defined by:

$$< \delta_{[p_j \to p_{j+1}]}, \phi > \; = \; \int_0^1 \phi((1-t)p_j + tp_{j+1}) \cdot \boldsymbol{v}_x \, \mathrm{d}t. \tag{C.6}$$

where the normal exterior to Ω on the interval $]p_j, p_{j+1}[$, noted $\boldsymbol{v}$, is defined by:

$$\begin{aligned} \boldsymbol{v} &= \mathcal{R}_{-\frac{\pi}{2}} \cdot (p_{j+1} - p_j) \\ &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot (p_{j+1} - p_j) \end{aligned}$$

Remark. The Fourier transform of a function f on $C_c^\infty(\mathbb{R}^2)$ is noted with the help of the distribution:

$$\hat{f}(\omega) =< f, e^{i(w,\cdot)} >$$

Theoreme. *Given Ω a polygon of $\mathbb{R}^2$, given χ^Ω the indicator function of Ω, the Fourier transform of χ^Ω is:*

$$\forall \boldsymbol{\omega} \in \mathbb{R}^2, \quad \hat{\chi}^\Omega(\boldsymbol{\omega}) = \begin{cases} -\frac{1}{\omega_x} \sum_{j=1}^{n} \boldsymbol{v}_x \left[\frac{e^{i\left(\boldsymbol{\omega},\, p_{j+1}\right)} - e^{i\left(\boldsymbol{\omega},\, p_j\right)}}{(\boldsymbol{\omega},\, p_{j+1} - p_j)} \right] & \text{if} \quad \omega_x \neq 0 \\ -\frac{1}{\omega_y} \sum_{j=1}^{n} \boldsymbol{v}_y \left[\frac{e^{i\left(\boldsymbol{\omega},\, p_{j+1}\right)} - e^{i\left(\boldsymbol{\omega},\, p_j\right)}}{(\boldsymbol{\omega},\, p_{j+1} - p_j)} \right] & \text{if} \quad \omega_y \neq 0 \\ \frac{1}{2} \left(x_n y_1 - y_n x_1 + \sum_{k=1}^{n-1} (x_k y_{k+1} - y_k x_{k+1}) \right) & \text{if} \quad \boldsymbol{\omega} = (0,0) \end{cases} \tag{C.7}$$

Proof. Given $\boldsymbol{\omega}$ in $(\mathbb{R}^*)^2$, let's compute $\widehat{\partial_1 \chi}^\Omega(\omega)$:

$$\begin{aligned} \widehat{\partial_1 \chi}^\Omega(\boldsymbol{\omega}) &= -\int_{\mathbb{R}^2} \chi(\boldsymbol{u}) \left(i\omega_x e^{i(\boldsymbol{w},\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\ &= -\int_{\Omega} i\omega_x e^{i(\boldsymbol{w},\boldsymbol{u})} \mathrm{d}\boldsymbol{u} \\ &=< T_\Omega^{(x)}, e^{i(\boldsymbol{\omega},\cdot)} > \\ &= -\sum_{j=1}^{n} < \delta_{[p_j \to p_{j+1}]}, e^{i(\boldsymbol{\omega},\cdot)} > \\ &= -\sum_{j=1}^{n} \boldsymbol{v}_x \cdot \int_0^1 e^{i(\boldsymbol{\omega},\, (1-t)p_j + tp_{j+1})} \, \mathrm{d}t. \\ &= -\sum_{j=1}^{n} \boldsymbol{v}_x \cdot \left[\frac{e^{i(\boldsymbol{\omega},\, (1-t)p_j + tp_{j+1})}}{i\,(\boldsymbol{\omega},\; p_{j+1} - p_j)} \right]_0^1 \end{aligned}$$

$$= -\sum_{j=1}^{n} \boldsymbol{v}_x \cdot \left[\frac{e^{i(\boldsymbol{\omega},\, p_{j+1})} - e^{i(\boldsymbol{\omega},\, p_j)}}{i\,(\boldsymbol{\omega},\; p_{j+1} - p_j)} \right]$$

where we used: the theorem of derivating under the integral (usable since χ has a compact support), then an integral by part, then the definition of χ as an indicator, then the definition of $T_\Omega^{(x)}$, then the proposition on $T_\Omega^{(x)}$ in the case of Ω is a polygon, then the definition of the distribution δ. The final result is obtained by a simple integral calculus.

Since $\widehat{\partial_1 \chi}(\omega) = i\omega_x \hat{\chi}(\omega)$, one gets the results for $\boldsymbol{\omega}$ in $(\mathbb{R}^*)^2$.

If $\boldsymbol{\omega} = (0,0)$, then:

$$\begin{aligned}
\hat{\chi}(\boldsymbol{\omega}) &= \int_\Omega 1 \mathrm{d}t \\
&= \frac{1}{2}\left(\begin{vmatrix} x_n & x_1 \\ y_n & y_1 \end{vmatrix} + \sum_{k=1}^{n-1} \begin{vmatrix} x_k & x_{k+1} \\ y_k & y_{k+1} \end{vmatrix} \right) \\
&= \frac{1}{2}\left(x_n y_1 - y_n x_1 + \sum_{k=1}^{n-1} (x_k y_{k+1} - y_k x_{k+1}) \right)
\end{aligned} \tag{C.8}$$

where we used the formula for the area of a non intersecting polygon, then the formula of the determinant of a 2x2 matrix. □

Remark. Given Ω a convex polygon of $\mathbb{R}^2$, we note, for $\boldsymbol{\omega} \neq (0,0)$, $\hat{\chi}_{int}^{\Omega}(\boldsymbol{\omega})$ the function defined by the previous theorem, where $\boldsymbol{v}$ is the *interior* normal and $\hat{\chi}_{ext}^{\Omega}(\boldsymbol{\omega})$ the function defined by the previous theorem, where $\boldsymbol{v}$ is the *exterior* normal. This notation is extended in $(0,0)$ by assuming $\hat{\chi}_{int}^{\Omega}((0,0)) = \hat{\chi}_{ext}^{\Omega}((0,0)) = \hat{\chi}^{\Omega}((0,0))$.

Theoreme (Fourier transform of a 2D function constant by part)**.** *Given $x_1, x_2, y_1, y_2, T_x, T_y$, $\epsilon_A, \epsilon_B \in \mathbb{R}$ such that $x_1 < x_2$, $y_1 < y_2$, $T_x := x_2 - x_1$ et $T_y := y_2 - y_1$.*

Given S the rectangle $\mathbb{R}^2$ define by $\{(x,y) \in \mathbb{R}^2 \mid (x,y) \in [x_1, x_2] \times [y_1, y_2]\}$.

Given Ω a convex polygon of $\mathbb{R}^2$, stricly included in S, (the indicator function of $\mathbb{R}^2$ are noted with the usual symbol $\mathbb{1}$).

If u, function T_x-periodic in x and T_y-periodic in y, is defined by:

$$\begin{aligned}
u \quad : \quad & \mathbb{R}^2 \to \{\epsilon_A, \epsilon_B\} \\
& (x,y) \mapsto \epsilon_B \mathbb{1}_{S \setminus \Omega}(x,y) + \epsilon_A \mathbb{1}_\Omega(x,y)
\end{aligned} \tag{C.9}$$

then its Fourier transform is:

$$\forall \boldsymbol{\omega} \in \mathbb{R}^2, \quad \hat{u}(\boldsymbol{\omega}) = \epsilon_B \hat{\chi}_{ext}^{S}(\boldsymbol{\omega}) + \epsilon_B \hat{\chi}_{int}^{\Omega}(\boldsymbol{\omega}) + \epsilon_A \hat{\chi}_{ext}^{\Omega}(\boldsymbol{\omega})$$

Proof. Given $\boldsymbol{\omega}$ in $(\mathbb{R}^*)^2$, let's compute $\partial_1\hat{u}(\omega)$:

$$\begin{aligned}\text{Aire}(S)\ \partial_1\hat{u}(\boldsymbol{\omega}) &:= \int_S u(\boldsymbol{t})\left(i\omega_x e^{i(\boldsymbol{w},\boldsymbol{t})}\right)\mathrm{d}\boldsymbol{t}\\ &= \epsilon_B\int_{S\setminus\Omega} i\omega_x e^{i(\boldsymbol{w},\boldsymbol{t})}\mathrm{d}\boldsymbol{t} + \epsilon_A\int_\Omega i\omega_x e^{i(\boldsymbol{w},\boldsymbol{t})}\mathrm{d}\boldsymbol{t}\\ &= \epsilon_B\int_{\partial(S\setminus\Omega)} e^{i(\boldsymbol{w},\sigma)}\cdot\boldsymbol{v}_x\ \mathrm{d}\sigma + \epsilon_A\hat{\chi}_{ext}(\boldsymbol{\omega})\\ &= \epsilon_B\hat{\chi}^S_{ext}(\boldsymbol{\omega}) + \epsilon_B\hat{\chi}^\Omega_{int}(\boldsymbol{\omega}) + \epsilon_A\hat{\chi}^\Omega_{ext}(\boldsymbol{\omega})\end{aligned}$$

Thanks to Stockes formula and the previous theorem for the jump from the second line to the third.
If $\boldsymbol{\omega} = (0,0)$, then:

$$\begin{aligned}\hat{\chi}(\boldsymbol{\omega}) &= \epsilon_B\int_{S\setminus\Omega} 1\mathrm{d}t + \epsilon_A\int_\Omega 1\mathrm{d}t\\ &= \epsilon_B\hat{\chi}^S_{ext}(\boldsymbol{\omega}) + \epsilon_B\hat{\chi}^\Omega_{int}(\boldsymbol{\omega}) + \epsilon_A\hat{\chi}^\Omega_{ext}(\boldsymbol{\omega})\end{aligned}$$

□

Remark. One easily generalizes this formula in the case of several non-intersecting polygons, stricly included in the rectangle S.

Bibliography

[1] J. Grebot, G. Mugny, R. Helleboid, I. Nicholson, F. Abbate, D. Rideau, H. Wehbe-Alause, C. Scheid, and S. Lanteri, "Semi-empirical model for optical properties of $Si_{1-x}Ge_x$ alloys accounting for strain and temperature," in *ESSDERC 2021 - IEEE 51st European Solid-State Device Research Conference (ESSDERC)*, 2021, pp. 267–270.

[2] P. Genevet, F. Capasso, F. Aieta, M. Khorasaninejad, and R. Devlin, "Recent advances in planar optics: from plasmonic to dielectric metasurfaces," *Optica*, vol. 4, no. 1, pp. 139–152, Jan. 2017, publisher: Optica Publishing Group. [Online]. Available: https://opg.optica.org/optica/abstract.cfm?URI=optica-4-1-139

[3] G. A. Sotiriou, F. Starsich, A. Dasargyri, M. C. Wurnig, F. Krumeich, A. Boss, J.-C. Leroux, and S. E. Pratsinis, "Photothermal Killing of Cancer Cells by the Controlled Plasmonic Coupling of Silica-Coated Au/Fe2O3 Nanoaggregates," *Advanced Functional Materials*, vol. 24, no. 19, pp. 2818–2827, 2014, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.201303416. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201303416

[4] "LiDAR drives forwards," *Nature Photonics*, vol. 12, no. 8, pp. 441–441, Aug. 2018. [Online]. Available: https://doi.org/10.1038/s41566-018-0235-z

[5] [Online]. Available: https://en.wikipedia.org/wiki/Diffraction.

[6] R. Helleboid, D. Rideau, I. Nicholson, J. Grebot, B. Mamdy, G. Mugny, M. Basset, M. Agnew, D. Golanski, S. Pellegrini, J. Saint-Martin, M. Pala, and P. Dollfus, "A Fokker-Planck-based Monte Carlo method for electronic transport and avalanche simulation in single-photon avalanche diodes," *Journal of Physics D: Applied Physics*, vol. 55, p. 505102, 2022, publisher: IOP Publishing. [Online]. Available: https://cnrs.hal.science/hal-03828806

[7] [Online]. Available: https://www.yolegroup.com/

[8] L. Pancheri and D. Stoppa, "Low-Noise CMOS single-photon avalanche diodes with 32 ns dead time," in *ESSDERC 2007 - 37th European Solid State Device Research Conference*, 2007, pp. 362–365.

[9] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, T.-J. Wang, K.-C. Huang, Y. Yamashita, D.-N. Yaung, and E. Charbon, "High-Performance Back-Illuminated

Three-Dimensional Stacked Single-Photon Avalanche Diode Implemented in 45-nm CMOS Technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, 2018.

[10] T. Mackay and A. Lakhtakia, "The Transfer-Matrix Method in Electromagnetics and Optics," *Synthesis Lectures on Electromagnetics*, vol. 1, pp. 1–126, Apr. 2020.

[11] K. Yee, "Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media," *IEEE Transactions on Antennas and Propagation*, vol. 14, no. 3, pp. 302–307, May 1966, conference Name: IEEE Transactions on Antennas and Propagation.

[12] J. Hesthaven and T. Warburton, "High-order nodal methods on unstructured grids. I. Time-domain solution of Maxwell's equations," *Journal of Computational Physics*, vol. 181, pp. 186–221, Sep. 2002.

[13] M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," *JOSA*, vol. 71, no. 7, pp. 811–818, Jul. 1981, publisher: Optica Publishing Group. [Online]. Available: https://opg.optica.org/josa/abstract.cfm?uri=josa-71-7-811

[14] P. Vines, K. Kuzmenko, J. Kirdoda, D. C. S. Dumas, M. M. Mirza, R. W. Millar, D. J. Paul, and G. S. Buller, "High performance planar germanium-on-silicon single-photon avalanche diode detectors," *Nature Communications*, vol. 10, no. 1, p. 1086, Mar. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-019-08830-w

[15] J. Vaillant, L. Masarotto, R. Paquet, V. Lecoutre, C. Pelle, N. Moussy, and S. Jouan, "SPAD array sensitivity improvement by diffractive microlens," p. 4.

[16] E. D. Palik, *Handbook of Optical Constants of Solids*, 1985.

[17] "Refractive index database." [Online]. Available: https://refractiveindex.info

[18] S. Botti, A. Schindlmayr, R. Sole, and L. Reining, "Time-dependent density-functional theory for extended systems," *Reports on Progress in Physics*, vol. 70, Mar. 2007.

[19] G. Onida, L. Reining, and A. Rubio, "Electronic excitations: density-functional versus many-body Green's-function approaches," *Reviews of Modern Physics*, vol. 74, no. 2, pp. 601–659, Jun. 2002, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.74.601

[20] S. Refaely-Abramson, M. Jain, S. Sharifzadeh, J. B. Neaton, and L. Kronik, "Solid-state optical absorption from optimally tuned time-dependent range-separated hybrid density functional theory," *Physical Review B*, vol. 92, no. 8, p. 081204, Aug. 2015, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.92.081204

[21] M. Gómez, P. González, J. Ortega, and F. Flores, "Si dielectric function in a local basis representation: Optical properties, local field effects, excitons, and stopping power," *Physical Review B*, vol. 90, no. 20, p. 205210, Nov. 2014, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.90.205210

[22] J. Noffsinger, E. Kioupakis, C. G. Van de Walle, S. G. Louie, and M. L. Cohen, "Phonon-Assisted Optical Absorption in Silicon from First Principles," *Physical Review Letters*, vol. 108, no. 16, p. 167402, Apr. 2012, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.108.167402

[23] C. Schinke, P. Christian Peest, J. Schmidt, R. Brendel, K. Bothe, M. R. Vogt, I. Kröger, S. Winter, A. Schirmacher, S. Lim, H. T. Nguyen, and D. MacDonald, "Uncertainty analysis for the coefficient of band-to-band absorption of crystalline silicon," *AIP Advances*, vol. 5, no. 6, p. 067168, Jun. 2015, publisher: American Institute of Physics. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.4923379

[24] H. Fujiwara and R. W. Collins, Eds., *Spectroscopic Ellipsometry for Photovoltaics: Volume 1: Fundamental Principles and Solar Cell Characterization*, ser. Springer Series in Optical Sciences. Cham: Springer International Publishing, 2018, vol. 212. [Online]. Available: http://link.springer.com/10.1007/978-3-319-75377-5

[25] C. Emminger, F. Abadizaman, N. S. Samarasingha, T. E. Tiwald, and S. Zollner, "Temperature dependent dielectric function and direct bandgap of Ge," *Journal of Vacuum Science & Technology B*, vol. 38, no. 1, p. 012202, Jan. 2020. [Online]. Available: http://avs.scitation.org/doi/10.1116/1.5129685

[26] Y. Niquet, D. Rideau, C. Tavernier, H. Jaouen, and X. Blase, "Onsite matrix elements of the tight-binding Hamiltonian of a strained crystal: Application to silicon, germanium, and their alloys," *Physical Review B: Condensed Matter and Materials Physics (1998-2015)*, vol. 79, no. 24, p. 245201, Jun. 2009, publisher: American Physical Society. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00992736

[27] G. Grosso and G. P. Parravicini, *Solid State Physics*. Elsevier, 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780123044600X50002

[28] L. Valerio, J. S. Jarkko, P. Kai-Erik, and V. Erik M., *Kramers-Kronig Relations in Optical Materials Research*, ser. Springer Series in Optical Sciences. Springer, Berlin, Heidelberg, 2005, https://doi.org/10.1007/b138913. [Online]. Available: X,162

[29] K. D. Moeller, *Optics: Learning by Computing, with Examples Using Maple, MathCad®, Matlab®, Mathematica®, and Maple®*, 2nd ed. New York: Springer-Verlag, 2007. [Online]. Available: https://www.springer.com/fr/book/9780387261683

[30] G. F. Bassani and G. P. Parravicini, *Electronic States and Optical Transitions in Solids*. Franklin Book Company, 1975, google-Books-ID: cGh5AAAAIAAJ.

[31] G. G. Macfarlane, T. P. McLean, J. E. Quarrington, and V. Roberts, "Fine Structure in the Absorption-Edge Spectrum of Ge," *Physical Review*, vol. 108, no. 6, pp. 1377–1383, Dec. 1957, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.108.1377

[32] M. Gómez, P. González, J. Ortega, and F. Flores, "Si dielectric function in a local basis representation: Optical properties, local field effects, excitons, and stopping power," *Physical Review B*, vol. 90, no. 20, p. 205210, Nov. 2014, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.90.205210

[33] V. Olevano, M. Palummo, G. Onida, and R. D. Sole, "Exchange and correlation effects beyond the LDA on the dielectric function of silicon," *Physical Review B*, vol. 60, no. 20, pp. 14 224–14 233, Nov. 1999, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.60.14224

[34] C. Kriso, F. Triozon, C. Delerue, L. Schneider, F. Abbate, E. Nolot, D. Rideau, Y. M. Niquet, G. Mugny, and C. Tavernier, "Modeled optical properties of SiGe and Si layers compared to spectroscopic ellipsometry measurements," *Solid-State Electronics*, vol. 129, pp. 93–96, Mar. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038110116303744

[35] P. Lautenschlager, M. Garriga, L. Vina, and M. Cardona, "Temperature dependence of the dielectric function and interband critical points in silicon," *Physical Review B*, vol. 36, no. 9, pp. 4821–4830, Sep. 1987, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.36.4821

[36] G. E. Jellison and F. A. Modine, "Parameterization of the optical functions of amorphous materials in the interband region," *Applied Physics Letters*, vol. 69, no. 3, pp. 371–373, Jul. 1996, publisher: American Institute of Physics. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.118064

[37] ——, "Erratum: "Parameterization of the optical functions of amorphous materials in the interband region" [Appl. Phys. Lett. 69, 371 (1996)]," *Applied Physics Letters*, vol. 69, no. 14, pp. 2137–2137, Sep. 1996, publisher: American Institute of Physics. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.118155

[38] H. Fujiwara and R. W. Collins, *Spectroscopic Ellipsometry for Photovoltaics: Volume 2: Applications and Optical Data of Solar Cell Materials*, 1st ed. Springer, Jan. 2019.

[39] W. Sellmeier, p. 271, 1871.

[40] B. Johs and C. M. Herzinger, "Brevet Modele paramétrique 1998," 1998.

[41] B. Johs, C. M. Herzinger, J. H. Dinan, A. Cornfeld, and J. D. Benson, "Development of a parametric optical constant model for Hg1-xCdxTe for control of composition by spectroscopic ellipsometry during MBE growth," *Thin Solid Films*, vol. 313-314, pp. 137–142, Feb. 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040609097008006

[42] Y. S. Ihn, T. Kim, T. Ghong, Y.-S. Kim, D. Aspnes, and J. Kossut, "Parametric modeling of the dielectric functions of Cd 1- x Mg x Te alloy films," *Thin Solid Films*, vol. 455, pp. 222–227, May 2004.

[43] M. Cardona, "Modulation spectroscopy of semiconductors," in *Festkörperprobleme 10: Plenary Lectures of the Professional Groups "Semiconductor Physics", "Low Temperature Physics", "Thermodynamics", "Metal Physics" of the German Physical Society Freudenstadt, April 6–11, 1970*, O. Madelung, Ed. Berlin, Heidelberg: Springer, 1970, pp. 125–173. [Online]. Available: https://doi.org/10.1007/BFb0108433

[44] P. Chaisakul, N. Koompai, and P. Limsuwan, "Theoretical investigation of a low-voltage Ge/SiGe multiple quantum wells optical modulator operating at 1310 nm integrated with Si3N4 waveguides," *AIP Advances*, vol. 8, no. 11, p. 115318, Nov. 2018, publisher: American Institute of Physics. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.5064701

[45] A. A. Shklyaev and A. V. Tsarev, "Broadband Antireflection Coatings Made of Resonant Submicron- and Micron-Sized SiGe Particles Grown on Si Substrates," *IEEE Photonics Journal*, vol. 13, no. 3, pp. 1–12, Jun. 2021, conference Name: IEEE Photonics Journal.

[46] W. Traiwattanapong, P. Chaisakul, J. Frigerio, D. Chrastina, G. Isella, L. Vivien, and D. Marris-Morini, "Design and simulation of waveguide-integrated Ge/SiGe quantum-confined Stark effect optical modulator based on adiabatic coupling with SiGe waveguide," *AIP Advances*, vol. 11, no. 3, p. 035117, Mar. 2021, publisher: American Institute of Physics. [Online]. Available: https://aip.scitation.org/doi/10.1063/5.0039129

[47] E. Nolot, J.-M. Hartmann, and J. Hilfiker, "Optical Constants Determination of Pseudomorphic Si1-XGex Layers on Si(001), with 0<x<0.54," *ECS Transactions*, vol. 64, no. 6, p. 455, Aug. 2014, publisher: IOP Publishing. [Online]. Available: https://iopscience.iop.org/article/10.1149/06406.0455ecst/meta

[48] G. Vuye, S. Fisson, V. Nguyen Van, Y. Wang, J. Rivory, and F. Abelès, "Temperature dependence of the dielectric function of silicon using in situ spectroscopic ellipsometry," *Thin Solid Films*, vol. 233, no. 1, pp. 166–170, Oct. 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/004060909390082Z

[49] J. Weber and M. I. Alonso, "Near-band-gap photoluminescence of Si-Ge alloys," *Physical Review B*, vol. 40, no. 8, pp. 5683–5693, Sep.

1989, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.40.5683

[50] Y. P. Varshni, "Temperature dependence of the energy gap in semiconductors," *Physica*, vol. 34, no. 1, pp. 149–154, Jan. 1967. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0031891467900626

[51] R. Braunstein, A. R. Moore, and F. Herman, "Intrinsic Optical Absorption in Germanium-Silicon Alloys," *Physical Review*, vol. 109, no. 3, pp. 695–710, Feb. 1958, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.109.695

[52] X. Liu, C. Geng, X. Ji, S. Lei, and B. Zhang, "Near-IR absorption enhancement and crosstalk reduction of a photodiode in a CMOS indirect time-of-flight sensor," *Applied Optics*, vol. 61, no. 22, pp. 6577–6583, Aug. 2022, publisher: Optica Publishing Group. [Online]. Available: https://opg.optica.org/ao/abstract.cfm?uri=ao-61-22-6577

[53] "Handbook of Optical Constants of Solids - 1st Edition." [Online]. Available: https://www.elsevier.com/books/handbook-of-optical-constants-of-solids/palik/978-0-08-055630-7

[54] K. M. McPeak, S. V. Jayanti, S. J. P. Kress, S. Meyer, S. Iotti, A. Rossinelli, and D. J. Norris, "Plasmonic Films Can Easily Be Better: Rules and Recipes," *ACS Photonics*, vol. 2, no. 3, pp. 326–333, Mar. 2015, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/ph5004237

[55] M. Kupresak, X. Zheng, G. A. E. Vandenbosch, and V. V. Moshchalkov, "Benchmarking of software tools for the characterization of nanoparticles," *Optics Express*, vol. 25, no. 22, pp. 26 760–26 780, Oct. 2017, publisher: Optica Publishing Group. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?uri=oe-25-22-26760

[56] J. R. de Lasson, L. H. Frandsen, P. Gutsche, S. Burger, O. S. Kim, O. Breinbjerg, A. Ivinskaya, F. Wang, O. Sigmund, T. Häyrynen, A. V. Lavrinenko, J. Mørk, and N. Gregersen, "Benchmarking five numerical simulation techniques for computing resonance wavelengths and quality factors in photonic crystal membrane line defect cavities," *Optics Express*, vol. 26, no. 9, p. 11366, Apr. 2018, arXiv:1710.02215 [physics]. [Online]. Available: http://arxiv.org/abs/1710.02215

[57] "Classical Electrodynamics, 3rd Edition | Wiley." [Online]. Available: https://www.wiley.com/en-us/Classical+Electrodynamics%2C+3rd+Edition-p-9780471309321

[58] J. Viquerat, "Simulation de la propagation d'ondes électromagnétiques en nano-optique par une méthode Galerkine discontinue d'ordre élevé," PhD Thesis, 2015. [Online]. Available: http://www.theses.fr/2015NICE4109/document

[59] D. M. Whittaker and I. S. Culshaw, "Scattering-matrix treatment of patterned multilayer photonic structures," *Physical Review B*, vol. 60, no. 4, pp.

2610–2618, Jul. 1999, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.60.2610

[60] J.-M. Jin, M. Zunoubi, K. C. Donepudi, and W. C. Chew, "Frequency-domain and time-domain finite-element solution of Maxwell's equations using spectral Lanczos decomposition method," *Computer Methods in Applied Mechanics and Engineering*, vol. 169, no. 3, pp. 279–296, Feb. 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045782598001583

[61] A. Ditkowski, K. Dridi, and J. S. Hesthaven, "Convergent Cartesian Grid Methods for Maxwell's Equations in Complex Geometries," *Journal of Computational Physics*, vol. 170, no. 1, pp. 39–80, Jun. 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999101967191

[62] Y. Hao and C. Railton, "Analyzing electromagnetic structures with curved boundaries on Cartesian FDTD meshes," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 46, pp. 82–88, Feb. 1998.

[63] A. Gansen, M. El Hachemi, S. Belouettar, O. Hassan, and K. Morgan, "A 3D Unstructured Mesh FDTD Scheme for EM Modelling," *Archives of Computational Methods in Engineering*, vol. 28, no. 1, pp. 181–213, Jan. 2021. [Online]. Available: https://doi.org/10.1007/s11831-019-09395-z

[64] A. Gobé, "Méthodes Galerkin discontinues pour la simulation de problèmes multiéchelles en nanophotonique et applications au piégeage de la lumière dans des cellules solaires," PhD Thesis, 2020. [Online]. Available: http://www.theses.fr/2020COAZ4011/document

[65] L. Fezoui, S. Lanteri, S. Lohrengel, and S. Piperno, "Convergence and stability of a discontinuous Galerkin time-domain method for the 3D heterogeneous Maxwell equations on unstructured meshes," *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, vol. 39, no. 6, pp. 1149–1176, 2005. [Online]. Available: http://www.numdam.org/articles/10.1051/m2an:2005049/

[66] J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods*, ser. Texts in Applied Mathematics, J. E. Marsden, L. Sirovich, and S. S. Antman, Eds. New York, NY: Springer, 2008, vol. 54. [Online]. Available: http://link.springer.com/10.1007/978-0-387-72067-8

[67] P. Monk and Y. Zhang, "Finite Element Methods for Maxwell's Equations," Oct. 2019, arXiv:1910.10069 [cs, math]. [Online]. Available: http://arxiv.org/abs/1910.10069

[68] S. Petersen, C. Farhat, and R. Tezaur, "A space–time discontinuous Galerkin method for the solution of the wave equation in the time-domain," 2000.

[69] L. Angulo, J. Alvarez, M. Pantoja, and S. Garcia, "An Explicit Nodal Space-Time Discontinuous Galerkin Method for Maxwell's Equations," *Microwave and Wireless Components Letters, IEEE*, vol. 24, pp. 827–829, Dec. 2014.

[70] J.-P. Bérenger, "Perfectly Matched Layer (PML) for Computational Electromagnetics," *Synthesis Lectures on Computational Electromagnetics*, vol. 2, pp. 1–117, Jan. 2007.

[71] "Phys. Rev. B 60, 2610 (1999) - Scattering-matrix treatment of patterned multilayer photonic structures." [Online]. Available: https://journals.aps.org/prb/abstract/10.1103/PhysRevB.60.2610

[72] V. Liu and S. Fan, "S4 : A free electromagnetic solver for layered periodic structures," *Computer Physics Communications*, vol. 183, pp. 2233–2244, Oct. 2012.

[73] M. Auer, "Numerical treatment of localized fields in rigorous diffraction theory and its application to light absorption in structured layers," PhD Thesis, Jan. 2016.

[74] P. Duhamel and M. Vetterli, "Fast fourier transforms: A tutorial review and a state of the art," *Signal Processing*, vol. 19, no. 4, pp. 259–299, Apr. 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016516849090158U

[75] S.-W. Lee and R. Mittra, "Fourier transform of a polygonal shape function and its application in electromagnetics," *IEEE Transactions on Antennas and Propagation*, vol. 31, no. 1, pp. 99–103, Jan. 1983, conference Name: IEEE Transactions on Antennas and Propagation.

[76] J. Wuttke, "Form factor (Fourier shape transform) of polygon and polyhedron," *Journal of Applied Crystallography*, vol. 54, no. 2, pp. 580–587, Apr. 2021, arXiv:1703.00255 [math-ph]. [Online]. Available: http://arxiv.org/abs/1703.00255

[77] M. A. Al-Rawhani, J. Beeley, and D. R. S. Cumming, "Wireless fluorescence capsule for endoscopy using single photon-based detection," *Scientific Reports*, vol. 5, no. 1, p. 18591, Dec. 2015.

[78] F. Remondino and D. Stoppa, *TOF range-imaging cameras*. Berlin, Heidelberg: Springer-Verlag, 2013.

[79] G. Agrawal, *Fiber-Optic Communication Systems: Fourth Edition*, Jan. 2012.

[80] T. Arnaud, F. Leverd, L. Favennec, C. Perrot, L. Pinzelli, M. Gatefait, N. Cherault, D. Jeanjean, J.-P. Carrere, F. Hirigoyen, L. Grant, and F. Roy, "Pixel-to-Pixel isolation by Deep Trench technology: Application to CMOS Image Sensor," 2011.

[81] Y. Li, H. Guo, Y. Yao, P. Dutta, M. Rathi, N. Zheng, Y. Gao, S. Sun, J.-H. Ryou, P. Ahrenkiel, and V. Selvamanickam, "Defect reduction by liquid phase epitaxy of germanium on single-crystalline-like germanium templates on flexible, low-cost metal substrates," *CrystEngComm*, vol. 20, no. 41, pp. 6573–6579, Oct. 2018.

[82] A. Ono, A. Ono, K. Hashimoto, N. Teranishi, and N. Teranishi, "Near-infrared sensitivity improvement by plasmonic diffraction for a silicon image sensor with deep trench isolation filled with highly reflective metal," *Optics Express*, vol. 29, no. 14, pp. 21 313–21 319, Jul. 2021.

[83] D. Giubertoni, G. Paternoster, F. Acerbi, X. Borrisé, A. Cian, A. Filippi, A. Gola, A. Guerrero, F. P. Murano, F. Romanato, E. Scattolo, and P. Bellutti, "Plasmonic Enhanced Photodetectors for Near Infra-red Light Detection," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2020.

[84] J. Ma, M. Zhou, Z. Yu, X. Jiang, Y. Huo, K. Zang, J. Zhang, J. S. Harris, G. Jin, Q. Zhang, and J.-W. Pan, "Simulation of a high-efficiency and low-jitter nanostructured silicon single-photon avalanche diode," *Optica*, vol. 2, no. 11, pp. 974–979, Nov. 2015.

[85] Y. Gao, H. Cansizoglu, K. G. Polat, S. Ghandiparsi, A. Kaya, H. H. Mamtaz, A. S. Mayet, Y. Wang, X. Zhang, T. Yamada, E. P. Devine, A. F. Elrefaie, S.-Y. Wang, and M. S. Islam, "Photon-trapping microstructures enable high-speed high-efficiency silicon photodiodes," *Nature Photonics*, vol. 11, no. 5, pp. 301–308, May 2017.

[86] E. Garnett and P. Yang, "Light Trapping in Silicon Nanowire Solar Cells," *Nano Letters*, vol. 10, no. 3, pp. 1082–1087, Mar. 2010.

[87] K. Zang, X. Jiang, Y. Huo, X. Ding, M. Morea, X. Chen, C.-Y. Lu, J. Ma, M. Zhou, Z. Xia, Z. Yu, T. I. Kamins, Q. Zhang, and J. S. Harris, "Silicon single-photon avalanche diodes with nano-structured light trapping," *Nature Communications*, vol. 8, no. 1, p. 628, Sep. 2017.

[88] Y. Cao, Z. Zhang, and K. X. Wang, "Photon management with superlattice for image sensor pixels," *AIP Advances*, vol. 11, no. 8, p. 5314, Aug. 2021.

[89] S. Yokogawa, I. Oshiyama, H. Ikeda, Y. Ebiko, T. Hirano, S. Saito, T. Oinoue, Y. Hagimoto, and H. Iwamoto, "IR sensitivity enhancement of CMOS Image Sensor with diffractive light trapping pixels," *Scientific Reports*, vol. 7, no. 1, p. 3832, 2017.

[90] M. M. R. Elsawy, S. Lanteri, R. Duvigneau, J. A. Fan, and P. Genevet, "Numerical Optimization Methods for Metasurfaces," *Laser & Photonics Reviews*, vol. 14, no. 10, p. 1900445, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lpor.201900445. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lpor.201900445

[91] N. Teranishi, T. Yoshinaga, K. Hashimoto, and A. Ono, "Near-infrared sensitivity enhancement of image sensor by 2nd-order plasmonic diffraction and the concept of resonant-chamber-like pixel," in *2022 International Electron Devices Meeting (IEDM)*, 2022, pp. 37.2.1–37.2.4.

[92] J. Mockus, "On Bayesian Methods for Seeking the Extremum," in *Optimization Techniques*, 1974.

[93] H. J. Kushner, "A versatile stochastic model of a function of unknown and time varying form," *Journal of Mathematical Analysis and Applications*, vol. 5, no. 1, pp. 150–167, 1962. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022247X62900112

[94] ——, "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise," *Journal of Basic Engineering*, vol. 86, no. 1, pp. 97–106, Mar. 1964, _eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/86/1/97/5763745/97_1.pdf. [Online]. Available: https://doi.org/10.1115/1.3653121

[95] D. Jones, M. Schonlau, and W. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, pp. 455–492, Dec. 1998.

[96] A. N. Shiryaev and R. P. Boas, *Probability (2nd Ed.)*. Berlin, Heidelberg: Springer-Verlag, 1995.

[97] R. V. MISES, "CHAPTER XI - MULTIVARIATE STATISTICS. CORRELATION," in *Mathematical Theory of Probability and Statistics*, R. V. MISES, Ed. Academic Press, 1964, pp. 566–614. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9781483232133500142

[98] R. Durrett, "Martingales," in *Probability: Theory and Examples*, 4th ed., ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010, pp. 221–273.

[99] I. A. Ibragimov, *Processus aléatoires gaussiens / par I. Ibrahimov et Y. Rozanov ; [traduit du russe par A. Sokova]*. Moscou: Éditions Mir, 1974, publication Title: Processus aléatoires gaussiens.

[100] G. Matheron, "Principles of geostatistics," *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, Dec. 1963, publisher: Society of Economic Geologists. [Online]. Available: https://doi.org/10.2113%2Fgsecongeo.58.8.1246

[101] D. G. Krige, S. A. I. o. Mining, and Metallurgy., *Lognormal-de Wijsian geostatistics for ore evaluation / D.G. Krige*, 2nd ed. South African Institute of Mining and Metallurgy Johannesburg [South Africa], 1981, type: Book.

[102] M. D. McKay, R. J. Beckman, and W. J. Conover, "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, May 1979.

[103] J.-S. Park, "Optimal Latin-hypercube designs for computer experiments," *Journal of Statistical Planning and Inference*, vol. 39, no. 1, pp. 95–111, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378375894901155

[104] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *Journal of Statistical Planning and Inference*, vol. 26, no. 2, pp. 131–148, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/037837589090122B

[105] O. Dubrule, "Cross validation of kriging in a unique neighborhood," *Journal of the International Association for Mathematical Geology*, vol. 15, pp. 687–699, 1983.

[106] D. J. C. MacKay, "Bayesian Interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, May 1992, _eprint: https://direct.mit.edu/neco/article-pdf/4/3/415/812340/neco.1992.4.3.415.pdf. [Online]. Available: https://doi.org/10.1162/neco.1992.4.3.415

[107] O. Roustant, D. Ginsbourger, and Y. Deville, "DiceKriging , DiceOptim : Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization," *Journal of Statistical Software*, vol. 51, Jan. 2013.

[108] N. Hansen, "The CMA Evolution Strategy: A Tutorial," 2016. [Online]. Available: https://arxiv.org/abs/1604.00772

[109] V. Picheny, T. Wagner, and D. Ginsbourger, "A Benchmark of Kriging-Based Infill Criteria for Noisy Optimization," *Struct. Multidiscip. Optim.*, vol. 48, no. 3, pp. 607–626, Sep. 2013, place: Berlin, Heidelberg Publisher: Springer-Verlag. [Online]. Available: https://doi.org/10.1007/s00158-013-0919-4

[110] D. D. Cox and S. John, "A statistical method for global optimization," *[Proceedings] 1992 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1241–1246 vol.2, 1992.

[111] D. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *J. of Global Optimization*, vol. 21, pp. 345–383, Dec. 2001.

[112] W. J. Welch and M. Schonlau, "Computer experiments and global optimization," 1997.

[113] D. Jones, M. Schonlau, and W. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, pp. 455–492, Dec. 1998.

[114] edited by Edward D. Palik, *Handbook of optical constants of solids.* Orlando : Academic Press, 1985., 1985. [Online]. Available: https://search.library.wisc.edu/catalog/999554063402121

[115] T. Bright, J. Watjen, Z. Zhang, C. Muratore, A. Voevodin, D. Koukis, D. Tanner, and D. Arenas, "Infrared optical properties of amorphous and nanocrystalline Ta2O5 thin films," *Journal of Applied Physics*, vol. 114, Aug. 2013.

[116] M. Kuss and C. E. Rasmussen, "Assessing Approximations for Gaussian Process Classification," in *NIPS*, 2005.

[117] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.*, ser. Adaptive computation and machine learning. MIT Press, 2006.

[118] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, no. 1, pp. 271–280, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378475400002706

[119] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur, "Asymptotic normality and efficiency of two Sobol index estimators," *ESAIM: Probability and Statistics*, vol. 18, pp. 342–364, 2014, publisher: EDP-Sciences. [Online]. Available: http://www.numdam.org/articles/10.1051/ps/2013040/

[120] L. Le Gratiet, C. Cannamela, and B. Iooss, "A Bayesian Approach for Global Sensitivity Analysis of (Multifidelity) Computer Codes," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 2, no. 1, pp. 336–363, 2014, _eprint: https://doi.org/10.1137/130926869. [Online]. Available: https://doi.org/10.1137/130926869

[121] "Diogenes: A Discontinuous-Galerkin based software suite for nano-optics." [Online]. Available: https://diogenes.inria.fr/

[122] C.-H. Hsu, S.-M. Liu, W.-Y. Wu, Y.-S. Cho, P.-H. Huang, C.-J. Huang, S.-Y. Lien, and W.-Z. Zhu, "Nanostructured pyramidal black silicon with ultra-low reflectance and high passivation," *Arabian Journal of Chemistry*, vol. 13, no. 11, pp. 8239–8247, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1878535220300083

[123] A. Loseille and R. Feuillet, "Vizir: High-order mesh and solution visualization using OpenGL 4.0 graphic pipeline," in *2018 - AIAA Aerospace Sciences Meeting, AIAA SciTech Forum*, kissimmee, United States, Jan. 2018, pp. 1–13. [Online]. Available: https://hal.inria.fr/hal-01686714

[124] C. Geuzaine and J.-F. Remacle, "Gmsh: A 3-D Finite Element Mesh Generator with Built-in Pre- and Post-Processing Facilities," *International Journal for Numerical Methods in Engineering*, vol. 79, pp. 1309 – 1331, Sep. 2009.